Fragebogen "Datasheets for Datasets"

Annalena Aicher, Lennart Eing 8. Mai 2024

Die Ihnen hier gestellten Fragen dienen der Schaffung einer klaren Übersicht über einen zu erstellenden, bzw. erstellten Datensatz. Sie sind direkt [1] entnommen. Das Beantworten dieser Fragen dient sowohl der Schaffung eines klareren Überblicks seitens der Ersteller eines Datensatzes, als auch seitens möglicher Nutzer. Sie zielen auf bestimmte Teilbereiche und Prozesse innerhalb des Workflows der Datensatzerstellung. Diese Fragen sind weder vollständig, noch minimal. Das heißt, das anhängig vom Anwendungsfall möglicherweise Fragen nur teilweise, bzw. gar nicht beantwortet werden können. Trotzdem bitten wie Sie alle Fragen, soweit dies möglich ist, zu beantworten.

1 Motivation zur Erstellung des Datensatzes

Im Folgenden werden Ihnen Fragen zur Motivation zur Erstellung des Datensatzes gestellt. Bitte beantworten Sie möglichst alle Fragen.

1. Zu welchem Zweck wird der Datensatz erstellt? Gibt es eine spezifische Aufgabe die mithilfe dieser Daten gelöst werden soll? Gibt es eine Lücke in der Datenverfügbarkeit, die damit geschlossen werden soll? Bitte beschreiben sie:

2. Von wem wird der Datensatz erstellt? Im Auftrag von wem wird der Datensatz erstellt?

Im Folgenden werden Ihnen Fragen zum Inhalt des Datensatzes ge-

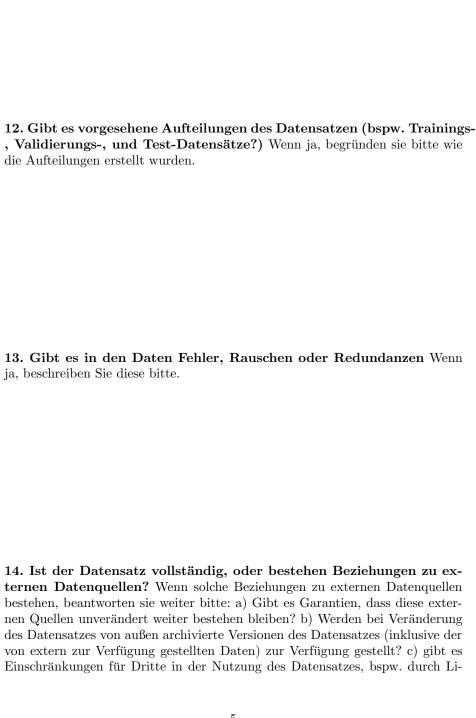
5. Wie setzen sich einzelne Proben des Datensatzes zusammen (bspw. Dokumente, Fotos, Menschen, Länder)? Gibt es verschiedene Probentypen (bspw. Filme und dazugehörige Bewertungen, Menschen und Interaktionen

stellt. Bitte beantworten Sie möglichst alle Fragen.

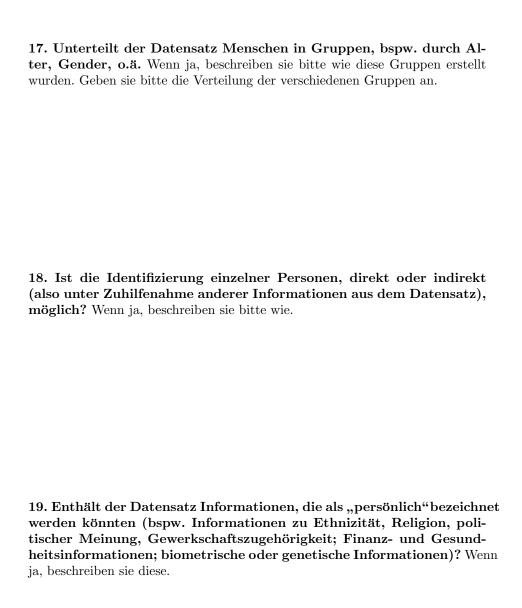
zwischen Ihnen, Kanten und Knoten?). Bitte beschreiben Sie:

6. Wie viele Proben sind insgesamt vorhanden? Wieviele Proben sind pro Probentyp vorhanden?	
7. Enthält der Datensatz vollumfänglich alle Probeninstanzen, oder	
ist er Teil eines größeren Datensatzes? Wenn der Datensatz Teil eines weiteren größeren Datensatzes ist, welcher ist dies? Ist der Datensatz repräsentativ für den größeren Datensatz. Wenn ja, wie wurde dies überprüft? Wenn nein, auf welche Art und Weise?	
8. Wie genau sieht eine eine einzelne Probe aus dem Datensatz aus? Beispiel: Werden Audio-Daten in Rohdaten-Formaten oder nur in vorm von vorberechneten Audio-Merkmalen abgespeichert?	
3	

9. Gibt es für jede Probe ein Trainings-Ziel, beziehungsweise zugehörige Label? Wenn ja, beschreiben sie bitte diese.
10. Gibt es Proben, für die Teile des Probenumfangs fehlen? Wenn ja, geben sie bitte eine Beschreibung warum diese Teil fehlen.
11. Sing Zusammenhänge zwischen verschiedenen Instanzen explizit? Wenn ja, wie werden diese Zusammenhänge dargestellt?



zenzen oder Einmalzahlungen, durch diese externen Quellen? Geben sie bitte alle externen Quellen an und beschreiben sie diese.
15. Enthält der Datensatz vertrauliche Informationen, bspw. aufgrund der ärztlichen Verschwiegenheitspflicht? Wenn ja, beschreiben sie diese bitte.
16. Enthält der Datensatz Dinge, die beim direkten Einsehen durch Menschen verstörend, beleidigend oder anderweitig belastend sein könnten? Wenn ja, beschreiben sie diese bitte.
Wenn der Datensatz keine Informationen enthält die sich direkt auf Personen bezieht können sie die folgenden Fragen überspringen.



20. Kommentare

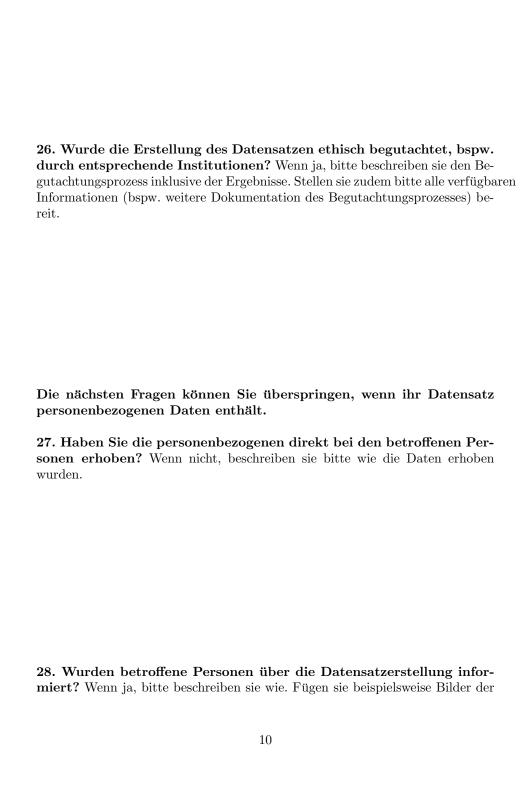
3 Erstellungsprozess des Datensatzes

Im Folgenden werden Ihnen Fragen zum Erstellungsprozess des Datensatzes gestellt. Bitte beantworten Sie möglichst alle Fragen.

21. Wie wurden einzelne Proben erstellt? Waren die Informationen direkt beobachtbar (bspw. Filme, Videos, etc.), wurden sie erfragt (bspw. mithilfe von Umfragen), oder indirekt erstellt (bspw. mithilfe von modellbasierter Altersschätzung)? Wenn Informationen indirekt erstellt wurden, wie wurde dies getan? Wie wurde sichergestellt, dass die erstellten Informationen korrekt sind?

22. Welche Methodik wurde verwendet um die Daten zu sammeln (bspw. Hardware-Spezifikationen, manuelles Erstellen von Menschen, Software, APIs)? Wie wurde die korrekte Funktionsweise der Methodik sichergestellt?

Die nächste Frage können Sie überspringen, wennn sie Frage 7 mit "Nein"beantwortet haben.
23. Wie haben sie Proben für den Datensatz aus dem größeren Datensatz ausgewählt?
24. Wer war am Sammeln der Daten beteiligt (bspw. Studenten, Vertragspartner, etc.) und wie wurden diese kompensiert?
25. In was für einem Zeitrahmen fand das Sammeln der Daten statt? Überschneidet sich der Zeitrahmen der Erstellung des Datensatzes auch mit der Erstellung einzelner Proben im Datensatz. Wenn nicht, beschreiben sie bitte auch den Zeitrahmen in welchem einzelne Proben erhoben wurden.



Mitteilung zur Datenerhebung an.
29. Haben die betroffenen Personen einer Sammlung ihrer Daten eingestimmt? wenn ja, beschreiben sie bitte wie diese Einstimmung eingeholt wurde.
30. Wenn Einstimmung eingeholt wurde, wird betroffenen Personen auch eine Möglichkeit zum Rückzug von ihrer Einstimmung gegeben? Dies gilt auch wenn der Rückzug nur für Teile der personenbezogenen Daten gilt. Beschreiben sie bitte wie ein Rückzug vollzogen werden kann.
31. Kommentare

4 Vorverarbeitung, Säubern und Labeling der Daten

Im Folgenden werden Ihnen Fragen zur Vorverarbeitung, Säuberung und dem Labeling der Daten gestellt. Diese dienen dem Zweck Nutzern des Datensatzes bessere Entscheidungsmöglichkeiten im Bezug auf die Eignung des Datensatzes für ihr Problem zu geben.

Literatur

[1] Timnit Gebru u. a. "Datasheets for Datasets". In: Communications of the ACM 64.12 (Dez. 2021), S. 86–92. ISSN: 0001-0782, 1557-7317. DOI: 10. 1145/3458723. (Besucht am 08.05.2024).