SMS Messages Dataset

Dataset Card

Version 1.1

Updated: 6th February 2024

Contents

How to use this document

Purpose

This document contains information about the SMS Messages dataset, to help users understand what the dataset contains, its intended use cases, how it was developed, associated metadata, and information on how to use the data responsibly.

The Dataset Card consists of four main parts: an overview of the dataset, dataset structure, data splits, and information about the curation methodology. This includes information on privacy, ethics, and any known biases and limitations of the dataset.

This document is intended for use primarily by the data science community.

Navigation

Use the hyperlinked tabs along the left-hand side to navigate through the document, and between pages and sections.

Dataset description

The **SMS Messages** for Authorship Attribution dataset is a sub-set of the Dialogue Summarisation dataset, with the addition of author labels.

It is a collection of text messages generated by over 1,000 authors. It has been designed to train and test authorship attribution models, that take a text input and predict the likely author of the text.

The messages in the dataset are all in English. A small number of messages were sent with media attachments, such as gifs or images, and these messages are marked with a relevant label.

The dataset contains over 1,000 authors in the training set, with a low number of messages per author (median = 39), and a minimum of 12 messages per author. There are also 'unknown' authors within the training data.

The messages aim to be realistic representations of natural text conversations generated by over 1,000 unique authors from a range of demographics (to maximize diversity and variance in the conversation styles and content) who were either friends or family. They cover a range of everyday topics, where some participants were given example topics to cover but others free reign on what to talk about. The messages are varied in length and therefore provide a range of complexity for the attribution task, but on average are very short at 11 words long.

Dataset structure

Data format

There is a single CSV file for this dataset which contains both SMS messages with the relevant author ID label.

A sample of rows from the training data is shown below:

message	author	attachment
And let's not forget the power of representation.	812	NaN
And everybody started imitating him{U+1f604}	590	NaN

Data fields within the dataset

Column name	Data type	Description
message	string	Text content of the message sent
author	int	ID of the person who sent the message
attachment	string	The file name of an attachment (i.e., gif or photo) accompanying a given message. If the message has no attachment, then this shows "nan"

Curation methodology

Curation rationale

Authorship attribution aims to determine which author, out of a number of already studied authors, an anonymous text belongs to. It aims to identify the author of a piece of content by considering the stylometric features present within the text. This technique is used in the wider field of digital forensic text analysis, policing, law, government intelligence, social network sites and academic research.

Identifying an author would traditionally require expertise in both the author (e.g., their habits and preferred topics of conversation) and their typical style of writing (e.g., common spelling mistakes, use of emojis, grammatical inconsistencies).

The length of text, total number of authors, and number of contributions per author are thought to have a significant impact on model performance. Messenger conversations, which are characterised by their short length of prose, unstructured text, frequent topic shifts, colloquial language, and grammatical errors, make the task of authorship attirbution much harder. Previous datasets in this field have tended to have both a smaller number of authors, as well as a larger number of contributions per author.

This dataset was curated via a crowdsourcing methodology to represent natural messaging as closely as possbile. To optimise the naturalness and diversity of the conversational data, as well as to have a large author pool, this dataset includes a high number of unique participants. Conversation partners all had a high level of familiarity to each other (i.e. either a friend or family members) to facilitate more natural conversations.

Curation approach

For the generation of the messages, the primary participant (speaker 1) was instructed to have a conversation with their chosen conversation partner (speaker 2).

They were required to think of a purpose for contacting their conversation partner (an example list of possible intentions was provided), the duration that the dialogue had to last and example topics that they could cover.

The number of topics participants had to try and cover was either pre-specified or participants could choose, this was done to create variety, and therefore different degrees of complexity in the summarisation task, in the amount of topics covered (i.e., with some conversations bouncing around up to 10 different topics, and others covering just a single topics).

The exact number of topics that participants were instructed to cover was not independently validated (as it was found to impact the quality of the conversations to restrict participants to such requirements), but rather this guidance was provided to naturally induce variation between conversation styles.

Participants were given freedom to deviate to other topics if the conversation went in that direction. Topic examples included:

- World Affairs
- Work Life
- UK Politics
- Travel
- Sport
- Past Events
- Music
- Hobbies
- Finance / economics
- Complaints / gossip
- Climate change
- Arts and Entertainment
- Unspecified - a topic of the participants choosing

Participant information

All participants were native or fluent English speakers, and were based mostly in the Philippines (39%), India (31%) and Great Britain (30%).

The participants included a distribution of genders and age groups, with no specific demographic making up a significant proportion of the dataset (see Biases section below for more details).

- The conversations consisted of 52% male participants and 48% female participants
- Each age group had a participation proportion below 30%

Data privacy and ethics

All participants gave informed consent to participate in this data collection, with the knowledge that the conversational data could be released publicly and used for training and developing data science models.

All participants were assigned an ID number, which is not traceable back to any individual.

All participants were paid the living wage for their region for participation in the data curation.

During data collection participants were instructed that they should **not** share any Personally Identifiable Information. This was validated through quality control checks on the conversational data. These are some examples of PII shown to participants for them to avoid disclosing:

- Full names
- Credit/bank information
- Identity documentation (ID Numbers for example Social Security, account numbers or other citizenship numbers)
- Personal addresses
- Contact information
- Medical information
- Information about education (such as schools attended) or profession (such qualifications, place of employment)
- Information about racial or ethnic origin, political opinions, religious or philosophical beliefs, trade union data
- Data concerning sexual orientation
- Data concerning criminal allegations, proceedings or convictions.

Participants were also instructed not to talk about anything deemed offensive. This was validated through quality control checks on the conversational data, with examples given being:

- Defamatory content
- Violence or inciting violence
- Hate speech
- Indecent images

PII disclaimer
Disclaimer: We have taken best endeavours to remove all Personally Identifiable Information (PII) from this dataset. If you identify any conversations with content that could be deemed as PII please contact challengeaiuk@deloitte.co.uk.

Note, there may be small amounts of low-risk personal data within the dataset (for example: the full name of a public figure).

Biases

To produce representative and limit bias in the data, participant demographics were tracked at an aggregated level to validate that no age groups, genders or regions made up a significant proportion of the dataset. The breakdown of these statistics can be found in Appendix A.

Bias notification
Disclaimer: We have taken best endeavours to design out any biases from this dataset.

Other known limitations

Given how participant ID numbers were generated, it cannot be guaranteed that authors with a '_' or '-' in their id are actually unique individuals.

Licensing Information

PLEASE READ THE TERMS BELOW CAREFULLY BEFORE USING THE DATASET

What's in these terms?

This data licence ("**Data Licence**") applies to

1. any and all Data (as defined below) that you download from the Site (as defined below) and
2. any content or materials you create that are based on, or contain (in whole or part), any such Data.

We own, or license the use of, the Intellectual Property Rights (as defined below) in the Data and agree to provide the Data to you on the terms set out in this Data Licence.

1. Who we are and how to contact us

1.1 The challengeai.com website (the "**Site**") is operated by the Ministry of Defence. Our main address is Whitehall, London SW1A 2HB.

1.2 To contact us, please use the Get in contact button at the bottom of this website.

2. By downloading the Data you accept these terms

2.1 By downloading Data, you confirm (on behalf of your organisation) that you accept the terms of this Data Licence and that you agree to comply with them. You also confirm that you have the authority from your organisation to accept this Data Licence.

2.2 If you do not agree to these terms, you must not download any Data.

2.3 We recommend that you print a copy of these terms for future reference.

3. Interpretation

3.1 The definitions and rules of interpretation in this clause apply in this Data Licence.

- **Challenge**: any Challenge listed on the Site in the 'Challenges' section, which can be found here.
- **Computer System**: any information technology hardware, software and/or system owned, controlled, or operated by you/your organisation to which Data is received in accordance with this Agreement.
- **Data**: the datasets and associated information, documentation and materials which relate to any Challenge and are made available for download by users of the Site in order to participate in any such Challenge.
- **Derived Data**: any Data (wholly or in part) which is Manipulated to such a degree that it: (a) cannot be identified as originating or deriving directly from the Data and cannot be reverse-engineered such that any part of the underlying Data can be identified; and (b) is not capable of use substantially as a substitute for the Data.
- **Distribute**: to make Data accessible including the provision of access through a database or other application populated with the Data, resell, sub-license, transfer, disseminate, or otherwise disclose or provide a copy of the Data by any means in any format whether electronic or hard copy.
- **Good Industry Practice**: in relation to any undertaking and any circumstances, the exercise of skill, diligence, prudence, foresight and judgement and the making of any expenditure that would reasonably be expected from a skilled person engaged in the same type of undertaking under the same or similar circumstances.
- **Intellectual Property Rights**: all patents, rights to inventions, copyright and related rights, trade marks, service marks, trade, business and domain names, rights in trade dress or get-up, rights in goodwill or to sue for passing off, unfair competition rights, rights in designs, rights in computer software, database rights and other rights in data, moral rights, rights in confidential information (including know-how and trade secrets) and any other intellectual property rights, in each case whether registered or unregistered and including all applications for and renewals or extensions of such rights, and all similar or equivalent rights or forms of protection in any part of the world.
- **Licence**: the licence granted in clause 6.
- **Manipulate**: to combine or aggregate the Data (wholly or in part) with other data or information, or to adapt or change the Data (wholly or in part).
- **Manipulated Data**: any Data which has been Manipulated.

- **Permitted Purpose**: the purpose in respect of which you are permitted to use the Data, as set out in clause 6.1.

3.2 A **person** includes a natural person, corporate or unincorporated body (whether or not having separate legal personality).

3.3 A reference to **writing** or **written** includes email, but not fax.

3.4 Any words following the terms **including**, **include**, **in particular** or **for example** or any similar phrase shall be construed as illustrative and shall not limit the generality of the related general words.

4. There are other terms that may apply to you

4.1 This Data Licence applies to your use of Data, however the following additional terms (Additional Terms), also apply to your use of our Site generally:

- Our Website Terms and Conditions (Website Terms) that apply to your use of our Site.
- Our Privacy Policy. See further under How we may use your personal information.
- Our Acceptable Use Policy, which sets out the permitted uses and prohibited uses of our Site. When using our Site, you must comply with this Acceptable Use Policy.
- Our Challenge Terms on each relevant Challenge Rules page, which sets out the rules and requirements of participation in any Challenge.

5. We may make changes to this Data Licence

5.1 We may amend this Data Licence from time to time. Every time you wish to download Data from this Site, please check these terms to ensure you understand the terms that apply at that time. These terms were most recently updated on 6th February 2024.

6. Licence

6.1 Subject always to your compliance with the terms of this Data Licence, we grant to your organisation a non-exclusive, non-transferable, worldwide, royalty-free, perpetual but revocable licence to:

(a) access, view, use, Manipulate the Data and create Derived Data; and

(b) store copies of the Data and Manipulated Data on your Computer Systems,

solely for: (i) the purposes of your participation in any Challenge; (ii) the purpose of demonstrating a solution or capability which does not form part of a Challenge ; and/or (iii) your organisation's own internal business purposes, the limitations on which are set out below ("**Permitted Purpose**"). For the avoidance of doubt 'internal business purposes' means the usage of the Data solely by you, only on Computer Systems controlled by you/your organisation, in relation to the internal business operations of your organisation (including internal research and development, and internal training). Please note that this is not a personal licence granted to you as an individual.

6.2 Except as expressly provided in this Agreement, you shall (personally and on behalf of your organisation):

(a) limit access to the Data to those individuals who need to access it in order to carry out the Permitted Purpose;

(b) only make copies of the Data to the extent reasonably necessary for the Permitted Purpose, including where necessary, disaster recovery and testing;

(c) not use the Data in your personal capacity, outside the performance of your duties for your organisation.

(d) not use the Data for any purpose contrary to any law or regulation or any regulatory code, guidance or request;

(e) not extract, reutilise, use, exploit, Distribute, modify, copy, publish or store the Data (in whole or in part), for any purpose not expressly permitted by this Agreement. In particular you will not Distribute any of the Data to the public, third-parties and/or any person outside of your company or organisation;

(f) not decompile, reverse engineer or create derivative works from the Data for purposes other than the Permitted Purpose;

(g) not use the Data (wholly or in part) in your products or services that are made available to the public;

(h) not use the Data in a manner that allows or facilitates any Data to fall into the public domain;

(i) you shall not, nor directly or indirectly assist any other person to, do or omit to do anything to diminish our rights in the Data.

7. Data and Materials

7.1 We shall supply the Data on the Site.

7.2 You acknowledge that the Data is confidential, and you shall not disclose it, in whole or in part, to any third party, except as expressly permitted by this Data Licence. This obligation shall continue to apply after termination of this Data Licence.

7.3 At any time, we may, with as much prior notice to you as is reasonably practicable, change: (a) the content, format or nature of the Data; and (b) the means of access to the Data.

8. Security and passwords

8.1 You shall ensure that Data is kept securely and shall implement security practices and systems in accordance with Good Industry Practice to prevent, and take prompt and proper remedial action against, unauthorised access, copying, modification, storage, reproduction, display or Distribution of the Data.

8.2 If you become aware of any misuse of any Data, or any security breach (i.e. any incident that results in unauthorised access to computer data, applications/software, networks or devices) in connection with this Data Licence that could compromise the security or integrity of the Data or otherwise adversely affect us: (a) you shall, promptly notify us and fully co-operate with us to remedy the issue as soon as reasonably practicable; and (b) you agree to co-operate with our reasonable security investigations.

9. Export

You shall not export, directly or indirectly, any of the Data (or any products, documents or materials, including software, incorporating any such Data) to any location (whether physical or digital) that is not within your direct control.

10. Intellectual Property Rights ownership

10.1 You acknowledge that we are the owner or the licensee of all Intellectual Property Rights, in and to the Data.

10.2 Subject to our rights in the Data and the terms of this Data Licence (which shall at all times prevail), your organisation shall own the Intellectual Property Rights to the Derived Data. Your use of the Derived Data shall at all times be subject to the Licence.

10.3 We draw your attention again to the Website Terms and Conditions and Challenge Terms which contain obligations that you have agreed to in the event that your entry is a Winning Challenge Response (as defined in the Challenge Terms). In particular, you have agreed that your organisation will automatically grant a 12-month licence to us of the Intellectual Property Rights in any Winning Challenge Response submitted by you (including all Derived Data and underlying documents). For further details see the Website Terms and Conditions and the Challenge Terms.

11. Warranties

11.1 We warrant that we have the right to license the receipt and use of Data.

11.2 Except as expressly stated in this Data Licence, all warranties, conditions and terms, whether express or implied by statute, common law or otherwise are hereby excluded to the fullest extent permitted by law.

11.3 We will take all reasonable steps, in accordance with Good Industry Practice not to introduce any vulnerability or virus into your Computer Systems, whether via the Site or Data or otherwise.

11.4 Without limiting the effect of this clause 11, we do not warrant that: (a) the supply of the Data will be free from interruption; (b) the Data will be suitable for your Computer Systems; or (c) the Data is accurate, complete, reliable, secure, useful, or fit for purpose.

12. Limitation of liability

12.1 Neither party excludes or limits liability to the other party for: (a) fraud or fraudulent misrepresentation; or (b) any matter in respect of which it would be unlawful for the parties to exclude liability.

12.2 Subject to clause 12.1, we shall not in any circumstances be liable whether in contract, tort (including for negligence and breach of statutory duty howsoever arising), misrepresentation (whether innocent or negligent), restitution or otherwise, for any of the following losses suffered by you as a result of your use of the Data: (a) any loss (whether direct or indirect) of profits, business, business opportunities, revenue, turnover, reputation or goodwill; (b) any loss or corruption (whether direct or indirect) of data or information; (c) loss (whether direct or indirect) of anticipated savings or wasted expenditure (including management time); or (d) any loss or liability (whether direct or indirect) under or in relation to any other contract.

13. Termination

13.1 Without prejudice to any rights that have accrued under this Data Licence or any of our rights or remedies, we may terminate this Data Licence with immediate effect by giving written notice to you if you commit a material breach of any term and fail to remedy that breach within a period of [30] days after being notified in writing to do so.

13.2 Any provision of this Data Licence that expressly or by implication is intended to come into or continue in force on or after termination shall remain in full force and effect.

13.3 On any termination of this Data Licence for any reason, you shall as soon as reasonably practicable, delete or destroy (as directed in writing by us) all copies (in whole or in part) of the Data and Manipulated Data whether stored in hard copy or electronically on your Computer Systems or otherwise. You will also delete the Data and Manipulated Data from any database, software program, documentation and other materials which contain (in whole or in part) copies of the Data and/or Manipulated Data.

14. Announcements

You shall not make, or permit any person to make, any public announcement concerning this Data Licence, any Challenges or Data without our prior written consent, except as required by law, any governmental or regulatory authority

(including any relevant securities exchange), any court or other authority of competent jurisdiction.

15. Assignment

This Data Licence is personal to you, and you shall not assign, transfer, mortgage, charge, sub-contract, sub-license (except as permitted above) declare a trust of or deal in any other manner with any of your rights and obligations under this Data Licence without our prior written consent.

16. Waiver

16.1 A waiver of any right or remedy is only effective if given in writing and shall not be deemed a waiver of any subsequent right or remedy.

16.2 A delay or failure to exercise, or the single or partial exercise of, any right or remedy shall not waive that or any other right or remedy, nor shall it prevent or restrict the further exercise of that or any other right or remedy.

17. Remedies

Except as expressly provided in this Data Licence, the rights and remedies provided herein are in addition to, and not exclusive of, any rights or remedies provided by law.

18. Inadequacy of damages

Without prejudice to any other rights or remedies that we may have, you acknowledge and agree that damages alone would not be an adequate remedy for any breach by you of the terms of this Data Licence. Accordingly, we shall be entitled to seek the remedies of injunction, specific performance or other equitable relief for any threatened or actual breach of the terms of this Data Licence.

19. Severance

If any provision or part-provision of this Data Licence is or becomes invalid, illegal or unenforceable, it shall be deemed deleted, but that shall not affect the validity and enforceability of the rest of this Data Licence.

20. No partnership or agency

20.1 Nothing in this Data Licence is intended to, or shall be deemed to, establish any partnership or joint venture between any of the parties, constitute any party the agent of another party, or authorise any party to make or enter into any commitments for or on behalf of any other party.

20.2 You confirm that you are acting on your own behalf and not for the benefit of any other person.

21. Third-party rights

This Data Licence does not give rise to any rights under the Contracts (Rights of Third Parties) Act 1999 to enforce any term of this Data Licence.

22. Governing law

This Data Licence and any dispute or claim arising out of or in connection with it or its subject matter or formation (including non-contractual disputes or claims) shall be governed by and construed in accordance with the law of England and Wales.

23. Jurisdiction

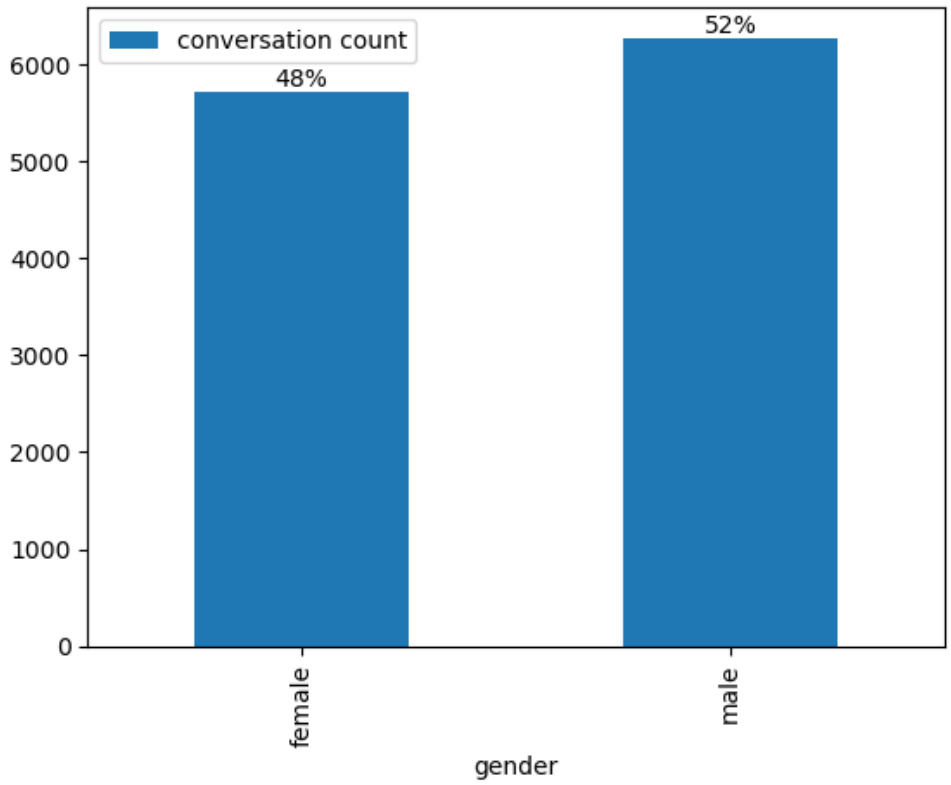
Each party irrevocably agrees that the courts of England and Wales shall have exclusive jurisdiction to settle any dispute or claim arising out of or in connection with this Data Licence or its subject matter or formation (including non-contractual disputes or claims).

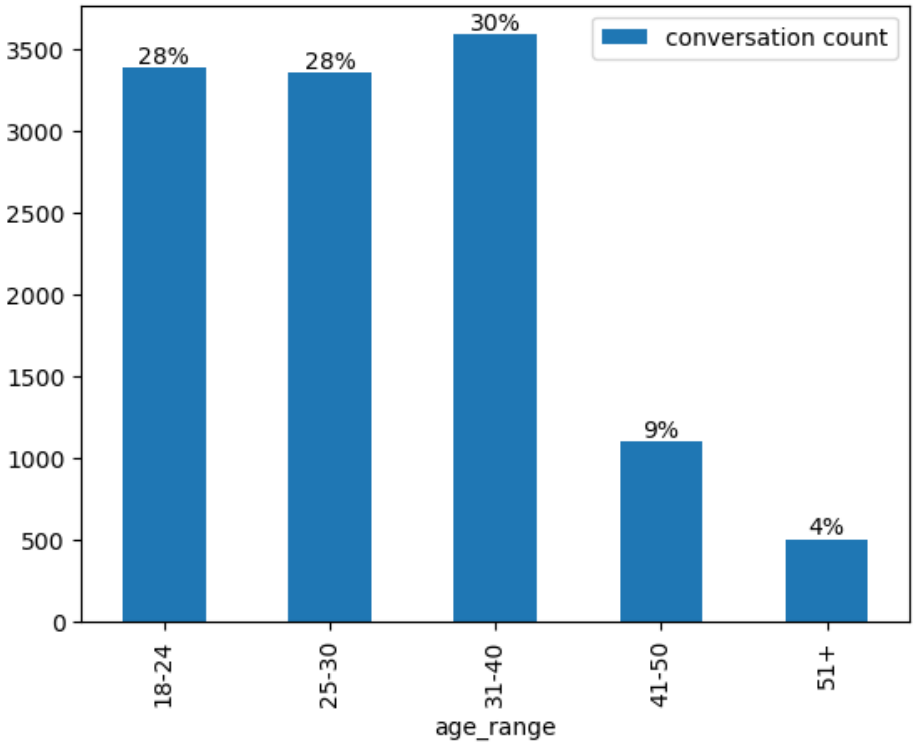
Additional information

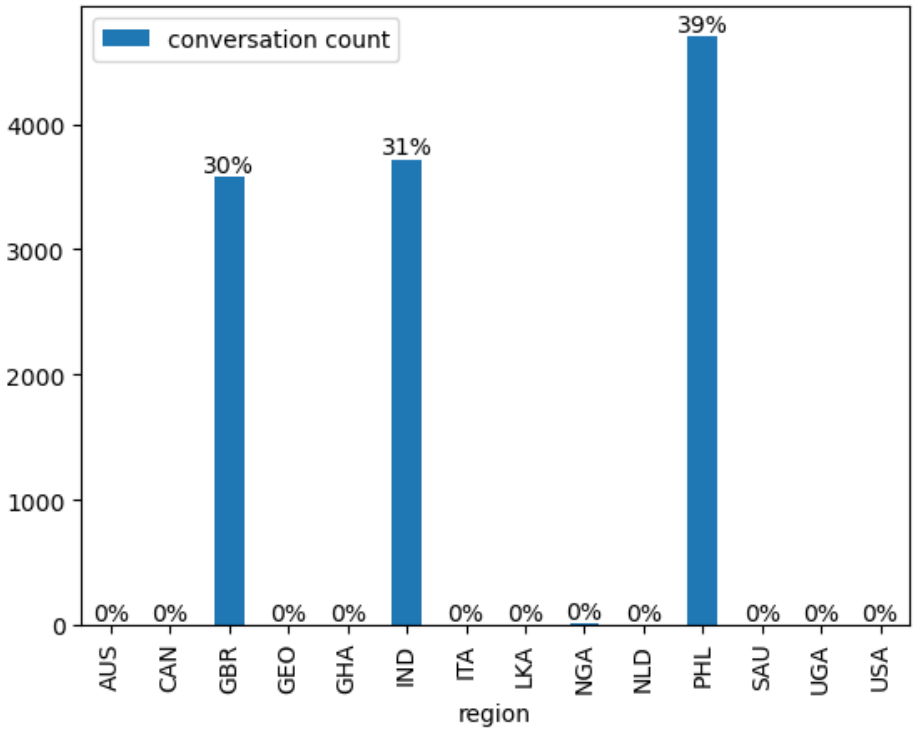
Appendix

A. Demographic descriptive statistics

The plots below show the demographic breakdown for participants (primary and secondary) across the data collection:




conversation count
28%
28%
30%
9%
4%
0
500
1000
1500
2000
2500
3000
3500
18-24
25-30
31-40
41-50
51+
age_range


conversation count
0%
0%
30%
0%
0%
31%
0%
0%
0%
0%
39%
0%
0%
0%
0
1000
2000
3000
4000
AUS
CAN
GBR
GEO
GHA
IND
ITA
LKA
NGA
NLD
PHL
SAU
UGA
USA
region