Potentielle Privatsphäreverletzungen aufdecken und automatisiert sichtbar machen

Bäumer, Frederik Simon
fbaeumer@mail.upb.de
Universität Paderborn, Deutschland

Buff, Bianca
bbuff@mail.upb.de
Universität Paderborn, Deutschland

Geierhos, Michaela
geierhos@mail.upb.de
Universität Paderborn, Deutschland

Das moderne Web basiert auf Interaktion, Diskussion und Austausch von Informationen. Durch die fortschreitende semantische Anreicherung wird das Web auch zu einer riesigen Informationsquelle für datengesteuerte Anwendungen, wie sie auch in Digital Humanities Verwendung finden. Dies stellt unter Umständen ein Risiko für einzelne Benutzer[1] dar. Da Daten immer effektiver mit bestehenden Ressourcen verknüpft werden, können selbst ungewollt (implizit) offenbarte Einzelinformationen schädliche Folgen für einzelne Nutzer haben. Obwohl *Serviceprovider* im Web die Pflicht und auch das Eigeninteresse haben, die Sicherheit und Privatsphäre von Benutzerdaten zu gewährleisten, gibt es Fälle, in denen Benutzerdaten missbraucht und kompromittiert oder öffentlich verfügbare Informationen gegen dessen ursprünglichen Verfasser verwendet werden (Gross, et al., 2005). Die bestehenden Datenschutzrichtlinien, Betreiberhinweise und (teil-)automatisierte Schutzmechanismen, welche die Privatsphäre von Personen schützen sollen, sind aber oftmals unzureichend. Es ist demnach im Interesse der Kommunizierenden, nur diejenigen Informationen in Textbeiträgen zu platzieren, die einen gewissen selbstbestimmten Grad an Anonymität wahren. Darüber hinaus ist es für alle die, die mit Daten arbeiten wollen (bspw. in der Forschung) im Interesse, private Daten filtern zu können. In diesem Beitrag stellen wir unsere bisherigen Arbeiten an *Text Broom* vor, einem ersten Prototypen, welcher potentielle Privatsphäreverletzungen in Form expliziter als auch inhärenter Angaben in online verfügbaren Fließtexten erkennen sowie sichtbarmachen kann und so eine Hilfe für Benutzer als auch für die gefahrlose Weiterverwendung von Daten darstellen kann.

Privatsphäreverletzungen: Erkennung als Herausforderung

Wie sich unwissentliche Informationspreisgaben in sprachlichen Ausdrücken manifestieren, wurde bisher unzureichend untersucht. Frühere Arbeiten zeigten jedoch auf, dass sprachliche Formulierungen oft mehr Informationen enthalten, als es zunächst den Anschein erweckt. Um diese zu erkennen, wurden vordefinierte Muster verwendet, die nur begrenzt dem Gestaltungsfreiraum natürlicher Sprache gerecht werden und nur offensichtliche (explizite) Informationspreisgabe feststellen (Bäumer, et al., 2017). In diesem Kontext existieren Vorarbeiten, wie die von Sweeney (1996), Dias (2016) sowie von Kleinberg und Mozes (2017). Dabei besonders erwähnenswert ist das Tool NETANOS (*Named Entity-based Text ANonymization for Open Science*) von Kleinberg und Mozes (2017), das benannte Entitäten in Fließtexten erkennen und hervorheben kann. Hierbei handelt es sich stets um benannte Entitäten (z. B. Personennamen), deren wörtliche Nennung zwar eine Gefahr für die Privatsphäre der Betroffenen darstellen kann, deren isolierte Erkennung jedoch trivial im Vergleich zur Behandlung der Ausdruckskomplexität von Privatsphäreverstößen in Fließtexten ist. Denn immer noch fehlt es an Wissen über die genaue sprachliche Manifestierung und an computerlinguistischen Verfahren, die darauf zurückgreifen können. Dies ist allerdings zwingend erforderlich, um entsprechende privatsphäregefährdende Textbestandteile zu identifizieren und mit einer Erläuterung möglicher Risiken zu versehen.

Arztbewertungen als Untersuchungsgegenstand

Service Provider nehmen im Web unterschiedliche Gestalt an, jedoch steht zumeist eine zentrale Dienstleistung im Mittelpunkt, wie es beispielsweise der Erwerb von Produkten bei Online-Shops oder entsprechende Meinungsäußerungen auf Bewertungsportalen sind. Ein medial vielbeachtetes Beispiel in diesem Zusammenhang sind sogenannte *Physician Review Websites* (PRWs), die es den Nutzern ermöglichen, medizinische Dienstleistungen und damit auch die bislang als sensibel geltende Arzt-Patienten-Beziehung anonym zu bewerten. Um eine authentische Bewertung zu erstellen, ergänzen Bewertende vielfach private Informationen, z. B. über Orte, Krankheiten oder Medikamente und sind so potentiell für Dritte identifizierbar (z. B. Ärzte, Freunde). Hier stehen nicht nur explizit genannte Informationen („Ich bin Diabetiker“) im Fokus, sondern auch inhärente Angaben („Ich bin Vater“ # männlich) sowie Metainformationen (Datum der Bewertung, Alter, Krankenkasse, Ort der Praxis). Oftmals ergibt sich eine Gefahr für die Privatsphäre nicht aus einer

einzelnen Information, sondern aus der Summe expliziter und inhärenter Angaben. Das Problem wird verschärft, wenn Patient und Bewertender nicht *in persona* auftreten und somit eine potentielle Privatsphäreverletzung an einer dritten Person (z. B. Kinder, Eltern) vorliegt (Geierhos & Bäumer, 2015). Arztbewertungen eignen sich somit auf Grund der hohen Gefahr unabsichtlich preisgegebener Informationen und auch auf Grund ihrer langjährigen und öffentlichen Zugänglichkeit. So steht uns ein Korpus zur Verfügung, welches ca. 900.000 deutschsprachige Patientenberichte inklusive Metainformationen enthält und die Zeitspanne von 2007 bis 2016 abdeckt (Bäumer et al., 2015).

Annotation privater Angaben in der medizinischen Domäne

Privatsphäreverletzungen können sich vielfältig im nutzergenerierten Texten manifestieren und sind zusätzlich auf Grund stark schwankender Textqualität schwer automatisiert zu erkennen. Deshalb werden neben präzisen, aber gering toleranten linguistischen Mustern auch Methoden des Maschinellen Lernens eingesetzt. Hier bedarf es umfangreicher Annotationen in Trainingsdaten, um private Angaben in unbekannten Texten automatisiert zu erkennen. Um eine große Anzahl an Texten annotieren zu können, nutzen wir das Annotationstool Prodigy, welches insbesondere binäre Annotationsentscheidungen mittels *Active Learning* merklich beschleunigen kann. Die Herausforderungen, die dennoch mit der Annotation einhergehen, werden im Folgenden an dem vermeintlich trivialen Beispiel der Annotation von Krankheiten im oben genannten Korpus aufgezeigt.

Bereits die Frage, was genau annotiert werden soll, ist nicht trivial: Oberbegriffe wie „Krankheit“ und „Symptom“ werden nur ergänzend mit einer Spezifizierung annotiert. Konkreter heißt dies, dass das Wort „Erkrankung“ nicht annotiert wird, jedoch das Kompositum bzw. die Nominalphrase „Herzerkrankung“ und „psychische Erkrankung“ schon. Personenbezeichnungen wie „Magersüchtige“, „Diabetiker“ und „Neurodermitiker“ und Adjektive wie „magersüchtig“ und „laktoseintolerant“ werden annotiert, da sie implizieren, dass eine Person an der jeweiligen Krankheit leidet. Die Annotation von Adjektiven stellt dabei eine Herausforderung dar. In den folgenden Beispielen enthält das attributive Adjektiv einen essentiellen Bestandteil der Semantik der Phrase: „kardiovaskuläre Erkrankung“, „bulimischer Anfall“. Ohne die Attribuierung würde das Substantiv „Erkrankung“ bzw. „Anfall“ nicht als „Krankheit“ annotiert werden. Durch das Adjektiv wird die Phrase jedoch (annähernd) gleichbedeutend mit den Ausdrücken „Herz-Kreislauf-Erkrankung“ und „Bulimie“. Da letztgenannte als „Krankheit“ annotiert werden würden, werden gleichermaßen auch alle Phrasen, die ein unspezifisches Substantiv (z. B. „Erkrankung“) sowie ein attributives Adjektiv (z. B. „kardiovaskuläre“), das den wesentlichen Teil der Semantik trägt, enthalten, als „Krankheit“ markiert. Im Gegensatz dazu werden attributive Adjektive, die eine Krankheit als solche nur näher beschreiben, nicht mit annotiert. Ein Beispiel dafür ist der „wässrige Durchfall“: In diesem Fall wird lediglich das Substantiv als „Krankheit“ annotiert und das Adjektiv nicht beachtet. In Bezug auf das Adjektiv „chronisch“, welches bei Krankheitsbildern oder Symptomen häufig attributiv verwendet wird, gilt, dass dieses Wort nicht mit annotiert wird, da es nicht maßgeblich zu der Semantik der Krankheit beiträgt. In dem Satz „Ich habe chronischen Durchfall und chronische Schmerzen.“ wird folglich nur „Durchfall“ als Krankheit annotiert.

Dies ist nur ein Beispiel für die Komplexität der Thematik und der Herausforderungen bei der Ressourcenerstellung. Ein weiteres Beispiel sind Ambiguitäten, wie sie u. a. bei „Die Ärztin sollte nicht wie meine Mutter sein“ und „Meine Mutter hat auch MC“ vorkommen. Während der erste Satz unkritisch ist, wird im zweiten Satz die Privatsphäre einer dritten Person gefährdet. Die Interpretation des Wortes „Mutter“ gelingt nur im Kontext und zeigt, dass eine rein wortbasierte Vorgehensweise nicht zielführend ist. Aus diesem Grund wird unser Tool Text Broom zwar wie dargestellt auf Basis domänenspezifischer Texte trainiert, nutzt aber eine umfangreiche NLP-Pipeline zur kontextspezifischen Interpretation.

Text Broom

Wie dargestellt, adaptiert unser Prototyp *Text Broom* die Idee von Kleinberg und Mozes (2017), die sich auf benannte Entitäten konzentrieren. Allerdings geht *Text Broom* darüber hinaus, indem es potentielle Privatsphäreverletzungen mit Hilfe einer Textverarbeitungspipeline (*Multi-Stage*-Ansatz) erkennt, die unterschiedliche Granularitätsstufen bietet. Somit verarbeitet die *Text-Broom*-Pipeline ein viel breiteres Spektrum an linguistischen Informationen, aufgeteilt in vier Phasen. Stufe I enthält eine Vorverarbeitung, die grundlegende Sprachverarbeitung wie *Part-Of-Speech* (POS) *Tagging* verwendet. Stufe II kombiniert *Semantic Role Labeling*, linguistische Muster und Eigennamenerkennung. Dies sind nicht-domänenspezifische Komponenten, die eine breite thematische Abdeckung ermöglichen. Im Gegensatz dazu enthält Stufe III eine domänenspezifische Informationsextraktion, eine Komponente zur Phrasenklassifizierung und die finale Bewertungskomponente, die alle bis zu diesem Zeitpunkt gesammelten Informationen zusammenfasst und auswertet. Die letzte Stufe IV enthält Komponenten zur Visualisierung und Erläuterung. Nutzer erhalten somit eine vielschichtige Sicht auf ihre Texte, in denen Privatsphäreverstöße explizit hervorgehoben werden.

Fußnoten

1. Aus Gründen der leichteren Lesbarkeit wird auf eine geschlechtsspezifische Differenzierung verzichtet. Entsprechende Begriffe gelten im Sinne der Gleichbehandlung für beide Geschlechter.

Bibliographie

Kleinberg, Bennet / Mozes, Maximilian (2017): *Web-based text anonymization with Node.js: Introducing NETANOS (Named entity-based Text Anonymization for Open Science)*. The Journal of Open Source Software.

Dias, Francesco (2016): *Multilingual Automated Text Anonymization*. Inst. Superior Técnico of Lisboa, Lissabon, Portugal.

Bäumer, Frederik S. / Geierhos, Michaela / Schulze, Sabine (2015): *A System for Uncovering Latent Connectivity of Health Care Providers in Online Reviews*. In **Dregvaite, Giedre / Damaševi#ius, Robertas (Hsg.)**: *Communications in Computer and Information Science*, Band 538. ICIST 2015, Litauen, Oktober 15-16, 2015. Proceedings (S. 3–15). Cham, Schweiz: Springer International.

Bäumer, Frederik / Grote, Nicolai / Kersting, Joschka / Geierhos, Michaela (2017): *Privacy Matters: Detecting Nocuous Patient Data Exposure in Online Physician Reviews*. In **Damaševi#ius, Robertas & Mikašyt#, Vilma (Hsg.)**: *Communications in Computer and Information Science*, Band 756. ICIST 2017, Litauen, 12.-14. Oktober 2017, Proceedings (S. 77–89). Cham, Schweiz: Springer International.

Sweeney, Latanya (1996): *Replacing personally-identifying information in medical records, the Scrub system*. Proceedings of the AMIA annual fall symposium. American Medical Informatics Association. S. 333-337.

Geierhos, Michaela / Bäumer, Frederik S. (2015): *Erfahrungsberichte aus zweiter Hand: Erkenntnisse über die Autorschaft von Arztbewertungen in Online-Portalen*. In: DHd 2015: Book of Abstracts, ZIM-ACDH, Graz, Österreich, 2015, S. 69-72.

Gross, Ralph / Acquisti, Alessandro (2005): *Information Revelation and Privacy in Online Social Networks (The Facebook case)*. In Proceedings of the ACM Workshop on Privacy in the Electronic Society, S. 71-80, Alexandria, USA, 2005.