Data in Brief 38 (2021) 107329

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

Semi-literate Texting (SLT): Survey based text message dataset from digitally semi-literate users in India

Prawaal Sharma[a,*], Navneet Goyal[b,*], Vinay MR[c,*]

[a] Infosys, Pune, India
[b] BITS Pilani, Pilani, Rajasthan, India
[c] Infosys, Bangalore, India

ARTICLE INFO

Article history:
Received 11 May 2021
Revised 12 August 2021
Accepted 24 August 2021
Available online 26 August 2021

Dataset link: digitally semi-literate text message dataset (Original data)

Keywords:
Text messages
Texting
Digitally Semi-literate
Emergent mobile phone users

ABSTRACT

The dataset explicates text messages and associated metadata from digitally semi-literate mobile phone users in India. A survey among urban and rural representatives conducted between July 2020 and November 2020 is the origin for this dataset. The data has been collected through face to face interviews and online surveys across urban and rural geographies in India, largely from western region of Maharashtra. A total of 382 respondents, accumulating 3368 messages has been composed (approximately 90% through face to face surveys and 10% from online mode). To the best of our knowledge there is no factual text message data from digitally semi-literate users being available till date. This dataset can be used for bridging the digital divide in human computer interaction using machine learning, data mining, behavioural analysis as well as in other fields.

* Corresponding author.
E-mail addresses: Prawaal_sharma@infosys.com (P. Sharma), goel@pilani.bits-pilani.ac.in (N. Goyal), vinay.mr@infosys.com (V. MR).

https://doi.org/10.1016/j.dib.2021.107329

Specifications Table

Subject	Human-Computer Interaction
Specific subject area	Text messages and other metadata to understand the text message patterns of digitally semi-literate users in Indian context.
Type of data	Table Figure
How data were acquired	Field Survey (90%), Online Survey (10%)
Data format	The raw datasets are in Microsoft Excel format and is available on the Mendeley data repository.
Parameters for data collection	• All the respondents should qualify to be called as digitally semi-literate (not very conversant on usage of technology). • Collection of forward messages, motivation messages and generic messages where there is no exchange of information to be discouraged. • Only one-to-one message have been considered. No group conversations recorded. • All Personally identifiable information (PII) data is anonymized. • Face to face collection is limited to 10 messages per participant, however for online surveys only 5 messages has been considered due to short attention span while filling up online forms.
Description of data collection	Professional market research agencies have been employed to conducted face-to-face survey along with online surveys.
Data source location	Data collected from urban and rural regions of Maharashtra, India via face to face surveys (90%) and online mode (10%).
Data accessibility	Repository name: Mendeley Data Data identification number: 10.17632/4b53nj78tv.8 Direct URL to data: https://data.mendeley.com/datasets/4b53nj78tv/8

Value of the Data

- Digitally semi-literate emergent mobile phone users in India are customarily not on social media. These users communicate via text messages for information exchange frequently. The dataset is valuable as it records the in-use vocabulary by these users.
- India is aspiring for digitization to empower people. One of the pillars of “Digital India” campaign is information for all. Our dataset can be used to design content for digitally semi-literate users for various government initiatives.
- The dataset can be used on investigation of social behaviour patterns of digitally semi-literates via text messages analysis.
- The dataset can contribute to further research on potential strategies for the facilitation of text messaging on mobile phones, in the field of human computer interaction.
- Digitally semi-literate users make a huge section of world population. Any step towards their digital enablement will increase their quality of life.

1. Data Description

Technology and innovation have transformed the ecosystem of information access and exchange. However, in India digital deprivation has been an ongoing issue specially for urban and rural poor sections. The primary factors to bridge the gap and facilitate digital enablement in-

cludes (1) Low cost hardware and internet access (2) Intuitive techniques for Human Computer Interaction (HCI) and (3) Building trust and security for use of technology.

Easy availability along with reasonably priced mobile phones and internet access in India has helped in penetration of this technology at grassroot level. However, most of the information on digital platforms which can possibly help emergent mobile phone users, use an enriched vocabulary and is beyond their comprehensibility. Therefore, the most appropriate use of technology for these users is to exchange information via text messages across their community to share and participate in issues appropriate for them.

Most mobile phones enable asynchronous text communication. Most users are conversant with local language and use the same for text communication. Sometimes they use roman script for communicating in their local language which creates challenges in sending and comprehending information. Any technological progress in ease of text messaging on mobile phones, can help in improving their digital adoption and bridge the digital divide.

There is no clear definition of digital literacy in the research world. In late 1900s Gilster [1] was the first one to coin the term 'digital literacy'. He measured this on parameters of education and information skills. Over a period, the definition of 'digital literacy' has modified. Chase and Laufenberg [2] have referred this to be 'inherently squishy'. The current definition ranges from being technology fluent to being able to use information on digital platforms without assistance.

In India, the government is running a flagship program of "Digital India". One of the key pillars of this program is "Information for all" [3]. In a recent study conducted by 'Ministry of Women and Child Development, Government of India' along with 'University of Delhi' it was observed that one of the key factors for being digitally illiterate is lack of tertiary education [4]. In another study a similar observation was made, that the poor literacy rate in India, particularly women and rural population is major impediment to the growth of digital literacy. "It is hard to think of universal digital literacy without universal literacy" [5]. Therefore it is evident from related literature that academic literacy is strongly co-related with digital literacy.

Mostly users are classified as digitally literate or illiterate. We believe that there is another classification of users, who are digitally semi-literate. These users have access to mobile phones and internet and have elementary education. However, they are not very conversant on usage of technology and don't have enriched English vocabulary which is primary language of internet. We have used the term "semi-literate" to refer to these individulas who have elementary education but face challenges with digital enablement in our paper.

Priya et al. [6] conducted a systematic literature review on usage of mobile phones by semi-literates and non-literates. Kntongo and Morakanyane [7] also conducted a similar study in Botswana to help semi-literates with digital inclusion and awareness in rural areas. They had conducted a survey in rural Botswana with 127 participants and collected data with their mobile usage and problems faced. However, their data did not contain text message data for analysis purpose.

We have collected actual text messages exchanged by digitally semi-literate users in our dataset. Our dataset can be used for designing interfaces for human computer interaction. It can also help understand the communication behaviour pattern of digitally semi-literate users. Organisations working towards their social upliftment can also derive insights from the messaging text. Nevertheless, due to the characteristics of the variables this can be used in variety of other functional uses, beyond the scope mentioned.

The datasets presented are the Microsoft Excel Workbook with two(2) sheets as below:

Sheet 1 (Summary): A high level summary of data, along with dates and participants involved.

Sheet 2 (Data): Survey data collected through face to face interviews and online mode. Table 1 describes the specifications of variables for this data. We have referred this dataset as "Semi-literate Texting(SLT) Data" in this paper.

We have observed that more than 70% of text messages (SLT Dataset) were in regional languages, hence a translation has been performed (by bilingual surveyors) during the process of

Table 1
Variables and their description (Semi-literate Texting (SLT) Data).

Variable	Type	Description
Mode	Categorical	Face to face Online
Demography	Categorical	Rural Urban
Agency	Categorical	Agency Code
Gender	Categorical	Male, Female, Others
Age	Numeric	Age of the respondent
Town/Village, City, State	Categorical	Town/Village, City and State of residency
Education	Categorical	0 – No formal education 5, 8, 10, 12 – Highest level of education
Do you use Smartphone	Categorical	Yes, No
Do you send text message	Categorical	Yes, No When selected 'No' these recipients only receive and read messages
Frequency	Categorical	Daily, Weekly, Monthly, Never
Recipients	Categorical, Multiple options can be selected	Family, Friends, Employer, Others
Outstation communication	Categorical	Yes, No
Profession	Categorical/ Test	Categorical profession of interviewee
Language	Categorical	English, Hindi, Marathi
Message (1-10)	Long Text	Translated English language text messages free from Personally identifiable information (PII) data. (The messages are not included in the shared data location and can be found through email request to authors)
Length	Numeric	Average length of the translated English message

data collection. This was done to make the dataset useful to a more universal researcher group. All the text messages have been verified for Personally Identifiable Information (PII) information and any PII content inside the messages has been anonymized.

A high-level co-relation analysis has been performed on this data. The relationship is established as a heat map (Fig. 1). (We have used Python and seaborn package to derive this, after doing data conversion from categorical to numeric)

Summary statistics for SLT Dataset (Impact of various parameters on Education and Length of Text message) is presented in Tables 2 and 3.

A high-level analysis of text messages was done to understand the focus and sentiment of conversations. We have used word cloud [8] and a rule-based sentiment analysis VADER [9] (Valence Aware Dictionary and sentiment Reasoner) to demonstrate these. Figs. 2 and 3 demonstrate the same.

No inferences or conclusions has been derived and mentioned in this paper to avoid creating any bias for any other researcher who wish to use this dataset. We are working on analysing this dataset and will publish our work independently.

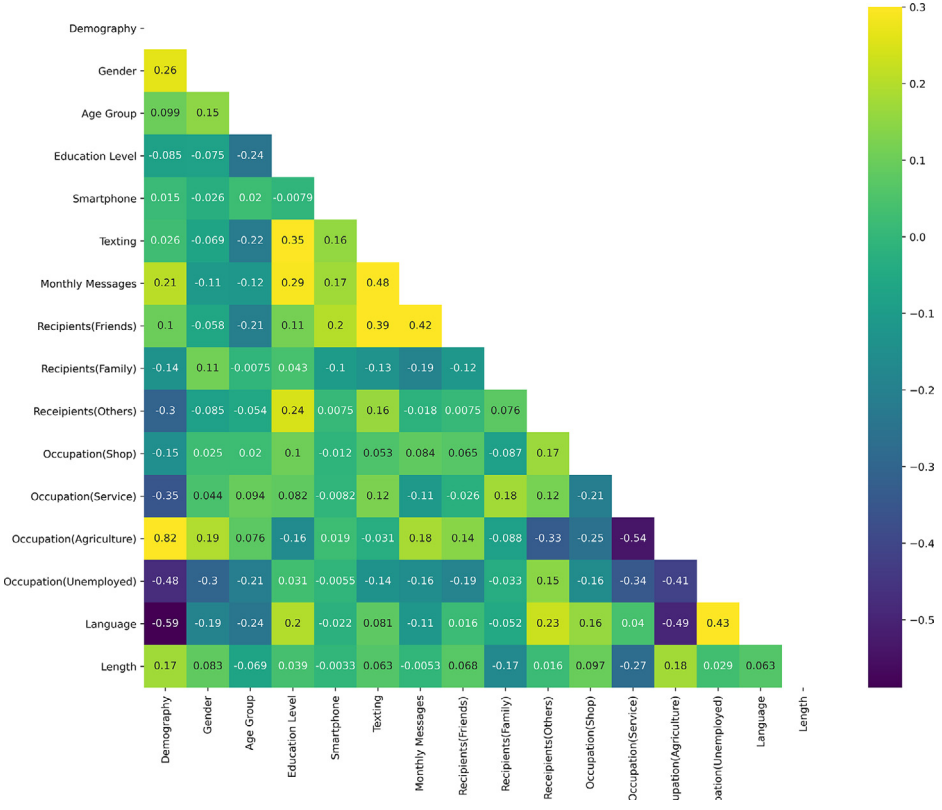

Fig. 1. SLT Dataset statistics – Heatmap for various parameters.

2. Experimental Design, Materials and Methods

2.1. SLT dataset: collection from survey

Surveys were conducted through (1) Face to face interview via hard copy collection of information by trained professionals and (2) Online survey with digitally semi-literate participants. For online participants, a brief video (in local language) specifying the intent for data collection, general directions, examples and best practices was created and shared. A total of 382 respondents were engaged out of which 342 were interviewed in face to face mode and 40 in online mode. We have observed that only a small fraction of online respondents shared meaningful text messages. This further verifies our hypothesis of problems with digital enablement for these people.

A survey form (refer Appendix) with 11 questions on metadata and 10 questions on actual text messages was designed. A team of 4 professional surveyors was hired and trained on the methods of data collection, questions to be asked to probe the respondents to extract relevant information. Both, the surveyors and the respondents were paid for this job. A random audit was performed by calling the respondents validating the confirmation of survey being conducted. A total of 3368 text messages across urban and rural India has been compiled.

Table 2
SLT Dataset statistics – Impact of gender, demography, age and profession on education.

		Education				
		Min	P25	Median	P75	Max
Gender	**Female**	0	10	10	12	12
	Male	0	8	10	12	12
Demography	**Urban**	0	10	12	12	12
	Rural	0	8	10	12	12
Age	**<=18**	8	10	10	12	12
	19-40	0	10	10	12	12
	41-60	0	8	10	12	12
	>=61	5	8	8	8	10
Profession	**Agriculture**	0	8	10	12	12
	Service	0	10	10	12	12
	Shop Owner	0	10	12	12	12
	Student	10	10	12	12	12
	Unemployed	0	8	10	12	12

Table 3
SLT Data statistics – Impact of gender, demography, age and profession on length of text messages (in characters).

		Length of Text Messages				
		Min	P25	Median	P75	Max
Gender	**Female**	15.75	28.5	34.7	42	66
	Male	13	30.9	36.3	41.4	59
Demography	**Urban**	13	26.5	34.1	42.3	65.1
	Rural	15.75	33	36.6	41	66
Age	**<=18**	28.25	36.5	42.2	44.7	59
	19-40	13	29.7	35.8	41.4	65.1
	41-60	19.5	30.9	35.7	40.6	66
	>=61	24.8	26.5	30.6	30.6	33.7
Profession	**Agriculture**	15.7	33	37.6	41.25	55.3
	Service	15.8	25.2	31.4	37.1	60.9
	Shop Owner	13	30.1	37.5	46.75	66
	Student	32.6	41.3	43.2	44.6	59
	Unemployed	18.9	26.8	32.5	40.4	50.7

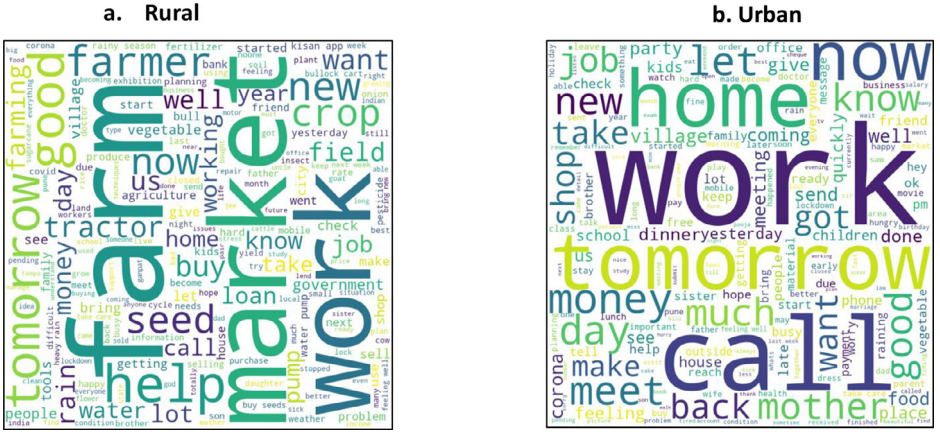

Fig. 2. SLT dataset statistics – Word cloud for rural and urban conversations.

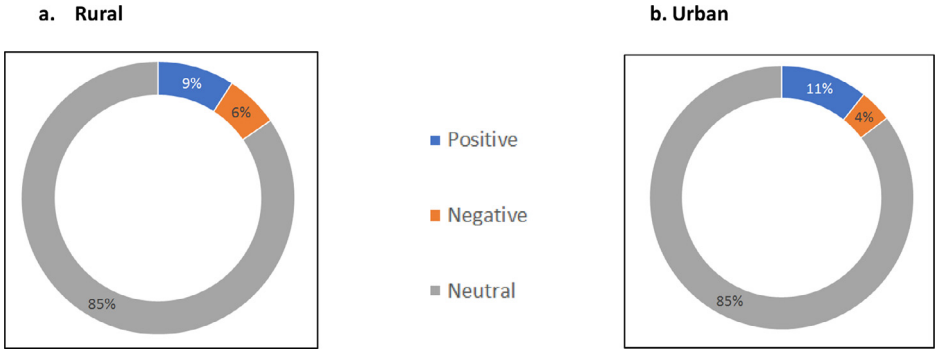

Fig. 3. SLT dataset statistics – Rule based, sentiment detection for rural and urban conversations.

Ethics Statement

Participation in the survey has been voluntary. Informed consent in writing has been obtained from all survey participants. The privacy rights of human subjects has been observed. No personal identifier information (PII) or clnical data has been collected as part of our survey. The research follows the "Guidelines for Ethical Considerations in Social Research & Evaluation in India" as described by CMS [10]. A self-administered Ethics Sensitivity Test was conducted based on the guidelines and grade of "Very Good" has been observed. From the best of our understanding there is no additional approval needed on ethics to use this data for research purpose.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships which have, or could be perceived to have, influenced the work reported in this article.

Data Availability

digitally semi-literate text message dataset (Original data) (Mendeley Data).

Acknowledgments

The authors would like to thank to all the volunteers who have given inputs to the survey.

Supplementary Materials

Supplementary material associated with this article can be found in the online version at doi:10.1016/j.dib.2021.107329.

References

[1] P. Glister, Digital Literacy, Wiley Computer Pub., New York, 1997.
[2] J. Coldwell-Neilson, T. Cooper, N. Patterson, Capability Demands of Digital Service Innovation, In: *Leadership, Management, and Adoption Techniques for Digital Service Innovation*, IGI Global, 2020, pp. 45–64.
[3] digitalindia| Digital India Programme | Government of India, 2020 https://www.digitalindia.gov.in/.

[4] T. Nayyar, S. Aggarwal, D. Khatter, K. Kumar, S. Goswami, L. Saini, Opportunities and Challenges in Digital Literacy: Assessing the Impact of Digital Literacy Training for Empowering. Urban Poor Women
[5] A.S. Khokhar, Digital literacy: how prepared is India to embrace it? International Journal of Digital Literacy and Digital Competence (IJDLDC) 7.3 (2016) 1–12.
[6] P.G. Rao, J. Ramey, Use of mobile phones by non-literate and semi-literate people: A systematic literature review, In: Proceedings of the IEEE International Professional Communication Conference, IEEE, 2011.
[7] A. Masizana-Katongo, R. Morakanyane, Representing Information for Semi-Literate Users: Digital Inclusion Using Mobile Phone Technology, Department of Computer Science, University of Botswana, Gaborone (nd), 2009.
[8] M. Burch, et al., Prefix tag clouds, in: Proceedings of the 17th International Conference on Information Visualisation, IEEE, 2013.
[9] C.J. Hutto, E. Gilbert, "Vader: a parsimonious rule-based model for sentiment analysis of social media text." Proceedings of the Eighth International Conference on Weblogs and Social Media (ICWSM 2014).
[10] Center for Media Studies (CMS). https://cmsindia.org/.