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

Ankush Agarwal
Gurmukh Singh
Ria Agarwal
Sania Rawat

Indian Statistical Institute, Bangalore Centre
21st April, 2023

ABSTRACT

The aim of this report is to replicate and build upon the analyses presented in the paper 'Semi-literate Texting (SLT): Survey based text message dataset from digitally semi-literate users in India' by Prawaal Sharma, Navneet Goyal and Vinay MR. In this report, we outline the methods used and the sources we consulted to carry out this project. By utilizing a combination of text-based analysis methods and linear models, we were able to expand on the original findings and provide new insights into the topic.

Contents

1. Introduction

Technology has changed how we access and exchange information, but in India, many urban and rural poor people still lack digital access. Mobile phones and internet are widely available and affordable in India, which has helped spread their use to grassroots levels. However, much of the information available on digital platforms is too complex for new mobile phone users to understand. As a result, the best way for these users to utilize technology is by exchanging text messages within their community to share and discuss relevant issues.

Mobile phones allow people to communicate through text messages, but users in India often use local languages or roman script, which can make it difficult to comprehend information. Improvements to text messaging technology could help bridge the digital divide and improve adoption of digital tools.

Digital literacy is a term that lacks a clear and universally agreed upon definition in research. It was first coined by Gilster [1] in the late 1900s, who measured it in terms of education and information skills. However, the definition of digital literacy has evolved over time, and is now generally considered to range from being technologically fluent to being able to use digital platforms without assistance. This view has been expressed by Chase and Laufenberg [2], among others.

In India, the government's "Digital India" [3] program aims to provide information to all citizens, but lack of tertiary education [4] and poor literacy rates, particularly among women and rural populations, are major barriers to digital literacy.

Users are typically labeled as digitally literate or illiterate. However, there exist people who are neither. These users have basic education, access to mobile phones and the internet, but struggle with technology and have limited English vocabulary. We refer to them as "semi-literate" in our paper.

So, by understanding how digitally semi-literate users communicate, it can help make designing interfaces for human-computer interaction easier.

We attempted to do that in the following ways:

1. Analysing all the categorical and numerical variables individually in section 3.1, and with respect to one another in section 3.2.

2. Performing high-level correlation analysis on the variables in section 3.3.

3. Conducting text based visual and sentimental analysis , as well as a readability test in sections 4.2, 4.1 and 4.3, respectively.

4. In section 5, we address two important questions concerning digital semi-literacy in India.

2. Data Description

The data was collected through a survey conducted between July 2020 and November 2020, largely across urban and rural geographies of western region of Maharashtra, India. It was mostly a face to face survey conducted by trained professionals with approximately 90% through them and the rest through the online mode. A total of 3368 messages were collected from a total of 382 respondents.

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, 5, 8, 10, 12
Do you use Smartphone	Categorical	Yes, No
Do you send text message	Categorical	Yes, No
Frequency	Categorical	Daily, Weekly, Monthly, Never
Recipients	Categorical	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	Text messages
Length	Numeric	Average length of the English message

Table 2.1: Variables and their description

There are several variables, out of which arguably the most important one is the actual text message data that was acquired, wherein each respondent gave not more than 10 one-to-one messages.

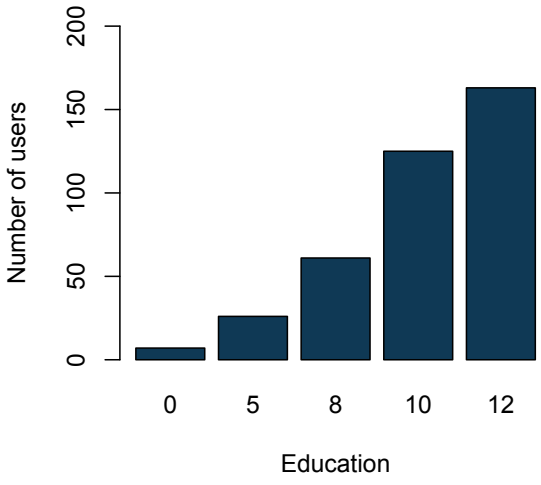
Steps taken to ensure the data collected could be analysed and pre-processed.

- Only the people who identifed as digitally semi-literate were surveyed.
- To make sure no unintended access was gained to personal information, all personally identifiable information data was anonymized.
- For all the text messages (70% of the total) that were in regional languages, bilingual surveyors were present to perform translations.
- Collection of forward messages, motivation messages and generic messages where there is no exchange of information was discouraged.

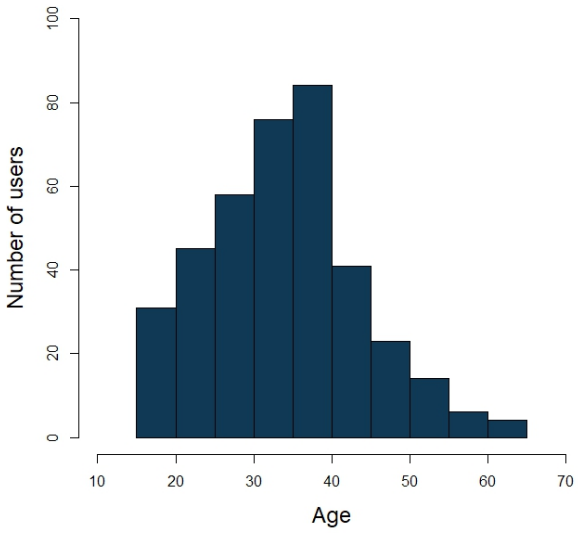
3. Descriptive Statistics

3.1 Variable Distributions

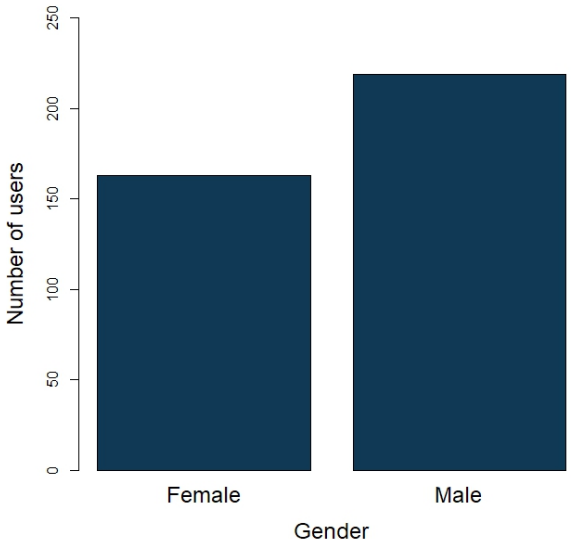
The following are barplots and histograms displaying individual distributions of the numeric and categorical variables.



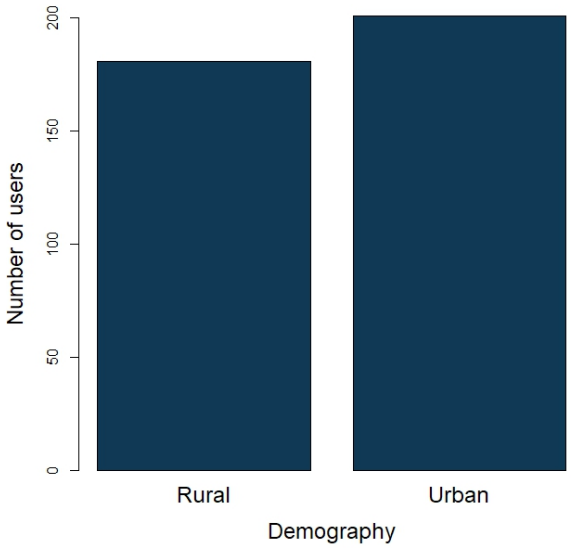
(a) Barplot of Education



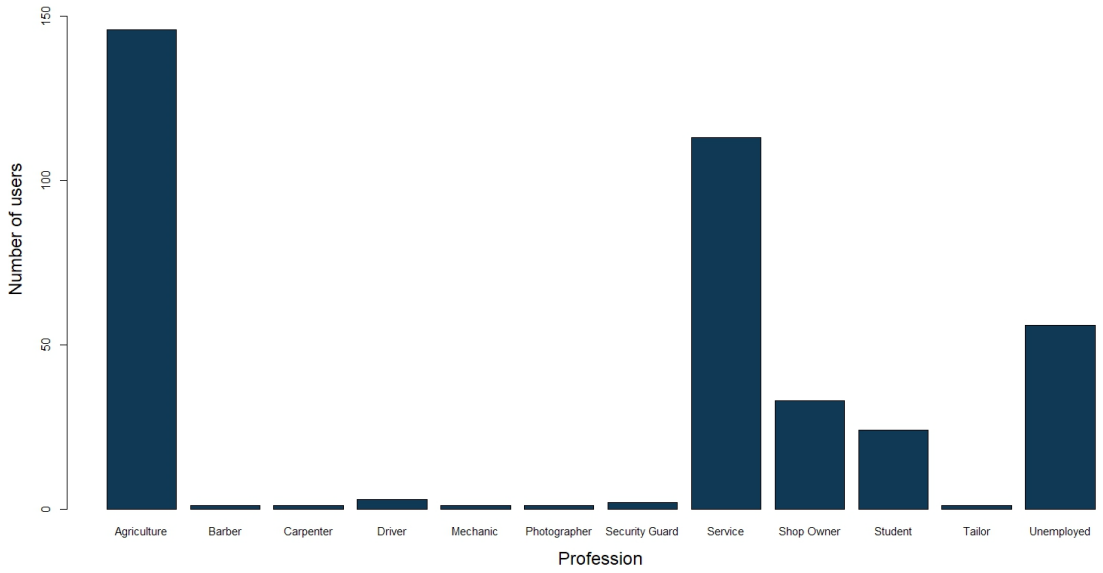
(b) Histogram of Age



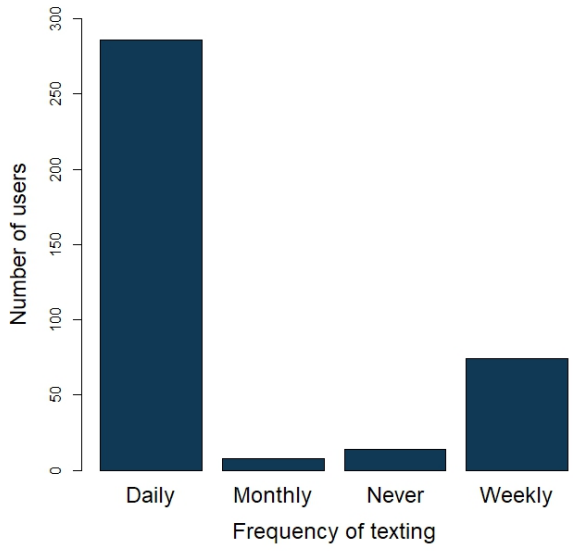
(c) Barplot of Gender



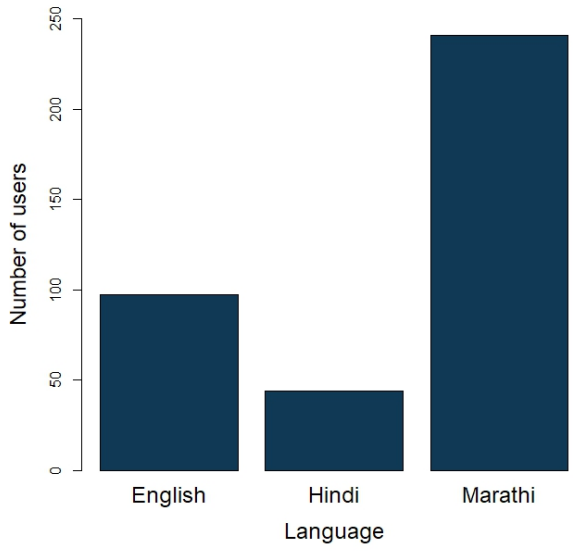
(d) Barplot of Demography



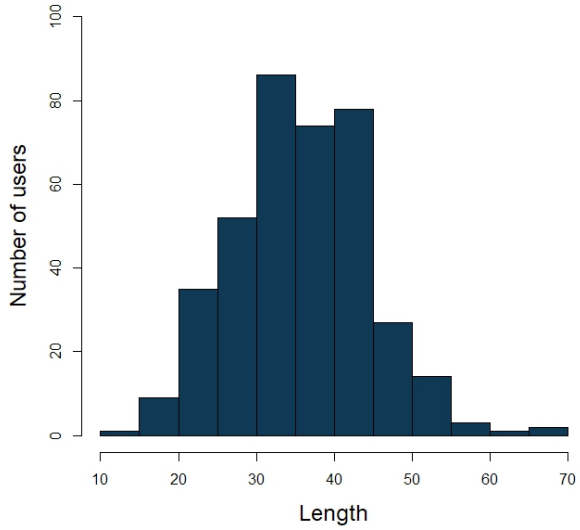
(e) Barplot of Profession



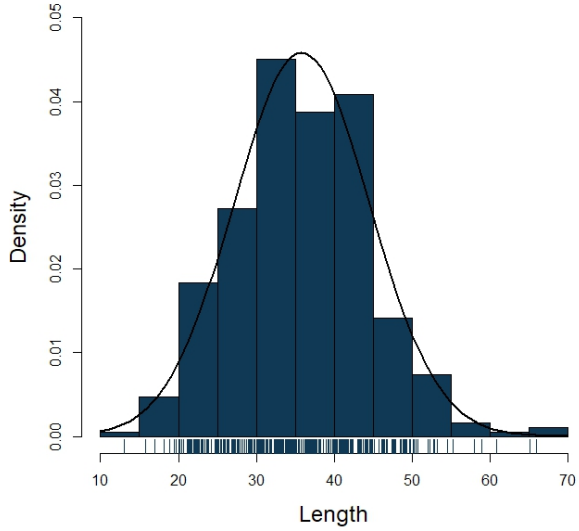
(f) Barplot of Frequency



(g) Barplot of Language



(h) Histogram of Average Length of Text Messages



(i) Histogram with Density Plot of Average Length of Text Messages

Observations:

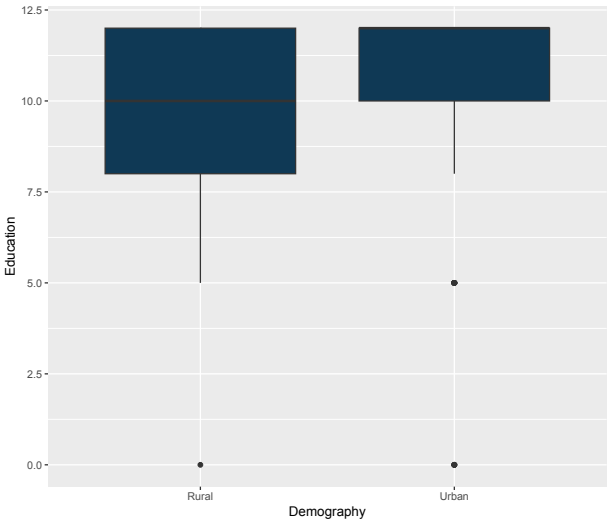
- Most of the semi-literate bear 12th standard education, with fewer people bearing further less education, conveniently giving the graph, in figure 3.1a, an increasing look.
- We observe from the figure 3.1b that the respondents belong to various age groups, spanning from the youngest to the oldest, providing us with a comprehensive representation of the population under study. The average age of the respondents is roughly 35 years, with the youngest volunteer being of age 18 whereas the oldest one stands at 65 years old.
- Figure 3.1f shows us that most of the people, irrespective of demography, text daily. The representation of the people who text daily is 74.69 percent, with the remaining partitions being left to the people who text weekly (19.37 percent), monthly (2.09 percent) and never(3.85 percent)
- Not so surprisingly, most of the people use their regional language to converse on the phone, while english seems to be dominant among the remaining groups, as shown in figure 3.1g.
- The figure 3.1h shows that the distribution of average length of text messages tends to crowd around 30 to 45 characters, with fewer people typing in the ranges as we sidetrack from the mode of the data. The data overall bears similarity to the normal distribution, which is demarcated by the overdrawn curve in the second graph in the figure 3.1i.

3.2 Impact of various parameters on Education and Length of Text message

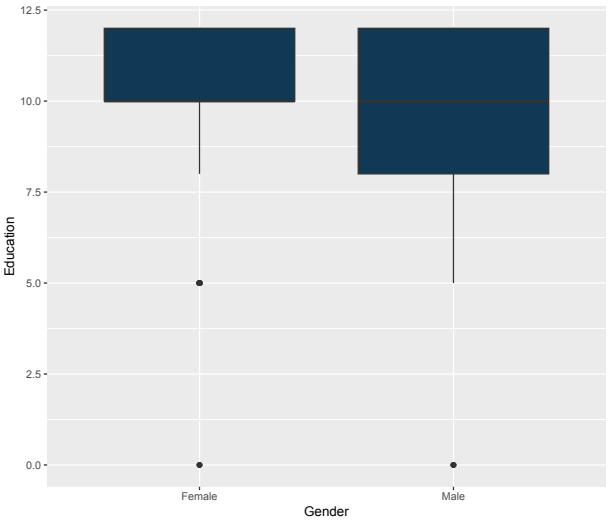
The following tables and boxplots accurately depict the distribution of users on their education level and average length of text messages on the basis of gender, demography, age and profession.

Variable	Category	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	$\leqslant 18$	8	10	10	12	12
	19-40	0	10	10	12	12
	41-60	0	8	10	12	12
	$\geqslant 61$	5	7.25	8	8.5	10
Profession	Agriculture	0	8	10	12	12
	Service	0	10	12	12	12
	Shop Owner	0	10	12	12	12
	Student	10	10	12	12	12
	Unemployed	0	8	10	12	12

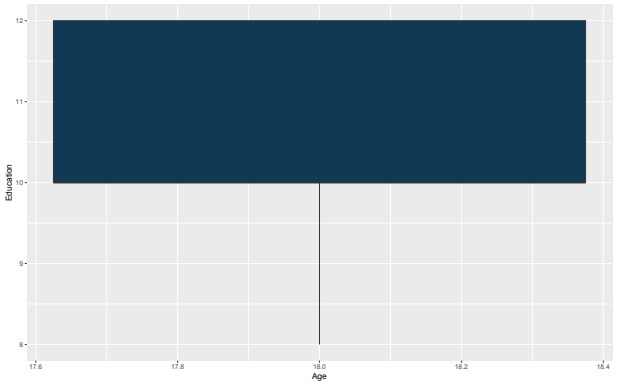
Table 3.1: Education



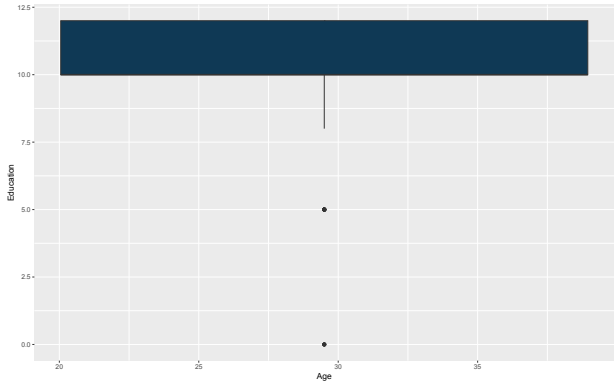
(a) Box plots for education categorized by demography



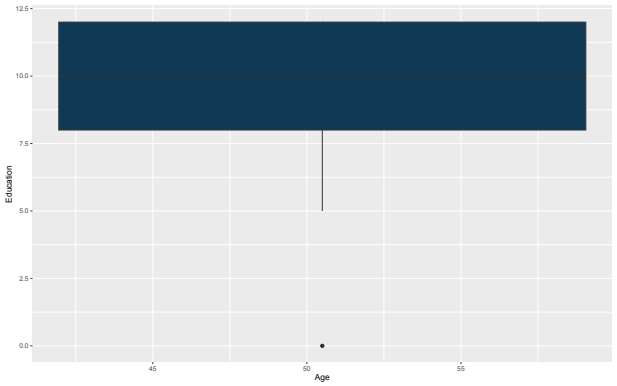
(b) Box plots for education categorized by gender



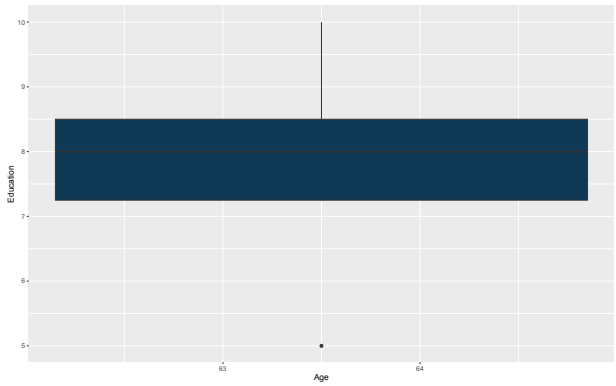
(c) ≤ 18



(d) 19 to 40

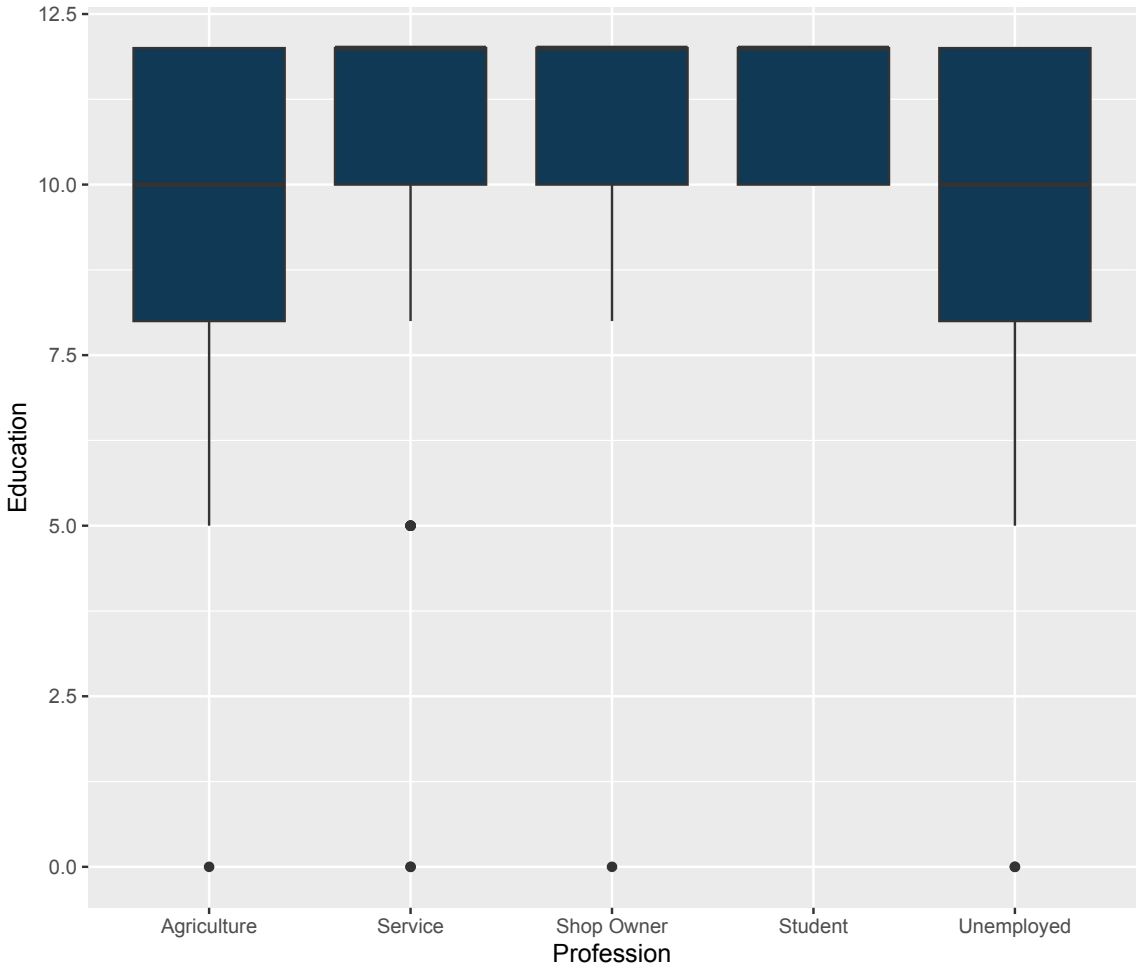


(e) 41 to 60



(f) ≥ 61

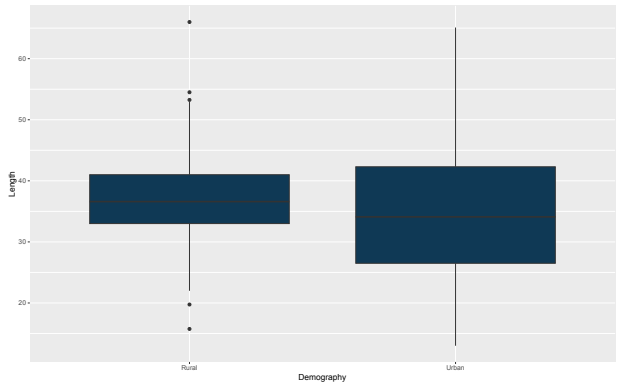
Figure 3.2: Boxplots for education categorized by age



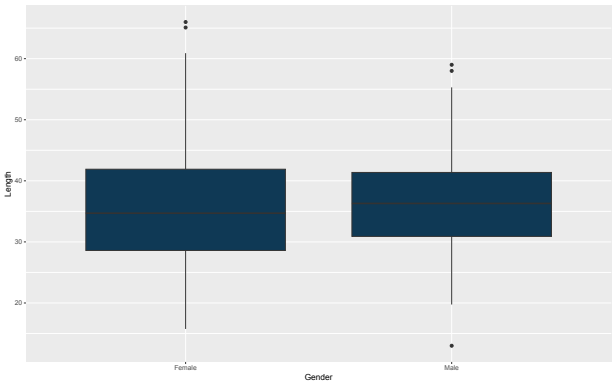
(g) Box plots for education categorized by profession

Variable	Category	Min	P25	Median	P75	Max
Gender	Female	15.75	28.6	34.7	41.9	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	$\leqslant 18$	28.25	36.65	42.71	44.46	59
	19-40	13	29.77	35.77	41.44	65.1
	41-60	19.5	30.9	35.55	40.73	66
	$\geqslant 61$	24.8	26.07	28.55	31.38	33.7
Profession	Agriculture	22	33.42	37.67	41.24	55.3
	Service	15.8	24.9	31.38	36.75	60.9
	Shop Owner	13	30.1	37.5	46.1	66
	Student	32.6	41	43.1	44.64	59
	Unemployed	18.9	26.9	33.55	40.98	50.7

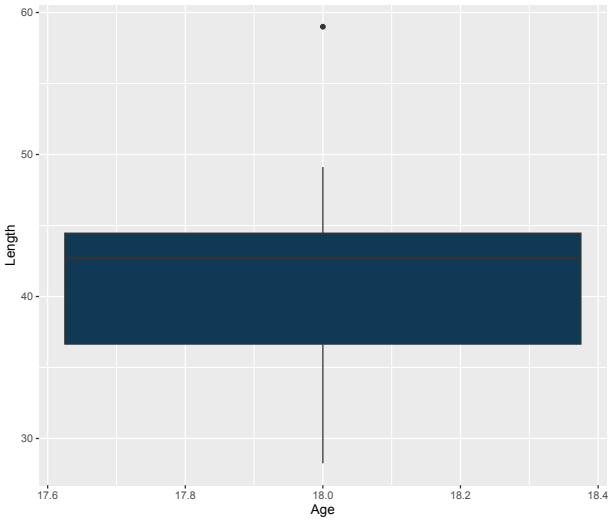
Table 3.2: Length of Text Messages



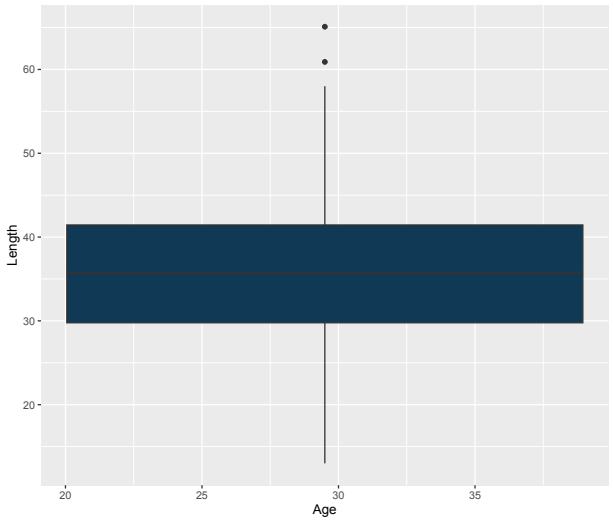
(a) Box plots for length of text messages categorized by demography



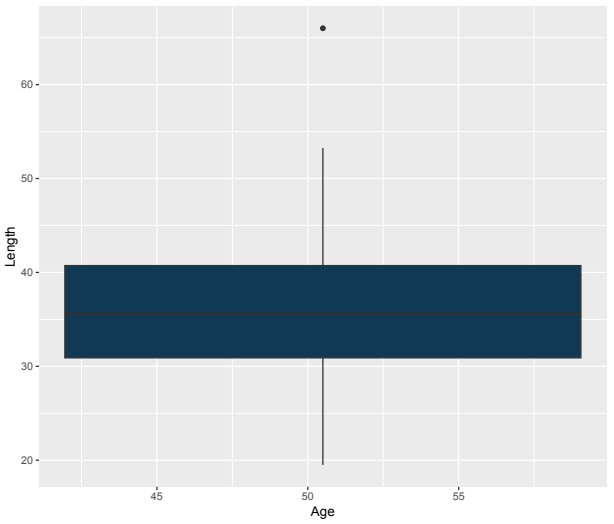
(b) Box plots for length of text messages categorized by gender



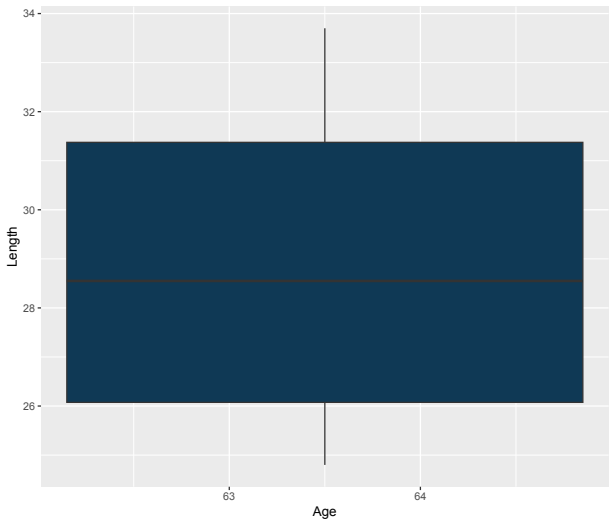
(c) ≤ 18



(d) 19 to 40

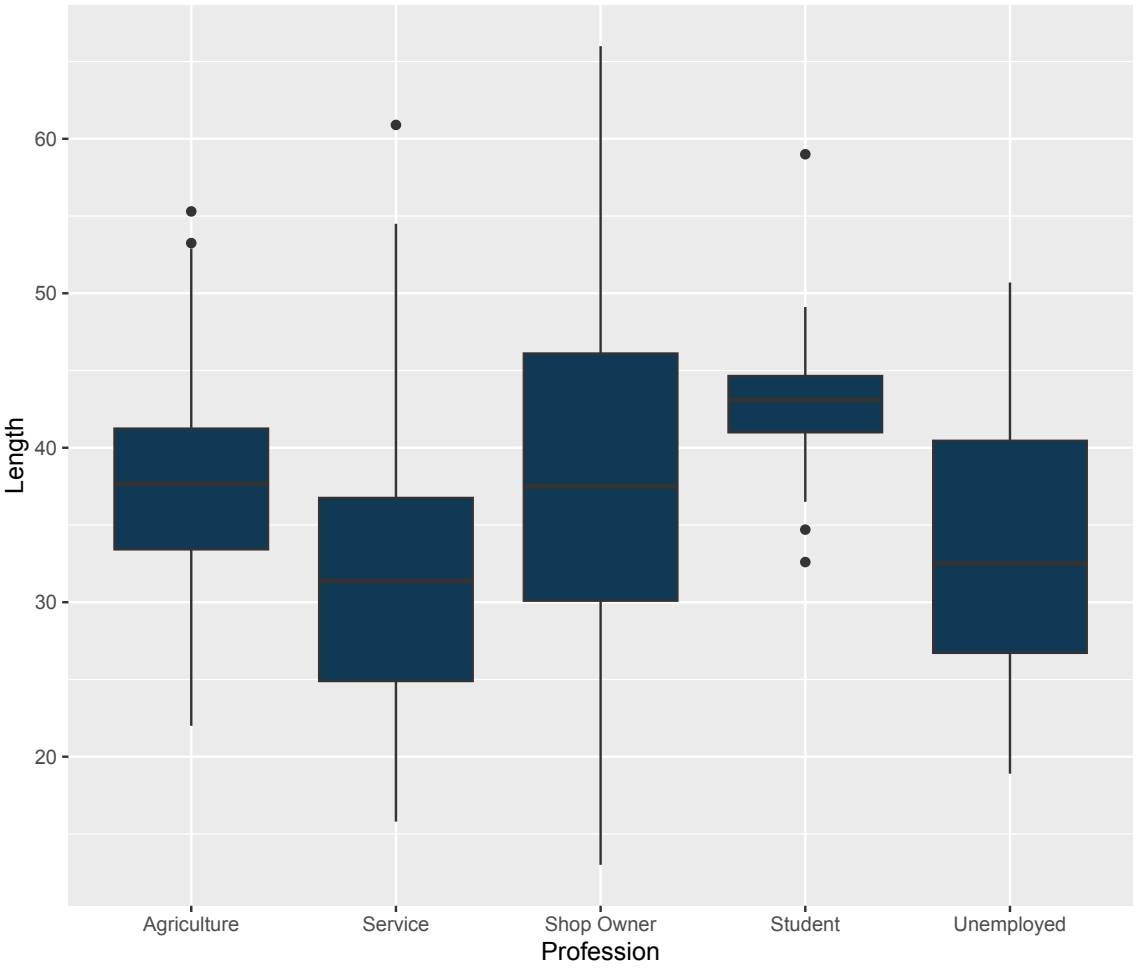


(e) 41 to 60



(f) ≥ 61

Figure 3.3: Boxplots for education categorized by age



(g) Box plots for length of text messages categorized by profession

3.3 Correlation Heatmap

A correlation heatmap is a visual representation of the correlation matrix, which shows the correlation between different variables in a dataset. It consists of tiles, and its color indicates the strength and direction of the correlation between the variables that it correponds to. Typically, a darker color indicates a stronger correlation, while a lighter color indicates a weaker correlation.

As correlation coefficient can only be calculated between numerical variables, it is impractical to consider it for categorical variables, like Gender and Demography, in our case. One of the main problems encountered was multicollinearity. It refers to a condition in which the independent variables are correlated to each other.

Thus, we adopted the method of creating dummy variables to help eliminate multicollinearity [8]. Dummy variable coding involves creating a set of binary (0 or 1) variables, with one variable representing each category of a categorical variable. Without using dummy variables, it would be necessary to use the categorical variable itself as a predictor variable in the model, which can result in collinearity among the predictors. This is because the different categories of the categorical variable are not independent of each other, and there is a linear relationship among them.

The following heatmap displays the correlation between the variables in the dataset.

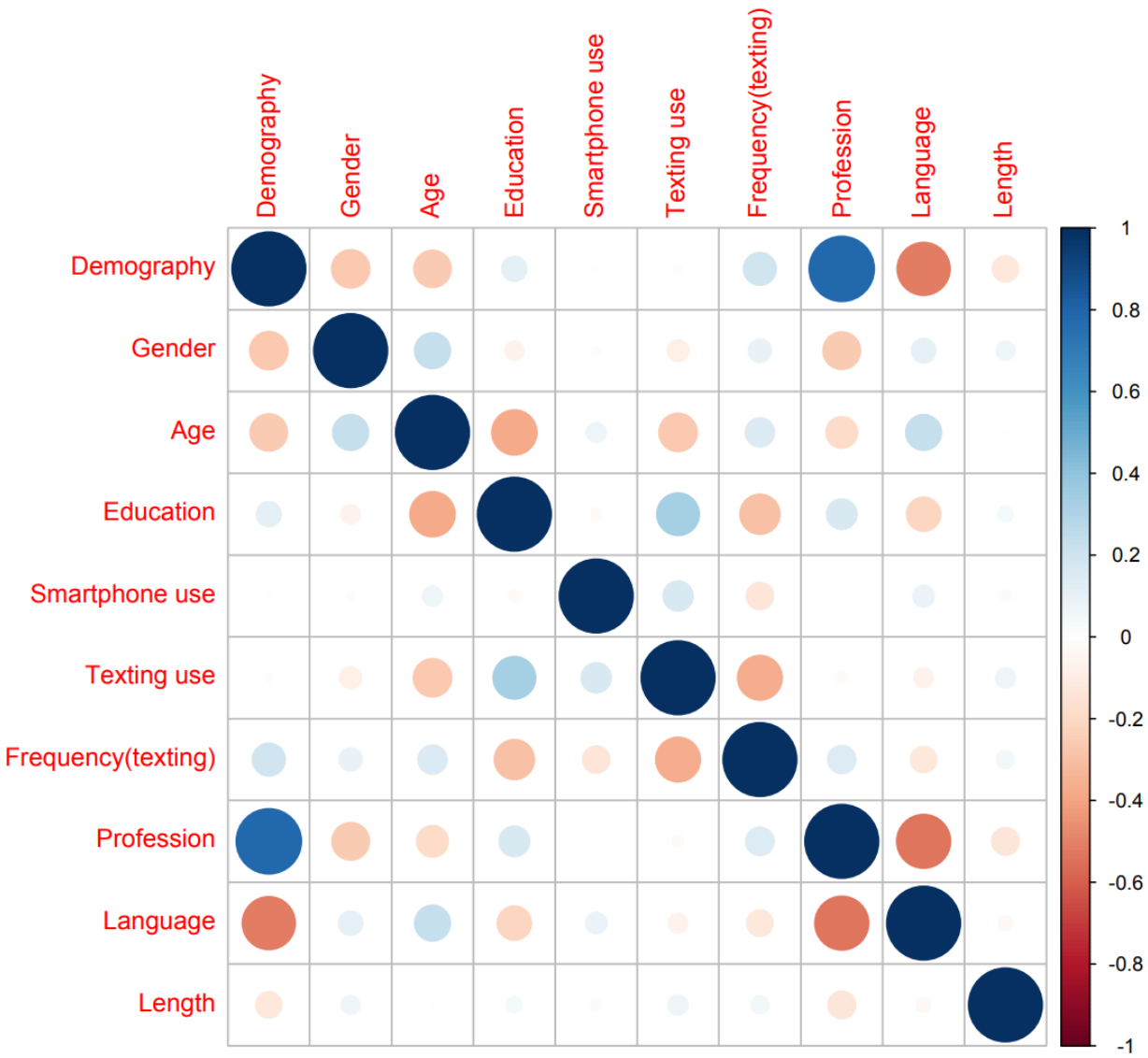


Figure 3.4: Correlation Heatmap of the variables labelled according to the colours

4. Text-based Analysis

Now we branch into the essential part of the paper where we apply text-based analysis methods that let us further understand and interpret the main components of the provided data, i.e., the text messages.

Text preprocessing is an essential step in text analysis. It involves transforming raw text data into a format that can be easily analyzed and processed by machine learning algorithms. The following steps were taken for the same

1. **Tokenization:** The text was broken down into individual words.
2. **Spell Checking and Correction:** All spelling mistakes and word repetitions in the text were corrected.
3. **Removal of Punctuation:** Punctuation such as commas, periods and question marks were removed from the text.
4. **Lowercasing:** All text was converted into lowercase to ensure consistency.
5. **Stop Word Removal:** All common words such as "and", "the" and "a" were removed since they do not provide significant meaning to the text.
6. **Stemming:** Words such as "running" were reduced to their root or base form of "run".
7. **Removal of Special Characters:** Special characters such as hashtags, backslash, and mentions were removed from the text.
8. **Term Frequency-Inverse Document Frequency (TF-IDF) weighting:** The importance of all words in the text data was evaluated for the purpose of sentiment analysis, using this statistical technique, wherein importance of a word in a document is evaluated by measuring the frequency of its occurrence in the document and comparing it with its frequency in other documents in the corpus.

Finally, after all these preprocessing techniques were applied, the following text analysis methods were used on the text data.

4.1 Sentiment Analysis

We started by conducting sentiment analyses on the data, which is a method that identifies the emotional tone behind a body of text. It is used to determine whether a given text contains negative, positive or neutral emotions. For achieving that, the polarity of a certain body of text is determined. Polarity refers to the overall sentiment conveyed by a certain text, phrase or word. This polarity is expressed with a numerical rating known as a "sentimental score" that ranges positive to negative with 0 being the neutral value. Multiple ways of determining and interpreting this score exist.

There are two approaches to this method, the lexical approach and the machine learning approach, VADER being the prime example of the former. The data given is only eligible to having sentiment analysis done via the lexical approach.

So, we conducted VADER on the text data and extended the sentiment analysis using the Syuzhet method.

4.1.1 VADER

Valence Aware Dictionary for sEntiment Reasoning, or VADER [6], is a model that is sensitive to both polarity (positive/negative) and intensity (strength) of emotion.

This method uses a lexicon of words and phrases that have been manually annotated with their associated positive or negative sentiment score. The VADER lexicon includes over 7,500 lexical features, including adjectives, adverbs, conjunctions, and slang. Each feature is assigned a sentiment score between -1 (negative) and +1 (positive), with 0 representing neutral sentiment.

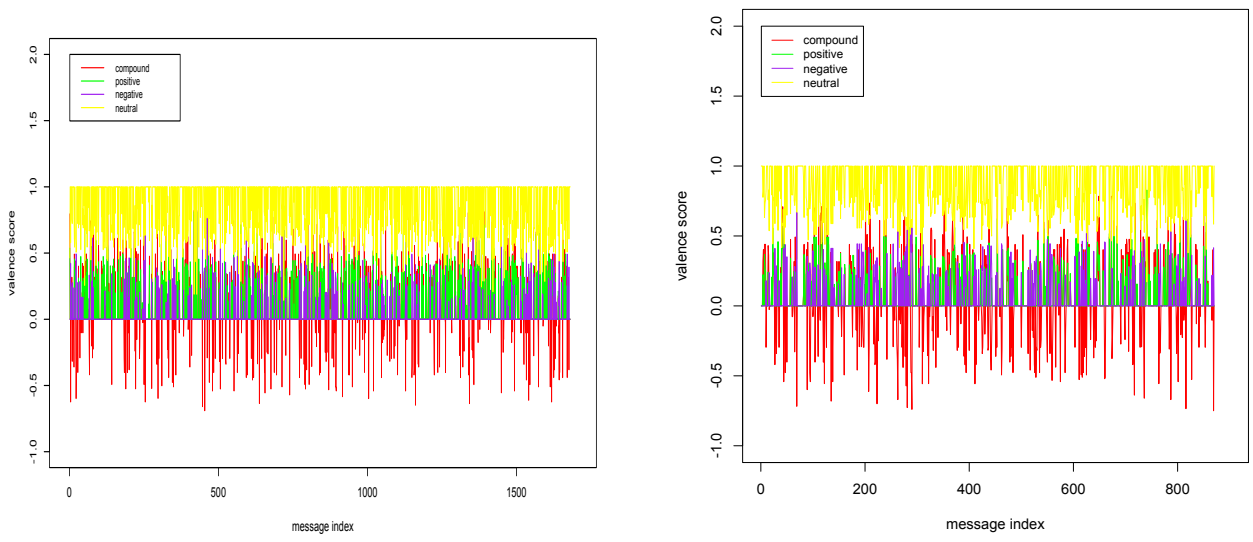
The VADER method generates sentiment scores for each sentence in a piece of text and combines them to calculate an overall sentiment score for the entire text. The output of the VADER method includes the following four sentiment scores.

- Positive score: a score from 0 to 1 that represents the proportion of the text that is classified as positive.
- Negative score: a score from 0 to 1 that represents the proportion of the text that is classified as negative.
- Neutral score: a score from 0 to 1 that represents the proportion of the text that is classified as neutral.
- Compound score: a normalized, weighted composite score from -1 to 1 that represents the overall sentiment of the text. The normalization is

$$\frac{x}{\sqrt{x^2 + \alpha}}$$

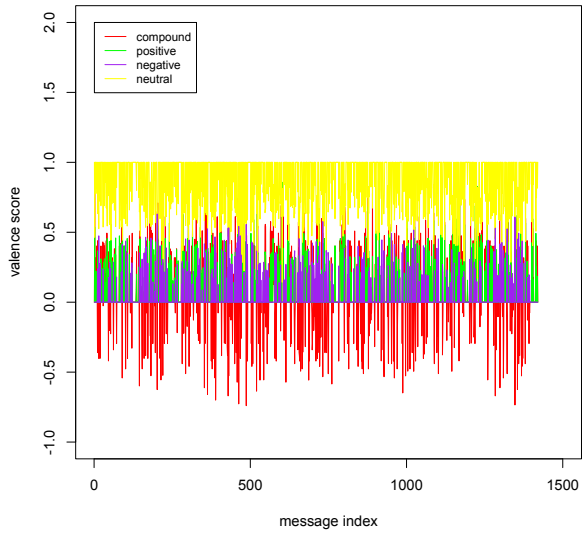
where x is the sum of the sentiment scores of the constituent words of the sentence and α is the normalization parameter that we set to 15.

We performed VADER analysis on the text data differentiated on the basis of the four variables: Gender, Demography, Age and Profession and the results are as follows:

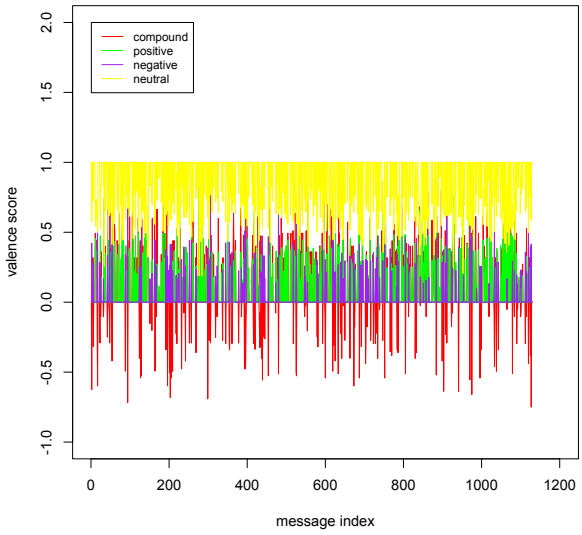


(a) Urban (b) Rural

Figure 4.1: Plots depicting VADER scores categorized by demography

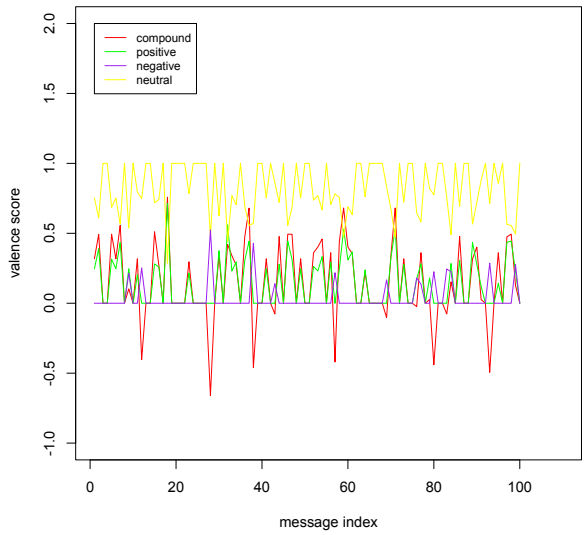


(a) Male

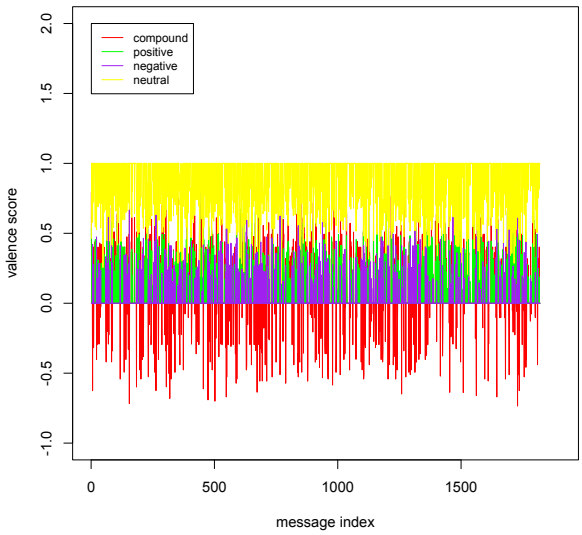


(b) Female

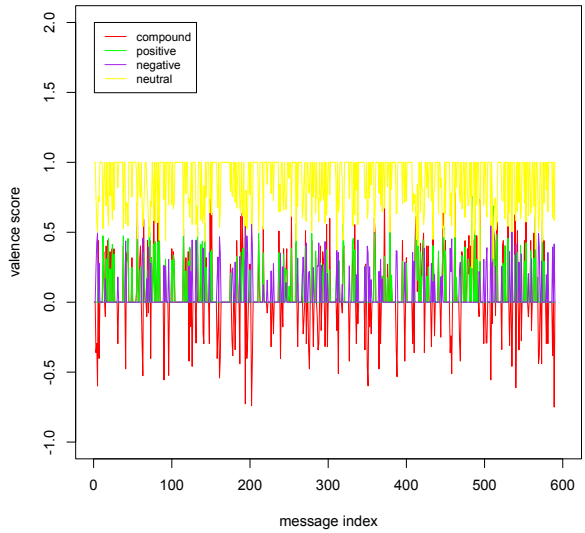
Figure 4.2: Plots depicting VADER scores categorized by gender



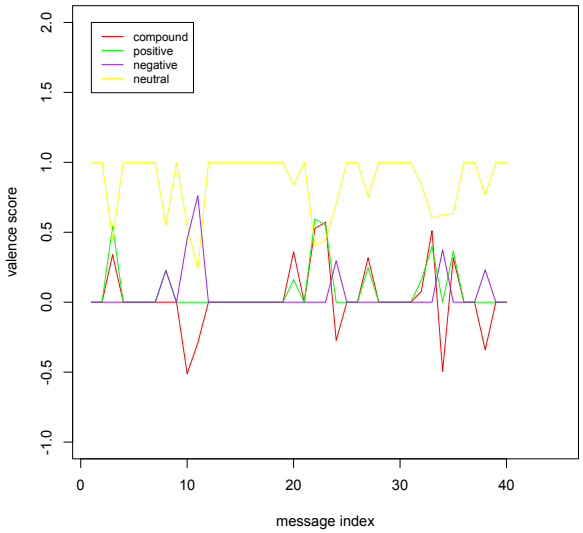
(a) $\leqslant 18$



(b) 19 to 40

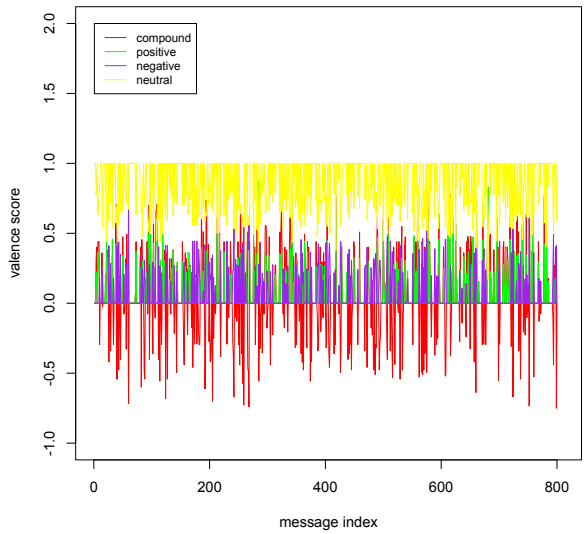


(c) 41 to 60

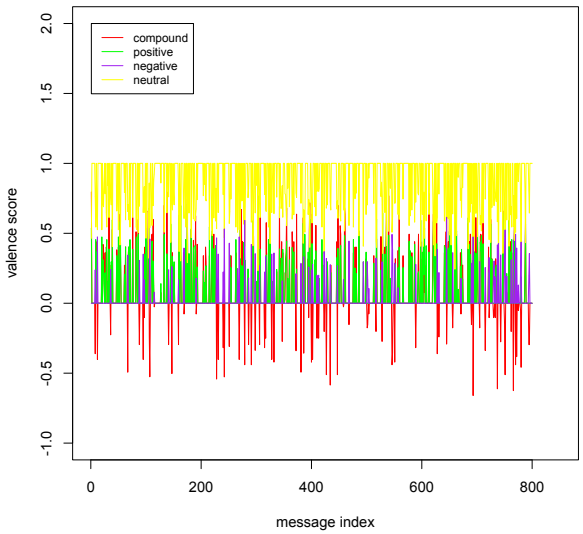


(d) $\geqslant 61$

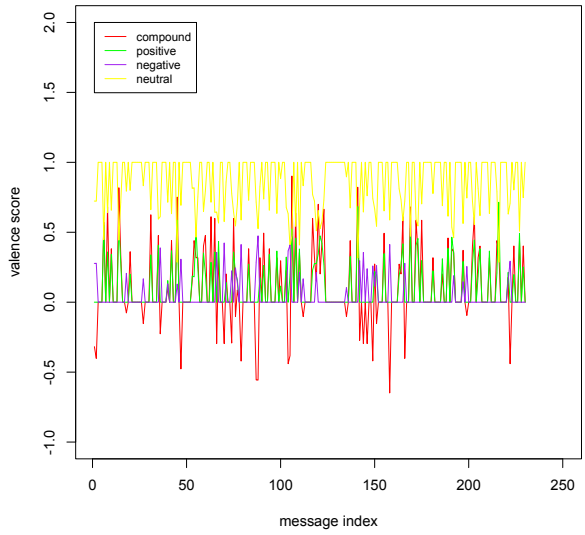
Figure 4.3: Plots depicting VADER scores categorized by age



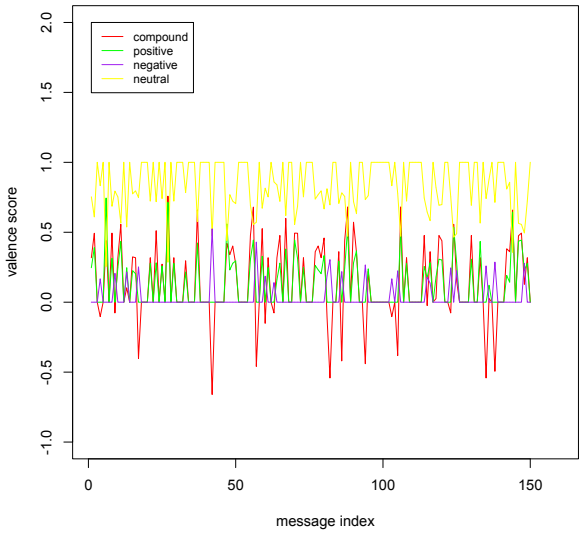
(a) Agriculture



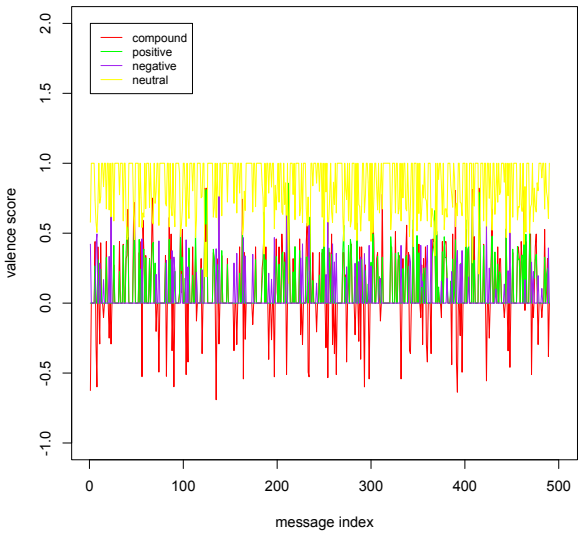
(b) Service



(c) Shop Owner



(d) Student



(e) Unemployed

Figure 4.4: Plots depicting VADER scores categorized by profession

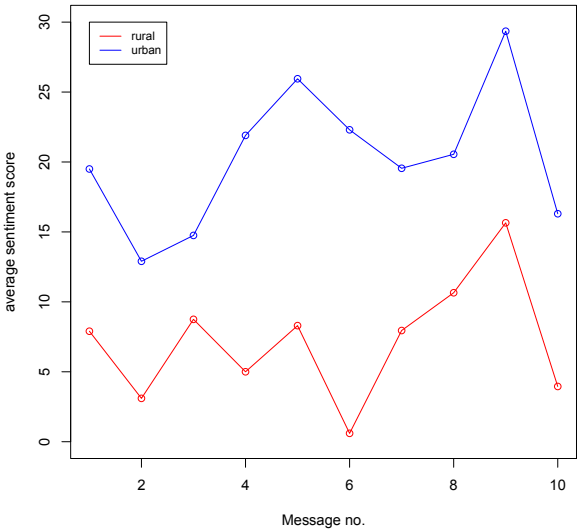
4.1.2 Syuzhet

The term Syuzhet stems from the Russian formalists Victor Shklovsky and Vladimir Propp who divided narrative into two components, the 'fabula' and the 'syuzhet' to depict narrative structures of a story, with 'fabula' being the raw material of the story and 'syuzhet' being the way the story is organized. Syuzhet intends to provide the latent structure of narrative by means of sentiment analysis and specifically the emotional shifts that serve as proxies for the narrative movement between conflict and conflict resolution.

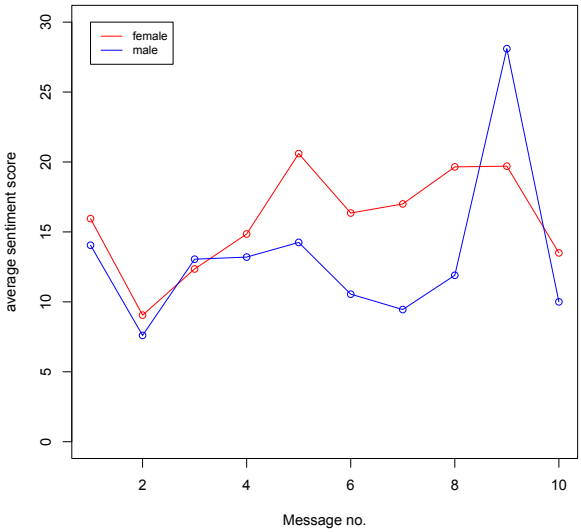
The Syuzhet method is a sentiment analysis technique that aims to discern the emotional tone or valence of a given text. This involves examining the linguistic and semantic properties of the text to ascertain its emotional character. By assigning numerical values to words based

on their emotional associations, the method can gauge the text's sentiment as a whole. This lexical technique relies on the concept that emotions can be represented as a spectrum that ranges from positive to negative, and that language can be used to detect and measure these emotional states.

Syuzhet scores was determined for messages (1-10) differentiated based on the four variables: Gender, Education, Age and Profession as seen in the following four figures.

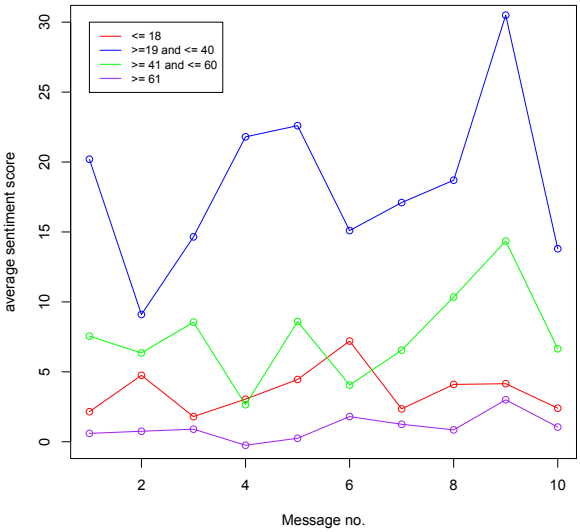


(a) Demography

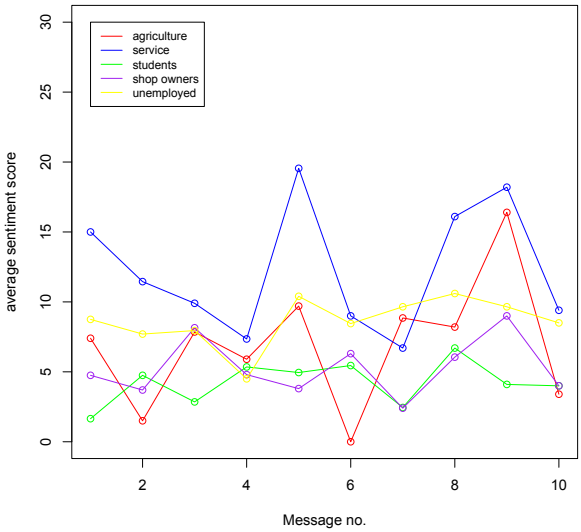


(b) Gender

Figure 4.5: Syuzhet scores



(c) Age



(d) Profession

Figure 4.5: Syuzhet scores

4.2 Wordcloud

Word clouds [5], also referred to as text clouds or tag clouds, operate on a straightforward principle: words that appear more frequently in a particular source of textual data, such as a speech, blog post, or database, are displayed in larger and bolder font within the word cloud. Essentially, a word cloud is a group or cluster of words arranged in different sizes. The larger and bolder the word, the more frequently it is mentioned in the given text and the more significant it is considered to be and therefore, these are effective tools for highlighting the most relevant portions of textual data.

The utilization of word clouds results in greater simplicity and clarity, and hence, they are a powerful tool for communication due to their ease of comprehension, shareability, and ability to make a significant impact as compared to table data.

Word clouds based on the text data were generated for all the four categories, demonstrating the different range of vocabularies used by the people are as follows:

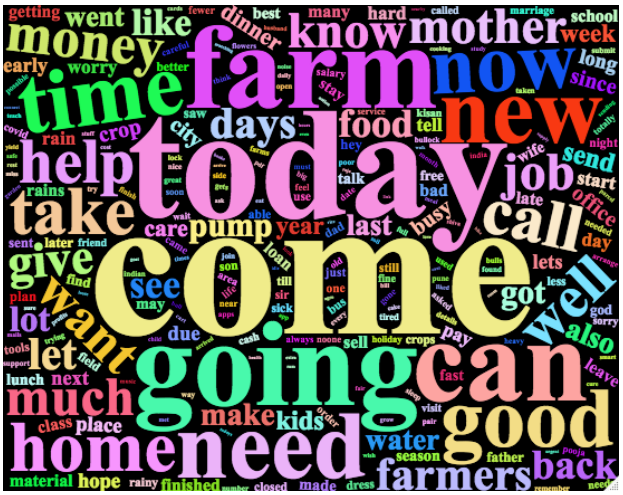
(a) Urban (b) Rural

Figure 4.6: Word clouds categorized by demography

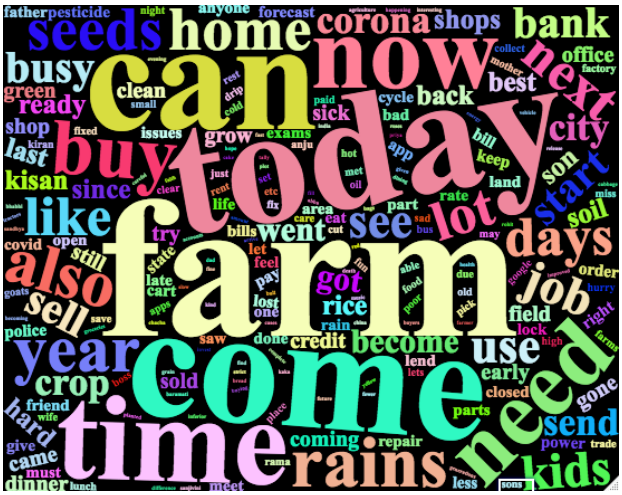
(a) Male (b) Female

Figure 4.7: Word clouds categorized by gender

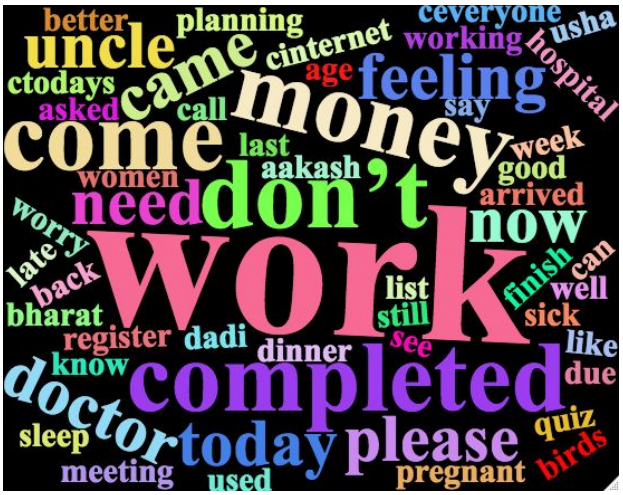
(a) ⩽ 18



(b) 19 to 40

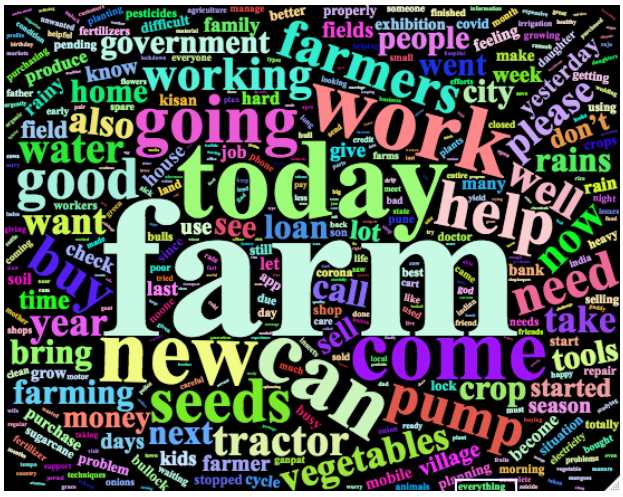


(c) 41 to 60

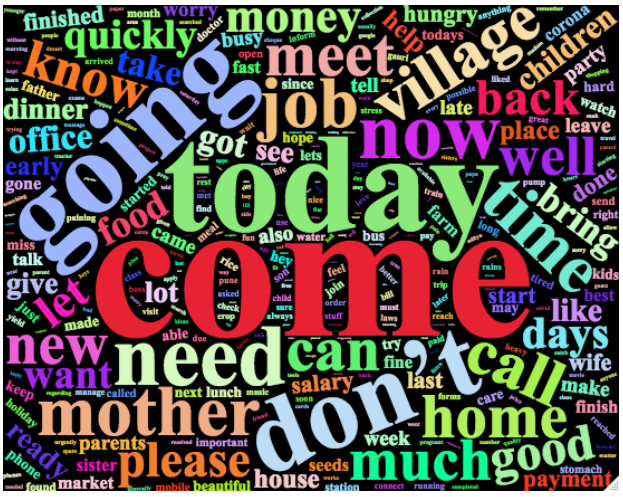


(d) ⩾ 61

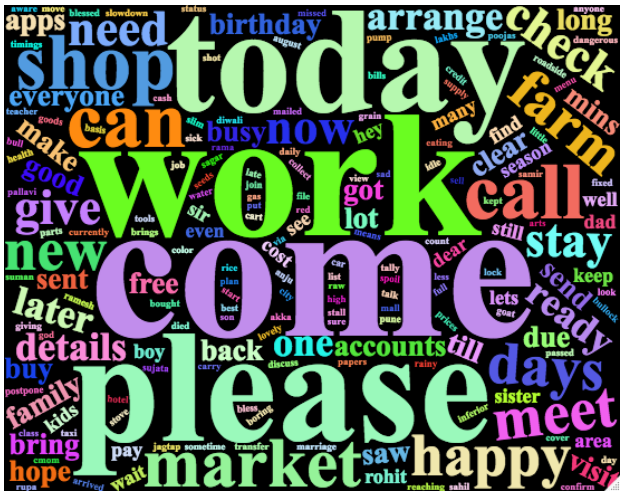
Figure 4.8: Word clouds categorized by age



(a) Agriculture

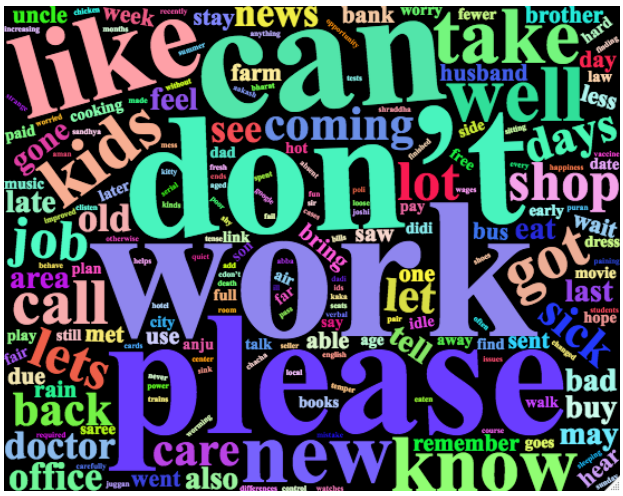


(b) Service



(c) Shop Owner

(d) Student



(e) Unemployed

Figure 4.9: Word clouds categorized by profession

4.3 Readability

The concept of readability pertains to the level of clarity and comprehension that a written material offers to its readers. It encompasses a range of factors, including sentence structure, vocabulary, and the presence of jargon or technical terms, all of which can influence the ease

or difficulty of reading and understanding the text.

Numerous formulas and algorithms exist for evaluating the readability of a text. Among the most widely recognised indices are the Flesch-Kincaid Grade Level, the Gunning Fog Index, and the Coleman-Liau Index. Each of these measures considers various factors, such as sentence length, word complexity, and grammatical structure, to assign a numerical score that reflects the text's level of difficulty.To assess the complexity of a written work, various metrics are employed that use numerical scales to assign a score to the text. The higher the score, the more challenging the text is considered to be.

We used the Traenkle-Bailer Method for evaluating the readability scores, the formula for which is given by

$$224.6814 - (79.8304 \times AWL) - (12.24032 \times ASL) - (1.292857 \times 100 \times \frac{\eta_{prep}}{\eta_w})$$

where

$$AWL = \text{Average Word Length } = \frac{\text{number of characters}}{\text{number of words}}$$
$$ASL = \text{Average Sentence Length } = \frac{\text{number of words}}{\text{number of sentences}}$$
$$\eta_w = \text{number of words}$$
$$\eta_{prep} = \text{number of prepositions.}$$

Variable	Category	Min	P25	Median	Mean	P75	Max
Gender	Female	-47609.9	-23866.8	-123.6	-15946.8	-115.3	-106.9
	Male	-66089.5	-33106.6	-123.6	-22106.7	-115.3	-106.9
Demography	Urban	-69517.4	-34820.5	-123.6	-23249.3	-115.3	-106.9
	Rural	-44145.5	-22134.6	-123.6	-14792.0	-115.3	-106.9
Age	$\leqslant 18$	-5506.1	-2814.9	-123.6	-1912.2	-115.3	-106.9
	19-40	-80216.4	-40170.0	-123.6	-26815.6	-115.3	-106.9
	41-60	-26809.9	-13466.8	-123.6	-9013.5	-115.3	-106.9
	$\geqslant 61$	-1650.7	-887.2	-123.6	-627.1	-115.3	-106.9
Profession	Agriculture	-40452.0	-20287.8	-123.6	-13560.9	-115.3	-106.9
	Service	-30388.0	-15255.8	-123.6	-10206.2	-115.3	-106.9
	Shop Owner	-11127.4	-5625.5	-123.6	-3786.0	-115.3	-106.9
	Student	-8226.0	-4174.8	-123.6	-2818.8	-115.3	-106.9
	Unemployed	-20568.5	-10346.1	-123.6	-6933.0	-115.3	-106.9

Table 4.1: Readability Scores

5. Inferential Statistics

5.1 Do variables such as age and gender significantly affect the average length of text messages of digitally semi-literate users in India?

We use a multiple linear regression model to address the question. Multiple linear regression is a statistical method used to model the relationship between multiple independent variables and a dependent variable. In other words, it is a technique for predicting a continuous outcome variable (also known as the response or dependent variable) based on two or more predictor variables (also known as independent or explanatory variables).

We took Length to be the dependent variable and Age and Gender to be the independent ones. As Gender is a categorical variable, we made dummy variable to ensure its accurate estimation by the model. A dummy variable [**??**] can be thought of as creating "two models," one for each category of a binary categorical variable.

The formula for regression in our model is

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon,$$

where Y is length, x_1 is age and x_2 is a dummy variable created by R that uses 1 for female.

We have used an interaction model that gives a unique slope for each gender [7]. An interaction model in multiple regression includes an interaction term between two or more predictor variables and allows for the possibility that the effect of one predictor variable on the outcome variable may depend on the level of another predictor variable.

Our hypotheses:

Null Hypothesis H_0: There exists a non-zero correlation between atleast two of the variables.

Alternate Hypothesis H_0: There is no correlation between any two variables.

We set $\alpha = 0.05$. The summary of the model is as follows:

	Estimate	Standard Error	t value	p value
(Intercept)	37.62967	2.35378	15.987	$< 2e-16$
Age	-0.07584	0.06962	-1.089	0.277
Male	-2.30043	3.37000	-0.683	0.495
Age:Male	0.10235	0.09403	1.088	0.277

Table 5.1: Summary of Regression Model

Conclusion: As the p-values corresponding to the variables are all greater than $\alpha = 0.05$, we infer that there isn't sufficient evidence to say that a non-zero correlation exists.

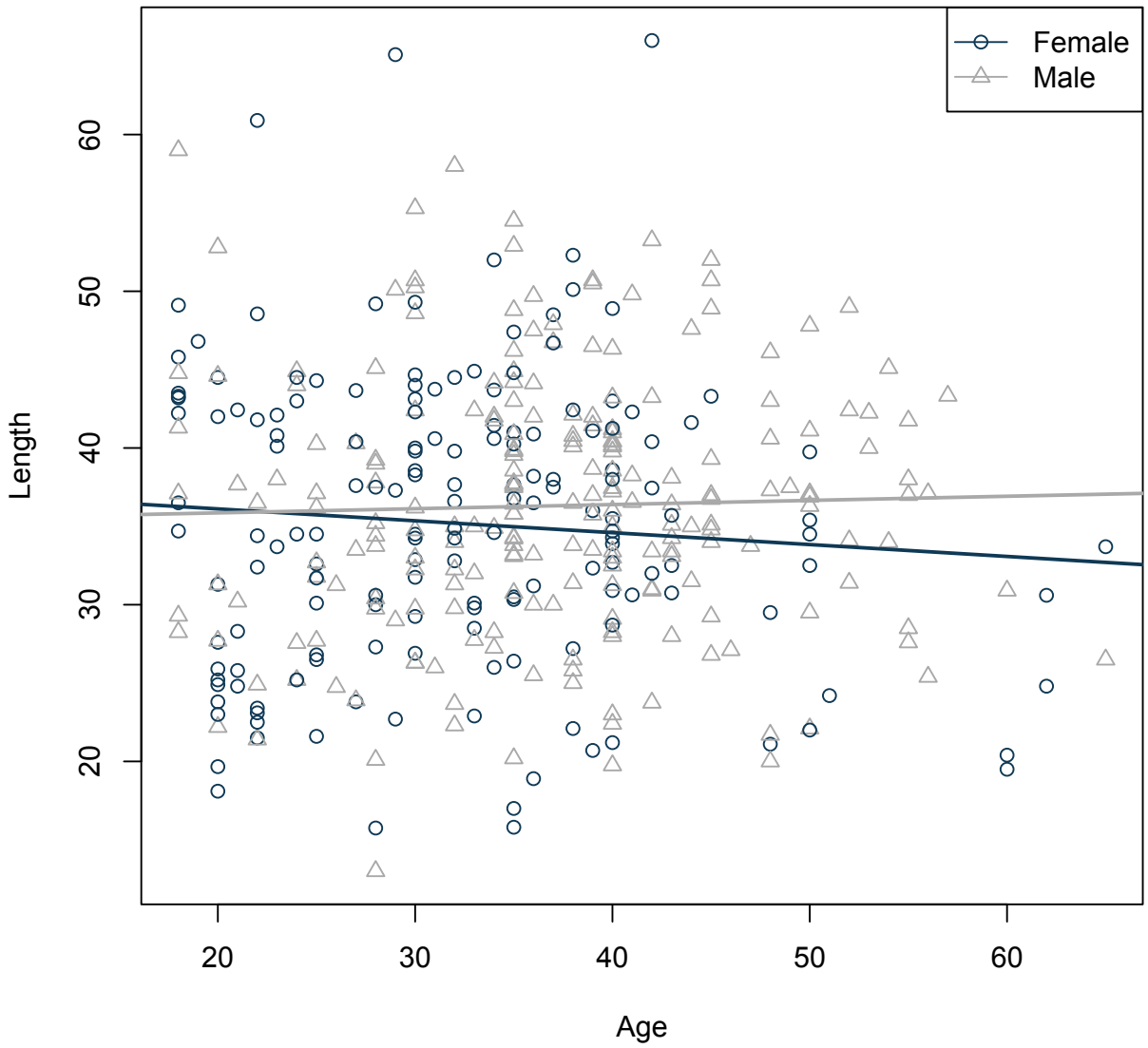


Figure 5.1: Interaction model plot

5.2 Does demography affect the average length of text messages significantly?

In order to determine whether there is a statistically significant difference between the average length of text messages of digitally semi-literate users of urban and rural demographies, we conducted the two tailed t-test. The formula for two tailed t-test is given by

$$t = \frac{(\overline{x_1} - \overline{x_2})}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$

where $\overline{x_1}$ is the mean of sample 1, $\overline{x_2}$ is the mean of sample 2, s_1 is the standard deviation of sample 1, s_2 is the standard deviation of sample 2, n_1 is the sample size of sample 1 and n_2 is the sample size of sample 2.

Here, sample 1 is digitally semi-literate users of the urban demography and sample 2 is digitally semi-literate users of the rural demography.

Our hypotheses:

Null Hypothesis H_0: There is no difference between the mean length of text messages between urban and rural digitally semi-literate populations of India

Alternate Hypothesis H_A: There is a difference between the mean length of text messages between urban and rural digitally semi-literate populations of India.

We set $\alpha = 0.05$. Now, the p-value can be determined using the equation

$$p = P(|T| \geqslant |t|)$$

where t is the t-value calculated in the previous formula and T is the t-distribution.

mean of x	mean of y	degrees of freedom	t-value	p-value
36.97468	34.79221	351.05	2.5123	0.01244

Table 5.2: x- rural, y- urban

Conclusion: As the p-value is less than $\alpha = 0.05$, we reject the null hypothesis and infer that there is a statistically significant difference between the mean length of text messages between urban and rural digitally semi-literate populations of India.

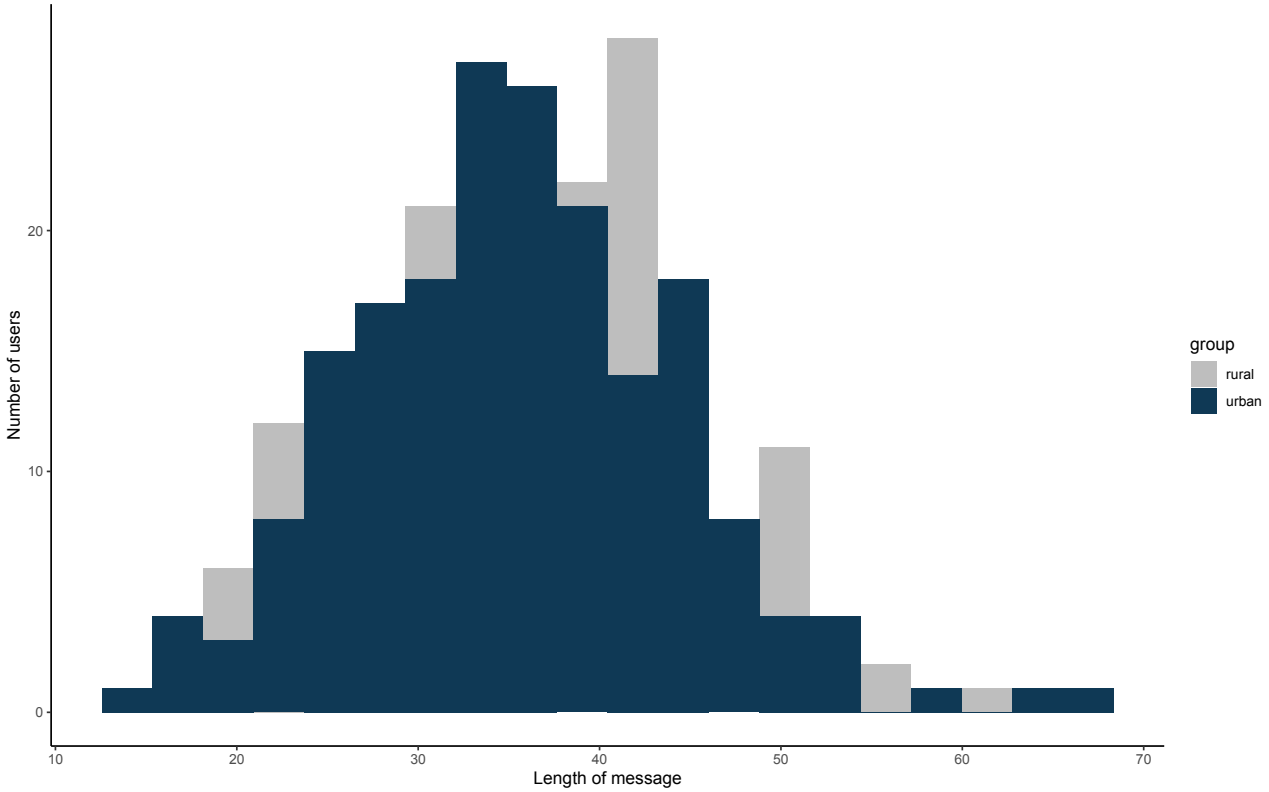


Figure 5.2: Overlapping histograms of average length of text messages corresponding to users from the urban and rural demographics

6. 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. M. Burch, et al., Prefix tag clouds, in: Proceedings of the 17th International Conference on Information Visualisation, IEEE, 2013.
6. 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).
7. David Dalpiaz, Applied Statistics with R, Chapter 11: Categorical Predictors and Interactions.
8. Satyam Kumar, How to avoid multicollinearity in Categorical Data, 2020