Richtlinie zum „Software-Sprint“

BIASLYZE – The NLP Bias Identification Toolkit

Schlussbericht

Zuwendungsempfänger*innen:

Tobias Sterbak & Stina Lohmüller

Das diesem Bericht zugrundeliegende Vorhaben wurde mit Mitteln des Bundesministeriums für Bildung und Forschung unter dem Förderkennzeichen 01IS23S04 gefördert. Die Verantwortung für den Inhalt dieser Veröffentlichung liegt bei den Autor*innen.

Kurze Darstellung der Aufgabenstellung und Motivation

Was war Deine Motivation? Welches Problem wolltest Du mit Deinem Projekt lösen? Wie war die geplante Vorgehensweise zur Problemlösung (auch Angabe der wichtigsten Meilensteine)?

Große Sprachmodelle werden zunehmend für die Entwicklung von NLP-Anwendungen, gerade von weniger erfahrenen Entwickler*innen und Teams mit geringen Ressourcen eingesetzt. Mögliche Risiken und Diskriminierungen, die durch die Modelle entstehen können, werden dabei häufig nicht berücksichtigt. Unter anderem, weil sich ein Bewusstsein für die Risiken erst langsam aufbaut, ebenso wie die soziotechnischen Kompetenzen, die nötig sind, um diesen zu begegnen. Mit dem Projekt “BIASLYZE - The NLP Bias Identification Toolkit” entwickeln wir ein zugängliches Tool für Entwickler*innen, um große Sprachmodelle mit modernen Methoden auf unterschiedliche gesellschaftlich relevante Biases zu untersuchen. Auf diese Weise wollen wir Entwickler*innen befähigen sich mit Quellen möglicher Diskriminierung im Bereich NLP auseinanderzusetzen und sie zu minimieren.

Unser Umsetzungsplan anhand von Meilensteinen:

- Implementierung von ersten Methoden als Python Package
  - Zusammenstellung von Wort- und Konzeptlisten schützenswerter Attribute
  - Implementierung von “targeted LIME”
  - Darstellung eines Anwendungsfalls mit Beispieldaten (Tutorial)
- Validierung und Erweiterung der Wort- und Konzeptlisten mit Betroffenengruppen
- Entwicklung eines ersten React-Frontends
- Erprobung unterschiedlicher Visualisierungsansätze der Analyseergebnisse
- Verbesserung der Ergebniseinordnung sowie Vorschläge von Mitigationsmaßnahmen und weiterführenden Infos.

Beitrag des Projektes zu den Zielen der Förderinitative „Software-Sprint“

Welche Bezüge gibt es zu den Themenfeldern „Civic Tech“ und „Data Literacy“ des Software Sprints oder zu weiteren gesellschaftlich relevanten Zielen bzw. Lösungsansätzen?

Als Open Source Toolkit wurde BIASLYZE von Beginn an als ein kollaboratives Projekt geplant und umgesetzt. Ziel des Projektes war es, ein Werkzeug zu entwickeln, dass aufwandsarm in bestehende Arbeitsprozesse von Entwickler*innen eingebracht werden kann. BIASLYZE bietet im Bereich NLP einen Einstieg in Bias Testing und Mitigierung sowie in die Auseinandersetzung mit gesellschaftlich relevanten Fragen zu Diskriminierungsrisiken algorithmischer Entscheidungssysteme. Das Tool ermöglicht und konkretisiert damit auch die Bearbeitung soziotechnischer Probleme. Zudem können Organisationen und Unternehmen BIASLYZE nutzen, um bereits trainierte Modelle zu analysieren und so eine informierte Entscheidung über deren Anwendung treffen.

Ausführliche Darstellung der Ergebnisse

Welche konkreten Ergebnisse hast Du erzielt? Konnten alle Meilensteine erreicht werden? Welche zusätzlichen Erkenntnisse hast Du aus der Projektarbeit gewonnen, auch im Hinblick auf die Begleitung durch die Open Knowledge Foundation?

Wir konnten unsere Ziele und Meilensteine weitestgehend wie geplant umsetzen. Dabei haben wir eine Toolbox zur Erkennung von Hinweisen auf Bias in NLP-Modellen und LLMs in python implementiert. Dabei haben wir auf flexible Darstellung der Ergebnisse in Visualisierungen wert gelegt. Die verwendeten Methoden und Visualisierungen basierend auf “counterfactual token fairness“ wurden auf der zugehörigen Website dokumentiert. Zudem gibt es Anwendungsbeispiele und Tutorials um den Einstieg in die Toolbox zu erleichtern.

Für Open Source Projekte ist Marketing und ein professioneller Auftritt und Logo wichtig und wir haben dabei alles richtig gemacht und uns frühzeitig darum gekümmert. Die Betreuung durch das Team der Open Knowledge Foundation war hervorragend und schnell. Die kurze, regelmäßige Reflexion über den Projektfortschritt und Herausforderungen in Form der Montagsupdates war sowohl hilfreich als auch motivierend.

Zielgruppe, Nutzen und mögliche Weiterentwicklungen

Welcher Nutzen ergibt sich für die Zielgruppe aus den Ergebnissen Deines Projekts? Welche weitergehenden Effekte ergeben sich aus der Open-Source-Stellung der Ergebnisse? Gibt es Ideen für die Weiterentwicklung Deiner Lösung und Pläne zu deren Umsetzung?

Mit dem Projekt unterstützen wir Entwickler*innen von Software mit NLP-Komponenten um deren ethisch verantwortungsvollen Einsatz zu ermöglichen. Unsere Toolbox bietet weniger erfahrenen Entwickler*innen, Studierenden und Teams mit begrenzten Ressourcen einen aufwandsarmen Zugang zu Bias-Testing und Mitigation in NLP-Use Cases. Zudem soll sie in Projekten des gemeinnützigen Vereins Beyond AI Collective e.V. eingesetzt werden und so für Risiken algorithmenvermittelter Diskriminierung sensibilisieren. Darüber hinaus haben wir uns mit dem Projekt für eine Förderung der Mozilla Foundation beworben um die Entwicklung und

Implementierung von Methoden zur Untersuchung von Trainingsdaten auf Bias durchzuführen. Auch die Unterstützung von weiteren Sprachen ist geplant.

Hat die Arbeit in dem Projekt Dich in Deiner persönlichen, fachlichen Weiterentwicklung unterstützt?

Tobias: Durch das Projekt konnte ich mein Wissen im ethischen Einsatz von KI und Bias-Erkennung erweitern und vertiefen.

Stina: Durch das Projekt konnte ich mein Wissen über Softwareentwicklung in python, sowie angewandte Methoden der Bias-Erkennung im Bereich NLP erweitern und vertiefen.

Kurze Darstellung der Arbeiten, die zu keiner Lösung geführt haben

Gab es Arbeiten bzw. Lösungsansätze, die nicht weiter verfolgt wurden? Was waren die Hintergründe, und wie bist Du alternativ vorgegangen?

In einer ersten explorativen Phase hatten wir die LIME Methode als beste Option für bias detection identifiziert. Bei der weiteren Recherche und in den Tests erwies sich dann aber eine andere Methode, die auf „counterfactual token fairness“ beruht als präziser und schneller.

Wir haben außerdem in der Entwicklung mit unterschiedlichen Arten der Visualisierung der Ergebnisse experimentiert. Dabei haben wir einige Varianten und Methoden, die zunächst vielversprechend schienen, verworfen.

In den Usertestings hat sich schließlich herausgestellt, dass unsere Zielgruppe direkt mit Python-Code arbeiten möchte. Wir haben uns deshalb entschieden kein umfangreicheres Frontend zu entwickeln, da sich das für die Zielgruppe als nicht interessant erwiesen hat und stattdessen ein umfangreiches Python package entwickelt.

Kurze Angabe von Präsentationsmöglichkeiten für mögliche Nutzer

Wo können sich Interessenten detailliert über Deine Projektergebnisse informieren (z.B. Webseite, GitHub, Veröffentlichungen)?

Code (Github): https://github.com/biaslyze-dev/biaslyze
Website/Dokumentation: https://biaslyze.org/

Kurze Erläuterung zur Einhaltung der Arbeits- und Kostenplanung

Gab es im Projektverlauf Ereignisse, die eine Anpassung der Planung erforderlich machten – z.B. Mehr- oder Minderaufwand bei der Bearbeitung von Teilaufgaben?

Einige Experimente zu vorgeschlagenen Methoden waren nicht vielversprechend, bzw. hatten Probleme. Nachdem die LIME-basierte Methode nicht gut funktioniert hat mussten wir eine neue Methode finden und erproben. So sind wir beim Counterfactual Score gelandet.

In den Usertestings hat sich schließlich herausgestellt, dass unsere Zielgruppe in python arbeiten möchte. Wir haben uns deshalb entschieden kein umfangreicheres Frontend zu entwickeln, da sich das für die Zielgruppe als nicht interessant erwiesen hat und stattdessen ein umfangreiches Python package entwickelt.

Unsere Planung konnte diese Planänderungen gut verkraften, da durch den Verzicht auf das Frontend Ressourcen für die Entwicklung und Erprobung neuer Methoden frei wurden. So wir sind bei einem guten Stand des Projekts angekommen.

Kurze Darstellung von etwaigen Ergebnissen bei anderen Stellen

Gab es Entwicklungen anderer Personen oder Institutionen, die Einfluss auf Deine Arbeiten und die Zielsetzung hatten? Wenn ja, worin bestand dieser und wie bist Du damit umgegangen?

Die verwendeten Methoden der Bias-Erkennung bauen auf unterschiedlichen wissenschaftlichen Veröffentlichungen auf. Diese haben unsere Arbeit und auch unsere Ziele direkt beeinflusst.