Fragebogen „Datasheets for Datasets“

Annalena Aicher, Lennart Eing

13. 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?

3. Wer hat die Erstellung des Datensatzes finanziert? Wenn die Finanzierung mithilfe Zuschüsse Dritter erstellt wurde nennen sie bitte diese.

4. Weitere Kommentare:

2 Inhalt des Datensatzes

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

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 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?

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?

12. Gibt es vorgesehene Aufteilungen des Datensatzen (bspw. Trainings-, Validierungs-, und Test-Datensätze?) Wenn ja, begründen sie bitte wie die Aufteilungen erstellt wurden.

13. Gibt es in den Daten Fehler, Rauschen oder Redundanzen Wenn ja, beschreiben Sie diese bitte.

14. Ist der Datensatz vollständig, oder bestehen Beziehungen zu externen Datenquellen? Wenn solche Beziehungen zu externen Datenquellen bestehen, beantworten sie weiter bitte: a) Gibt es Garantien, dass diese externen Quellen unverändert weiter bestehen bleiben? b) Werden bei Veränderung des Datensatzes von außen archivierte Versionen des Datensatzes (inklusive der von extern zur Verfügung gestellten Daten) zur Verfügung gestellt? c) gibt es Einschränkungen für Dritte in der Nutzung des Datensatzes, bspw. durch Li-

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.

17. Unterteilt der Datensatz Menschen in Gruppen, bspw. durch Alter, Gender, o.ä. Wenn ja, beschreiben sie bitte wie diese Gruppen erstellt wurden. Geben sie bitte die Verteilung der verschiedenen Gruppen an.

18. Ist die Identifizierung einzelner Personen, direkt oder indirekt (also unter Zuhilfenahme anderer Informationen aus dem Datensatz), möglich? Wenn ja, beschreiben sie bitte wie.

19. Enthält der Datensatz Informationen, die als „persönlich“bezeichnet werden könnten (bspw. Informationen zu Ethnizität, Religion, politischer Meinung, Gewerkschaftszugehörigkeit; Finanz- und Gesundheitsinformationen; biometrische oder genetische Informationen)? Wenn ja, beschreiben sie diese.

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.

26. Wurde die Erstellung des Datensatzen ethisch begutachtet, bspw. durch entsprechende Institutionen? Wenn ja, bitte beschreiben sie den Begutachtungsprozess inklusive der Ergebnisse. Stellen sie zudem bitte alle verfügbaren Informationen (bspw. weitere Dokumentation des Begutachtungsprozesses) bereit.

Die nächsten Fragen können Sie überspringen, wenn ihr Datensatz personenbezogenen Daten enthält.

27. Haben Sie die personenbezogenen direkt bei den betroffenen Personen erhoben? Wenn nicht, beschreiben sie bitte wie die Daten erhoben wurden.

28. Wurden betroffene Personen über die Datensatzerstellung informiert? Wenn ja, bitte beschreiben sie wie. Fügen sie beispielsweise Bilder der

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.

32. Wurde irgendwelche Vorverarbeitung, Säuberung oder Labeling der Daten vorgenommen, bspw. Diskretisierung, Part-of-Speech Tagging, SIFT Merkmal Extraktion, das Entfernen einzelner Instanzen oder automatisiertes Berechnen fehlender Werte? Wenn ja, beschreiben sie bitte ausführlich. Wenn nein, dann können sie die restlichen Fragen in diesem Abschnitt überspringen.

33. Wurden die Rohdaten zusätzlich zu den vorverarbeiteten Daten gespeichert, bspw. für zukünftig, nicht zu erwartende Zwecke? Wenn ja, geben sie bitte einen Weg an, über den die Rohdaten erhältlich sind.

34. Ist die Software, die verwendet wurde um die Rohdaten vorzuverarbeiten für beliebige Endnutzer erhältlich? Wenn ja, geben sie bitte einen Weg an diese Software zu erhalten.

35. Kommentare

5 Nutzung

Die folgenden Fragen richten sich maßgeblich an die Ersteller des Datensatzes. Sie sollen dazu anleiten sich Gedanken über die Aufgaben zu machen zu welchem der Datensatz verwendet, und auch nicht verwendet werden soll. Dies soll Nutzer des Datensatzes eine Unterstützung bei der Entscheidungsfindung bieten und dabei helfen potientielle Risiken zu vermeiden.

36. Wurde der Datensatz bereits für irgendwelche Zwecke verwendet?

37. Gibt es einen Überblick, der alle oder einem Teil der verfügbaren Papiere und Systeme an einem Ort zusammenfasst. Wenn ja, geben sie bitte einen Weg an, über welchen dieser Überblick erhältlich ist.

38. Für welche (anderen) Zwecke kann der Datensatz genutzt werden?

39. Könnten Teile dessen, wie der Datensatz zusammengestellt, gesammelt und vorverarbeitet wurde zu einer Verzerrung der Realität bei einer Nutzung führen? Beispielsweise können sie hier angeben, wenn es

durch die Nutzung des Datensatzes zur unfairen Behandlung von Individuen oder Gruppen, oder rechtlichen oder finanziellen Risiken kommen kann. Wenn ja, beschreiben sie bitte. Gibt es Vorkehrungen, die ein Nutzer des Datensatzes treffen kann um diese Risiken zu vermeiden?

40. Gibt es Zwecke, für welche der Datensatz nicht benutzt werden sollte? Wenn ja, beschreiben Sie bitte.

41. Kommentare

6 Verbreitung

Die folgenden Fragen richten sich maßgeblich an die Ersteller des Datensatzes. Sie sollten beantwortet werden, bevor der Datensatz in-

nerhalb der erstellenden Entität oder an Dritte weiterverbreitet wird.

42. Wird der Datensatz auch an Dritte außerhalb der erstellenden Entität weiterverbreitet werden? Wenn ja, beschreiben Sie bitte.

43. Wie wird der Datensatz verbreitet werden, bspw. tarball, API oder GitHub? Hat der Datensatz außerdem einen Digital Object Identifier (DOI)?

44. Ab wann wird der Datensatz weiterverbreitet werden?

45. Wird der Datensatz unter einen Copyright verbreitet? Werden die Rechte am geistigen Eigentum einbehalten? Wird der Datensatz unter einer Nutzungsvereinbahrung verbreitet? Wenn ja, beschreiben

Sie bitte die Lizenzen und/oder Nutzungsvereinbahrungen und geben sie einen Weg an diese einzusehen und/oder zu erhalten? Geben Sie außerdem auch alle Gebühren an, die potentiell aufgewendet werden müssen.

46. Schränken Dritte aufgrund geistigen Eigentums die Verbreitung des Datensatzes ein? Wenn ja, beschreiben Sie bitte die Einschränkungen. Geben Sie einen Weg an diese einzusehen. Geben sie außerdem alle Gebühren an, die aufgrund dieser Einschränkungen aufgewendet werden müssen.

47. Gibt es Exportbeschränkungen oder andere regulatorische Einschränkungen die den ganzen Datensatz oder Einzelinstanzen des Datensatzes betreffen? Wenn ja, schreiben Sie diese bitte und geben sie einen Weg an diese einzusehen.

48. Kommentare

7 Wartung

Die folgenden Fragen richten sich maßgeblich an die Ersteller des Datensatzes. Sie sollten beantwortet werden, bevor der Datensatz innerhalb der erstellenden Enität oder an Dritte weiterverbreitet wird. Sie zielen maßgelbich darauf ab, die Ersteller des Datensatzes anzuregen einen Wartungsplan aufzustellen und diesen den Nutzern des Datensatzes zu kommunizieren.

49. Wer wird die den Datensatz warten, zu Verfügung stellen, und Support übernehmen?

50. Wie kann Kontakt zu den Kuratoren des Datensatzes hergestellt werden?

51. Gibt es ein Erratum? Wenn ja, geben Sie bitte einen Weg an, über den es erhältlich ist.

52. Wird der Datensatz Updates erhalten, bspw. um Labelingfehler zu korrigieren? Wenn ja, beschreiben Sie bitte, wie oft diese Korrekturen stattfinden werden, durch wen diese Korrekturen durchgeführt werden und wie die Verfügbarkeit von Korrekturen an die Datensatznutzer kommuniziert werden wird.

53. Diese Frage kann übersprungen werden, wenn der Datensatz keine personenbezogenen Daten enthält: Gibt es zeitliche Begrenzungen wie lang personenbezogene Daten gehalten werden dürfen, bspw. weil den Personen mitgeteilt wurde, dass ihre Daten nur für einen bestimmten Zeitrahmen gehalten werden würden? Wenn ja, beschreiben

Sie diese. Beschreiben Sie weiter wie die Wahrung dieser zeitlichen Begrenzungen gewährleistet werden soll.

54. Werden alte Versionen des Datensatz weiterhin verfügbar sein und gewartet werden? Wenn ja, beschreiben Sie bitte wie. Wenn nicht, dann beschreiben Sie bitte, wie Obsolenszenz den Datensatznutzern mitgeteilt wird.

55. Ist es Dritten möglich zum Datensatz beitragen/beizufügen? Wenn ja, beschreiben Sie bitte wie. Wenn nicht, warum nicht? Wie werden Datensatznutzer über solche Beifügungen informiert?

56. Kommentare

Literatur

[1] Timnit Gebru u. a. „Datasheets for Datasets“. In: *Communications of the ACM* 64.12 (Dez. 2021), S. 86–92. DOI: 10.1145/3458723.