Einleitung

(Der Bereich der Künstlichen Intelligenz (KI) ist ein rasant wachsender Forschungszweig, mit unzähligen wertvollen Beiträgen aus der ganzen Welt. Dabei sind die Anwendungsmöglichkeiten genauso vielfältig wie die verschiedenen Datenquellen dazu. Da diese Datenquellen allerdings alle direkt oder indirekt durch Menschen erschaffen sind, repräsentieren sie teilweise auch deren Vorurteile und Weltbilder. Auch wenn diese Verzerrungen größtenteils unbeabsichtigt in die Daten projiziert werden, sind sie dennoch vorhanden und können dazu führen, dass sich die KI unfair verhält und bestimmten Gruppen Vorteile gegenüber anderen Gruppen verschafft. So kann es beispielsweise vorkommen, dass weißen Männern eher ein Kredit gewährt wird als Frauen oder dunkelhäutigen Menschen, wenn für diese Entscheidung ein Machine Learning (ML) Algorithmus verwendet wird. Unser Ziel dabei ist es, ein Artefakt zu entwickeln, welches solche Vorurteile/Verzerrungen (Bias) in einem tabellarischen Datensatz erkennt und ihn idealerweise gleich bereinigt.)

Der Bereich des Künstlichen Intelligenz und Machine Learnings ist ein rasant wachsender Forschungsbereich, welcher in vielen Bereichen Ausprägungen vorzuweisen hat. KI-Applikationen und KI-Unterstützung sind weit verbreitet, man findet sie heutzutage in vielen Bereichen, wie in der Medizin, der Produktion, der Lehre, dem Marketing oder auch im Finanz-Sektor. Die Anwendung und Umsetzung von KI-Applikationen sind sehr divers, man verwendet KI-Applikationen beispielsweise für Betrugserkennung (Brotcke, 2022) oder auch in der Entscheidungsfindung in verschiedenen Bereichen (Friedler et al, 2019).

Künstliche Intelligenzen und Machine Learning sind und werden immer mehr ein wichtiger Teil in der Entscheidungsfindung, bereichsübergreifend, jedoch sind künstliche Intelligenzen, Machine Learning und die aus ihnen folgenden Entscheidungen und Entscheidungsunterstützungen stark abhängig von den Trainingsdatensätzen, mit denen eben jene trainiert werden. Dementsprechend ist es von großer Relevanz für alle Beteiligten, dass innerhalb von Testdatensätzen keinerlei Verzerrungen (Bias) oder Diskriminierungen enthalten sind, da diese sich in dem Modell und den Ausgaben der künstlichen Intelligenzen wiederspiegeln können. Solche Verzerrungen können zu unfaireren Behandlungen und Diskriminierungen von Gruppen in unserer Gesellschaft führen. So könnte es passieren, dass Menschen auf Grund ihrer Ethnie, ihres Geschlechts oder ähnlichen Attributen schlechter von der künstlichen Intelligenz bewertet werden, auf Grund von solchen Verzerrungen, was in direkte Diskriminierungen mancher Personengruppen führen kann. Solche Verzerrungen und Verzerrungen allgemein zu reduzieren und abzubauen und Fairness zu schaffen ist dementsprechend ein großes Forschungsfeld geworden (Friedler et al, 2019).

Das Thema dieser Arbeit befindet sich genau in dem eben angesprochenen Forschungsfeld der Fair AI. Das Ziel und Thema der Arbeit ist, eine Methode in Form eines technischen Artefaktes zu entwerfen und entwickeln, welche Verzerrungen, wie die Beschriebenen, und Quellen potentieller Diskriminierung in Datensätzen ausgewählter Kontexte, also Kategorien, erkennt und aufzeigen kann. Dies alles soll sich auf tabellarische Datensätze beschränken, also Verzerrungen von Datensätzen für Natural-Language-Pattern-KIs, Chatbots oder der weiteren Abwandlungen werden nicht von der Methode abgedeckt. Eine detailliere Beschreibung des Artefakts ist im Versuchsplan nachzulesen.