Digital Literacy and the Spread of Misinformation in Pakistan

Ayesha Ali and Ihsan Ayyub Qazi *

May 10, 2019

1 Introduction

This document contains a pre-analysis plan for evaluating two interventions designed to educate social media users with low digital literacy about misinformation (fake news) on social media in the city of Lahore, Pakistan. The plan is being written at a time when data collection has started, but cleaned data has not yet become available. Therefore, it will be used as a reference for the final data analysis. If needed, an addendum to the plan will be prepared after data collection is complete, in order to take into account any discrepancies in design and execution of the study. The rest of the document is structured as follows. Section 2 describes the motivation and context of the project, section 3 discusses the sampling and data collection, section 4 contains the details of the experimental design, and section 5 discusses the planned empirical strategy.

*Ali: Lahore University of Management Sciences, ayeshaali@lums.edu.pk. Qazi: Lahore University of Management Sciences, ihsan.qazi@lums.edu.pk. We thank XX for helpful comments, and Sara Obaid, Noor ul Islam and Noor Alam for project coordination and research assistance. We gratefully acknowledge funding from the Facebook Integrity Foundational Research Award carrying out this research. The study was approved by the IRB at Lahore University of Management Sciences (Protocol Number: LUMSIRB/03252019)

2 Motivation and Context

The increasing availability of low-cost mobile phones and mobile Internet access in emerging and developing markets has led to widespread use of social media platforms, making them an important source of news and place for social and political activity. This trend has brought many new users online, including those with limited exposure to technology. According to the "Digital in 2019" Report, there are an estimated 3.45 billion Internet users in Asia, Africa, Eastern Europe and Southern America. Out of this 2.81 billion people are connected to each other through social media platforms. Social media penetration in these regions has grown by 10.8% since 2018, while in South Asia alone it has increased by 19.7% over the same time period.

With such easy access to internet and social media, news can rapidly travel to millions of people in a short span of time. With the rising use of social media to share news and information, we have also observed the phenomenon of misinformation or fake news being disseminated on popular platforms such as Facebook, WhatsApp and Twitter. Allcott et al., (2017) find that around the time of US elections of 2016, fake news was widely shared on social media, and the average American adult was exposed to three fake news stories during this time. In a follow up paper, Allcott, Gentzkow and Yu (2018) show that the number of engagements on Facebook with news sites known to publish completely false content declined after 2016, due to policy and algorithmic reforms undertaken by Facebook. However, the number of engagements on Twitter increased. The number of engagements with partially trustworthy news sites has also increased on both social media platforms. Thus, fake news remains prevalent and is difficult to counter through regulation and content monitoring alone.

While it is difficult to find good estimates of the extent of misinformation on social media in developing countries, the recent crackdown on fake accounts suggest that it is indeed pervasive. For example, Facebook has reported taking down hundreds of fake accounts and suspicious pages in Russia, Iran, Brazil, Myanmar, India and Pakistan, especially around important events such as elections.[1]

[1] See for example,
`https://www.facebook.com/notes/mark-zuckerberg/preparing-for-elections/10156300047606634/`,
`https://www.dawn.com/news/1473284`.

The spread of misinformation can have serious negative consequences for the immediate recipients, communities, and the broader society. It may augment or create social, political, religious, and ideological differences. It may also affect outcomes such as political participation, voting behaviour, marginalization of victimized groups, and even lead to violence. Therefore, it is important to understand the ways through which we can reduce the distribution and virality of misinformation. In order to do so, we need to learn about the characteristics of social media users and how they interact with news received on social media.

We choose to focus our study on social media users in Pakistan, a large developing country where approximately 18 % of the population or 37 million people use social media. Pakistan has experienced an exponential increase in the usage of mobile phones due to the widespread availability of low cost smart phones and data plans. According to the Digital in Pakistan reports, social media penetration has grown rapidly from just 4% in 2013 to 18% in 2019. Facebook.com and WhatsApp.com are among the websites with the largest average monthly traffic in the country (Top Websites Ranking, 2019). Furthermore, a recent 2018 Gallup poll of a nationally representative sample of internet users shows that 24% use social media several times a day to access news, while another 24% use social media at least once a day to access news. Given, the importance of social media as a popular news source, it is highly likely that misinformation continues to touch many social media users in Pakistan.

We have previously carried out an online survey before 2018 Pakistani general election to study misinformation on social media. Our data obtained from 537 users, shows that 85% have believed a news that later turned out to be fake, and 30% have shared a news that later turned out to be fake. We also find that younger users and those who spent most time reading news on social media were more likely to believe fake news stories when shown fake news headlines. Those who said that they typically verify the source of news were less likely to believe fake news. Another finding which echoes the results of previous literature, is that ideologically aligned fake news was believed more by people (Allcott et. al, 2017; Pogorelskiy and Shum, 2018).

In this study, we will carry out a survey of social media users in 1,000 low to

middle income households drawn from a large urban center in Pakistan. We have chosen to focus on social media users in this segment, as many of them have low levels of digital literacy due to limited exposure to technology. Therefore, these users may be more vulnerable to believing and spreading fake news. Our survey will provide unique first-hand evidence on the trends in social media use, sharing of news, and spread of misinformation among low and middle income users in Pakistan. Using a list of actual true and fake news stories circulated on social media we will measure the extent to which users are able to correctly identify fake news. We also measure different ways users engage with the news including why they label an item as true or false, degree of emotional arousal, intent to verify and share on social media. Through the data collected from our baseline survey we can better understand the characteristics of social media users vulnerable to misinformation in developing countries.

We will also evaluate the effectiveness of two interventions for countering misinformation in a randomized control setting with 750 participants drawn from our baseline survey. The users in the first treatment group are shown a video which educates users about common features of misinformation. The users in the second treatment group are first shown the video and then given specific feedback about their own responses to news stories shown at baseline. The personal feedback points out the features of each fake news item that can enable the user to identify the news as fake. We measure the effectiveness of our interventions through an endline test by asking users about their beliefs about and engagement with a second set of news stories circulated on social media.

3 Sampling and Data Collection

3.1 Sampling

We implement our survey and randomized control trial (RCT) in the city of Lahore in Pakistan which has a population of 11 million. In order to draw our sample of low and middle income households in the city, we rely on population density. The pattern

of urban development in Lahore is horizontal, and dwelling size is inversely correlated with income levels, therefore, we use population density as a proxy for income levels. Specifically, we use the AsiaPop satellite data which provides population counts at a spatial resolution of 100 m by 100 m, to identify the low and middle income areas of Lahore from where we draw our sample. We have selected the areas of Lahore covered by seven national assembly constituencies, accounting for 35% of the city's total population. These areas cover the older parts of the city where the median population density is 109 persons per 100 m by 100 m grid.[2] As a comparison, the median population density in the outlying areas not covered by our sample is 28 persons.

We draw a random sample of 200 grids from the selected areas. In order to initiate the data collection within the chosen grids, we randomly drop a point (x and y coordinate) within the grid. The enumerators arrive at the point and use the left hand rule to survey five households within each grid. The definition of social media user for our project is that the respondent must be above 18 years of age and use at least Facebook or WhatsApp. Figure 1 shows the distribution of the 200 grid points to be used for baseline data collection. For the RCT, we randomly assign 50 grid points each to control, treatment 1, and treatment 2 status. One respondent per household will be randomly drawn to be part of the RCT from these chosen grid points and assigned to control, treatment 1 and treatment 2 depending on the grid assignment status. Figure 2 shows the randomization scheme. In order to ensure gender balance, for each grid, we have also pre-specified the number of male and female participants to be drawn for the RCT.[3] Therefore, at the end of baseline, we will have a sample of 750 social media users drawn from 750 unique households, who become part of our experiment and will be visited again at endline. The enumerators use tablets to enter data for all questions. The news items, video, and feedback is

[2]Urban development in Lahore has led to new housing schemes being developed outside the older parts of the city, and the richer and affluent households have moved into these new housing schemes.

[3]The enumerator is informed by the survey application, at the start of the baseline survey, whether a male or female respondent is to be selected for the experiment. If the required gender is not present, the household is replaced.

also shown on the tablets. Furthermore, the enumerators carry printed version of the news items and ear phones to listen to the video.

3.2 Baseline Survey

The baseline survey is to be answered by all social media users present in the household at the time of the survey. The survey consists of the following sections:

1. Household and demographic characteristics including household size, monthly expenditures, number of social media users, age, gender, and highest education level.

2. Social media use including phone ownership and sharing, social media applications used, number of hours spent per day using Facebook and WhatsApp, size and composition of the social media network, top three news sources, and sharing of news on social media.

3. Digital literacy assessed by self-reported ability to turn on mobile data, wifi, use google, use social media without assistance, read text on social media, and read English and Urdu on social media. We also ask about knowledge and use of common features of WhatsApp and Facebook. We also ask if they have heard of the term "fake news" and if they ever believed a news that later turned out to be false.

3.3 News items

After the baseline survey is completed, we will show each social media user a series of news items. We have compiled a list of fourteen recent popular true and fake news stories, covering politics, foreign policy, current events and general interest topics. Out of these six are true news stories and nine are fake news stories. We have also created three placebo news items which are used to measure and control for false recall. The news stories are shown in the form of screenshots of messages, posts, or tweets, similar to how users would typically receive news on social media. Some

of the news stories are in English only, while others are completely or partially in Urdu. We have also prepared audio recording of the news item, as we expect some participants will are not able to read text and/or read English.

At baseline, the respondents are shown a set of seven news items, out of which three are true, three are fake and one is a placebo. The respondent is first asked to carefully view a print out of the news item, and then asked to listen to the audio recording. After having viewed and heard the news, we ask the following questions:

- Whether they recall seeing the news before, and if so at that time, whether they thought it was true or false.
- Right now if they would consider the news to be true or false.
- How they felt upon hearing the news, ranging from Very Positive, Positive, Neutral, Negative, to Very Negative.
- Whether they would verify the news, by asking friends or family, or searching online.
- Why did they consider the news to be true or false right now.
- Whether they will share the news on social media.

4 Experiment Design

4.1 Treatment 1: Video only

At the end of the baseline survey, participants in the treatment 1 group watch a short three minutes video about fake news. The video explains what is fake news, why is it spread, negative consequences of fake news, and then focuses on the following three ways to identify fake news:

1. Source problem: Untrustworthy or missing source, for example unknown author, unverified account, or absence of an authentic link supporting the news.

2. Quality problem: Poor quality of news, that can be spotted visually, for example, altered images or videos, informal and/or incorrect language, or excessive use of hashtags and emoticons.

3. Biasedness: Content or language that reflects bias, often designed to provoke extreme emotions.

4.2 Treatment 2: Video plus personal feedback

The participants in the treatment 2 group first view the video, and then receive personal feedback based on how successful they were in identifying the three fake news items shown at baseline. We have selected one fake news item each for the three problems highlighted in the video, namely, source problem, quality problem and biasedness. The personal feedback for each news item consists of two components:

1. A score of 0 or 1, based on whether the participant correctly identified the news items as false. Figure 4 shows the scorecard which is read out to the participant by the enumerator.

2. Pointers on how to identify that the news is fake by highlighting the main problem (source, quality, or biasedness) while also mentioning secondary problems that will help the user identify the news as fake.

The personal feedback will be delivered verbally by the enumerators using feedback screen which appears on the tablet. We have prepared standardized score cards and feedback cards for each of the three fake news items included in the baseline, which will provide the score and the pointers. The feedback cards aim to be encouraging and informational, ensuring that the users understand why they got something right or wrong.

4.3 Endline

The endline survey for each participant is conducted on the third day after the baseline survey. The endline survey consists of a series of 11 news items (6 fake, 3

true and 2 placebo) shown in the same manner to the participant as before. The news items have the same characteristics as the news items in the baseline. For each news item, we repeat the same questions asked as before.

Figure 3 summarizes the timeline of events, which begins with the baseline visit, the administration of the baseline survey and news items, followed by the intervention for the treatment groups, after which the baseline visit ends. The users in the RCT are visited on the third day after the baseline and once more shown and asked questions about a second set of news items, after which the endline visit ends.

5 Data and Variables

5.1 Primary Outcomes

The primary outcomes of interest for this project is the ability to correctly identify a news item as false or true at endline after receiving the intervention relative to the control group. The variable takes the value 1 if the user correctly identifies fake and placebo news items as false, and true news as items as true, and 0 otherwise.

Furthermore, we will also calculate an overall score based on the proportion of news items correctly at baseline and endline.

5.2 Secondary Outcomes

Additionally, we will also examine the effect of the intervention on how respondents engage with the news items. Our intervention is designed to educate users about common features of fake news and therefore, we can study whether users in the treatment group are more careful in checking extreme emotions, verifying news, and sharing it on social media when exposed to fake news.

- Using the response to the question on why the user identified the news as true or false, we will create a variable that takes the value 1 if the user mentions any of the features of fake news discussed in the video and personal feedback, and 0 otherwise.

- Using the response to the question on how the user felt after hearing the news, we will create a variable that takes the value 0 if the user felt "extremely positive" or "extremely negative" and 1 otherwise.
- Using the response to the question on what the user will do after receiving the news, we will create a variable that takes the value 0 if user answers "Do nothing", 0.5 if the user answers "Ask friends and family", and 1 if the user answers "Search online"
- Using the response to the question on whether the user will share the news on social media, we will create a variable that takes the value 0 if the user shares a false news item and 1 otherwise.

5.3 Covariates

The following user characteristics measured at baseline will be included in the main regression specification evaluating the effect of our interventions on the outcomes of interest.

- Demographic characteristics: gender, age, and education
- Digital literacy indicators: Indicator variables measuring the ability to connect to WiFi, turn on mobile data, search on google, use social media without assistance, and read both languages on social media.
- Application knowledge and use: Fraction of common features of WhatsApp and fraction of common features of Facebook known and used by user.
- Active social media use: constructed on a scale of 1 to 3, based on the number of regular social media activities, namely, viewing content, sharing content, and creating content.
- Social media hours: the sum of hours spent per day using social media
- Social media primary news source: indicator equal to one if social media is the primary news source

- Number of friends on Facebook
- Number of members in top three WhatsApp groups.
- Fraction of social network that is not classified as family or friends on Facebook and WhatsApp groups
- Sharing news on social media: equals 1 if "All the time", 0.75 if "Often", 0.5 if "Sometimes", 0.25 if "Rarely" and 0 if Never. We take the average of the sharing propensity if both WhatsApp and Facebook used.
- False recall at baseline: indicator equal to one if the user falsely recalls having seen placebo news item
- Baseline score: fraction of correctly identified news items at baseline

6 Empirical Strategy

6.1 Average treatment effect

We will evaluate the effect of our educating users about how to identify fake news in the treated group relative to the control group using the outcomes measured at endline. The main specification that we will use to identify the average treatment effect of the intervention is as follows:

$$Y_{ik} = \alpha_0 + \alpha_1 X_i + \beta_1 T_1 + \beta_2 T_2 + \epsilon_{ik} \tag{1}$$

where Y_{ik} is the outcome of user i for news item k shown at endline, X_i are user characteristics captured at baseline, including the user's baseline score measured as the fraction of correctly identified news items shown at baseline, and T_1 is the treatment dummy for receiving the video intervention while, T_2 is the treatment dummy for receiving the video and personal feedback intervention. β_1 is the average treatment effect of showing the video and β_2 is the average treatment effect of showing the video and giving personal feedback.

Since we observe the responses to the news items for each user before and after the intervention, we can also look at the within user change in ability to correctly identify news as true or false. The outcome variable is the change in the score based on total number of correctly identified news items. This specification would account for any fixed unobserved characteristics of users that we are unable to control for and are unbalanced across the treatment and control groups.

$$\Delta Y_i = \delta + \beta_1 T_1 + \beta_2 T_2 + e_i \tag{2}$$

6.2 Sub-group Analysis

We expect the impact of the intervention to be larger for users who are more digitally literate. For both equations (1) and (2) we can estimate the treatment effect separately using an interaction between treatment dummies and the digital literacy score calculated as the sum of digital literacy indicators. We will also examine the existence of differential effects by using information on individual dimensions of the digital literacy score such as the ability to read both languages on social media and the ability to use social media without assistance. Furthermore, we will also examine the existence of heterogeneous treatment effects by the following moderators:

- Gender of the users
- Age of the users
- Social media hours
- Active social media use
- Application knowledge and use
- For equation (1) we will also present separate estimates after splitting our sample into true and false news items.

- For equation (1) in the sample of false news items, we will also test for the presence of heterogeneity in treatment effects by news type, i.e. source problem, quality problem, and biasedness.

- For equation (2), using the sample of fake news, we will also present separate estimates of the change in score by news type, i.e. source problem, quality problem, and biasedness

References

[1] Allcott, Hunt, Mathew Gentzkow (2017)"Social media and fake news in the 2016 election". *Journal of Economic Perspectives*, 31(2): 211236.

[2] Allcott, Hunt, Mathew Gentzkow, Chuan Yu (2018) "Trends in the Diffusion of Fake News on Social Media", Available at: http://web.stanford.edu/ gentzkow/research/fake-news-trends.pdf

[3] Pogorelskiy, Kirill, Mathew Shum, "News We Like to Share: How News Sharing on Social Networks Influences Voting Outcomes", (April 11, 2018). `https://ssrn.com/abstract=2972231`

[4] Digital in 2019, Accessed May, 2019, `https://wearesocial.com/global-digital-report-2019`

[5] Digital in 2018, Accessed May, 2019, `https://wearesocial.com/global-digital-report-2018`

[6] Digital 2019: Pakistan, Accessed May, 2019, `https://datareportal.com/reports/digital-2019-pakistan`

[7] Top Websites Ranking, Accessed March, 2019, `https://www.similarweb.com/top-websites/pakistan`

7 Figures

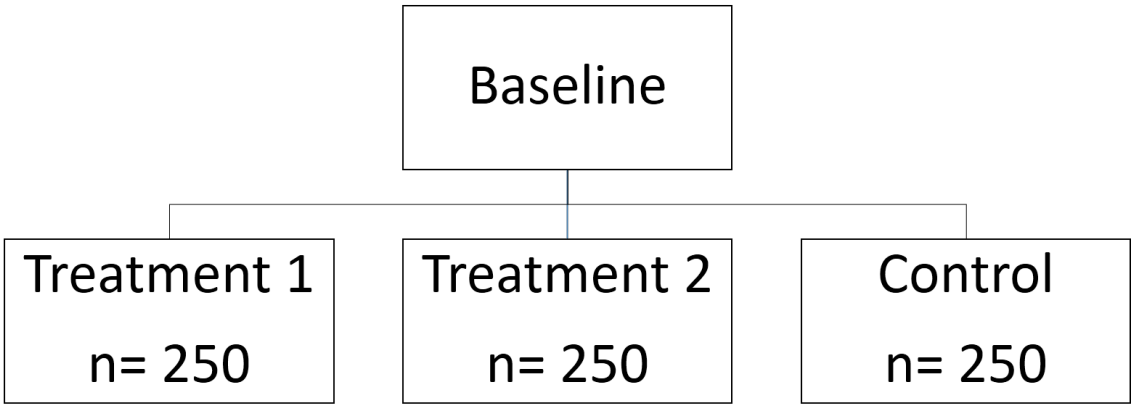

Figure 1: Randomization Scheme

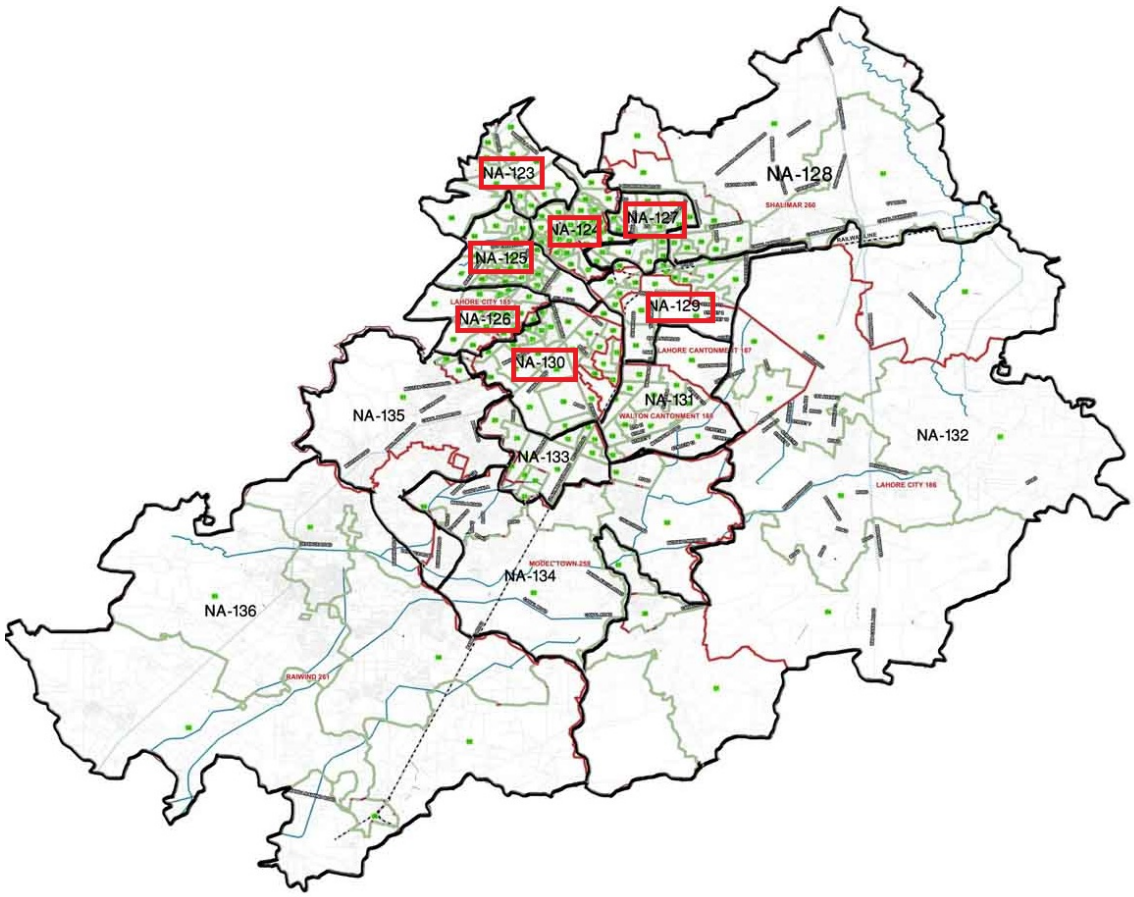

Figure 2: The map shows the city of Lahore.The constituencies from which the sample is drawn are highlighted in red.

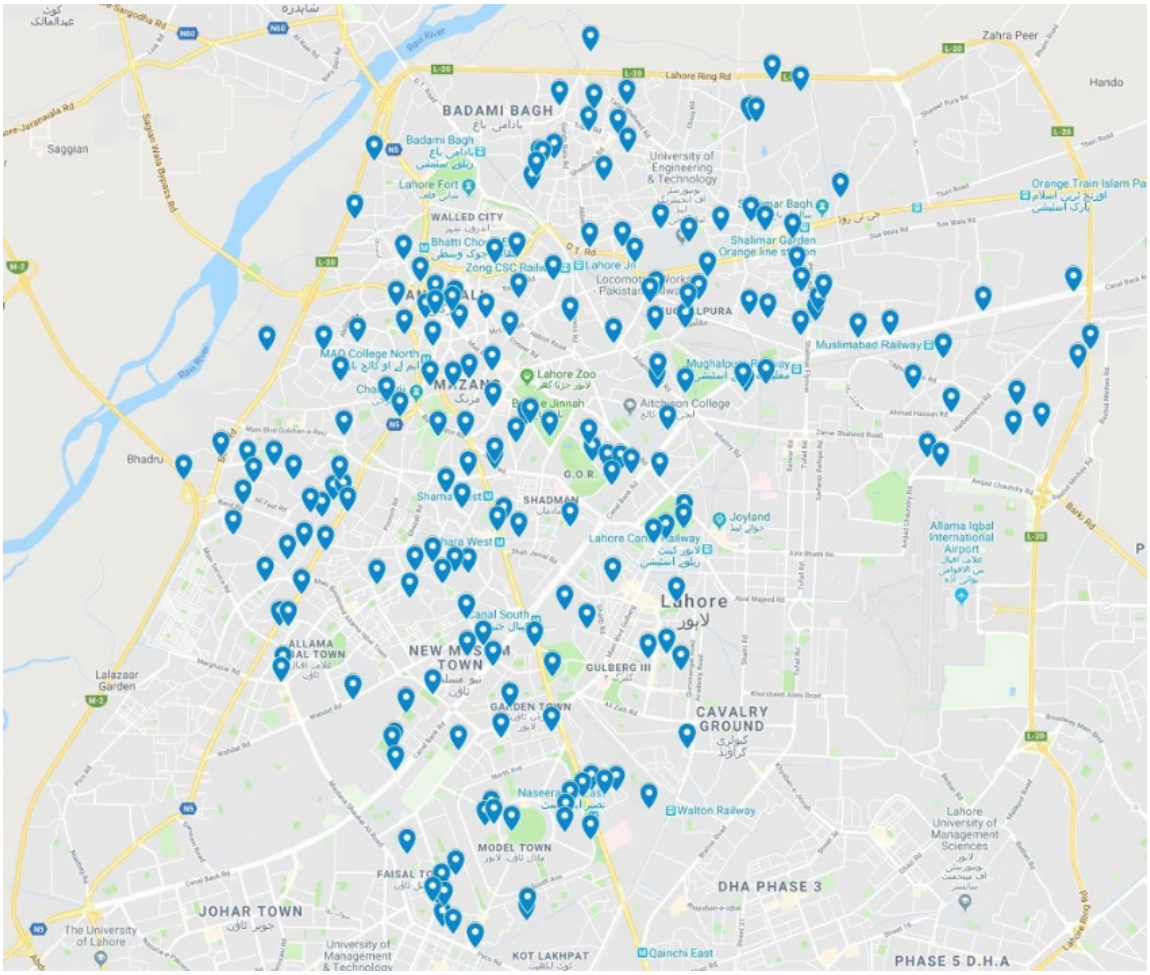

Figure 3: The map shows the grid points that are randomly drawn from the selected areas.

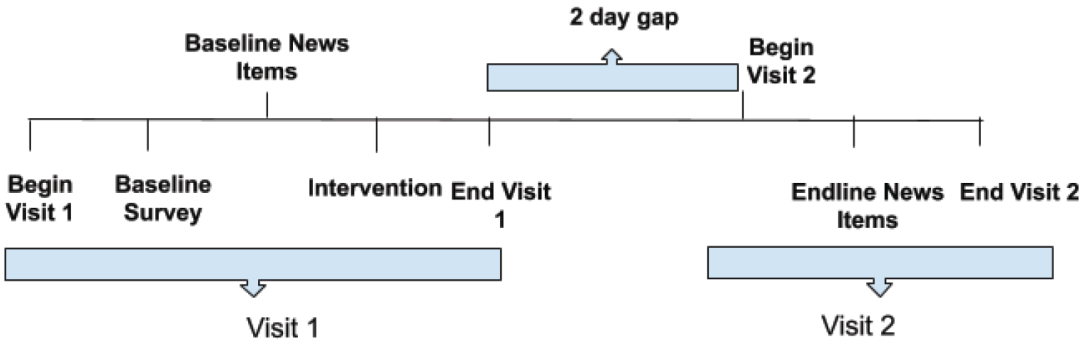

Figure 4: Timeline of Events