International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-7, Issue-6, March 2019

Identifying Trends in Facebook Usage: A Visual Approach

R. Sethuraman, Krishna Chaitanya Reddy V, M. Gautham Veer, R. Subhashini

Abstract*: A sharp surge in the usage of social media platforms like Facebook, Twitter, Snapchat etc., makes study on social phenomenon worthwhile. This analysis is an effort to understand the usage trends of people belonging to different age groups in Facebook. The main steps involving our analysis are (a) preprocessing the data into appropriate format and shape for easy analysis and (b) visualizing the patterns observed in the dataset. Visual approach is considered so that the representation of complex social data is simplified, also when something is visualized, we tend to identify the patterns just by looking at them. However, this visualization process takes a lot of time, but worth the time spent. Common visualization techniques include Bar graphs, Histograms, Distplots (Distribution Plots). The preprocessing stage which takes place before the visualization phase is a challenge in itself. Once we get the raw data from the source, we then look for the shape of dataset, null values, co-related features etc. For this analysis we used only the cleaning of null values in preprocessing. We normally have a preconception that a certain age group of people will use social media more than that of other age groups. We will explore all such intricacies in this paper. This paper also explores how various features and parameters affect the trends in dataset.

***Index Terms*: Data Preprocessing, Data Visualization, Distplots, Heatmaps.**

I. INTRODUCTION

The sudden explosion of usage of social media platforms have opened a wide variety of opportunities in term of understanding the behavior of users, providing them with the best possible services according to their liking and also curbing any acts of violence. This enabled the study of social behavior of users. The purpose of this analysis is to find out weather these so called 'paradoxes' are true or not. We shall proceed by collecting required data for analysis.

II. METHODOLOGY

A. Data Collection

Data about Facebook is collected from Kaggle [1]. Kaggle is a place where many data scientists and various companies post their datasets. This dataset consists of 15 features and 99003 entries. There are 175 null values in gender and 2 in no. of days of usage features.

I. Features and their null values

Features	Description	Null Values
userid	User ID	0
age	Age of User	0
dob_day	Date of Birth (Day)	0
dob_year	Date of Birth (Year)	0
dob_month	Date of Birth (Month)	0
gender	Gender of User	175
no_of_days_of_usage	Days of Usage	2
friend_count	Number of Friends	0
friend_request_sent	Friend requests sent by user	0
likes	Likes made by user on a post	0
likes_received	Likes received by user	0
mobile_likes	Likes made by user	0
mobile_received	Likes received on mobile	0
pages_liked	Pages liked by user	0
pages_likes_received	User pages likes	0

B. Data Preprocessing

The preprocessing step is the step that takes most time in a data science project. A typical preprocessing method will involve cleaning of null values, changing categorical features to numeric or binary (0 or 1). Generally, removal of null values can be done in three ways:

1. Removing an entire row in which a null value is present
2. Filling the null value with a mean, median or mode of that column or any other related column.
3. Filling the null value using methods like back filing and forward filing.

Revised Manuscript Received on 30 March 2019.
***** Correspondence Author
R. Sethuraman*, School of Computing, Sathyabama Institute of Science and Technology, Chennai, India.
Krishna Chaitanya Reddy V, School of Computing, Sathyabama Institute of Science and Technology, Chennai, India.
M. Gautham Veer, School of Computing, Sathyabama Institute of Science and Technology, Chennai, India.
R. Subhashini, School of Computing, Sathyabama Institute of Science and Technology, Chennai, India.

The first method causes data loss. Suppose we have a dataset of 100 records and of which 50 of them contains null value, removing all the 50 records will lead to a huge data loss. However, this step is useful in cases where the data loss is minimal. The second method can be used in most cases. We use the third method in this dataset as it is relevant to the feature in which the null values are present. In the dataset we have 175 null values in gender column which is a categorical feature which has only two values – 'M' and 'F' corresponding to Male and Female respectively. These values however can be encoded to 0 and 1. Encoding these values to binary helps us in model generation.

The gender field null values are handled by filling that particular column with a "Back Filling" technique, which is available as an attribute for fillna() method [] in Pandas Library. Pandas is a widely used data wrangling and preprocessing package. The 'no_of_days_of_usage' field has two null values. These are handled by filling the column with mean of all the values in that particular column.

C. Visualization

After the dataset is cleaned of all the null values, a heatmap [2] is generated to understand correlated features in dataset. This is done by using a library called seaborn. There are five features that are highly correlated with each other. Heatmap is a visualization type generated used to check if there are any correlated features.

II.Correlated Features

Sno.	Correlated Features
1	friend_count and friend_request_sent
2	likes and mobile_likes
3	likes_received and mobile_likes_received
4	likes_received and pages_likes_received
5	mobile_likes_received and pages_likes_received

Heatmaps can be generated only for features having numerical values. They can also be generated in various colors. Once the heatmap analysis is done, KDE plots [3] which are also known as Kernel Density Estimation Plots are made to understand the distribution of each feature. For this, the features used are:

1. age
2. gender
3. no_of_days_of_usage
4. friend_count
5. likes
6. pages_liked

Given below is a KDE plot for age distribution by gender.

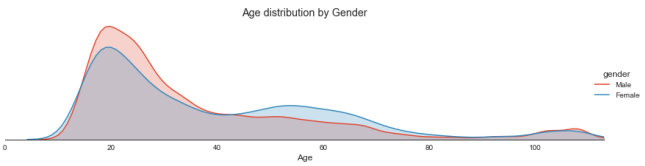

Fig 1: Age Distribution by Gender

These plots enable us to understand how the trends are and these plots are the heart of study. Five of these plots are made with various features. We used only two of them in this paper for understanding. The distplots are inbuilt class in seaborn package. The distplot shows distribution of a feature with respect to another. However, we cannot say that the particular plot generated follows Gaussian Distribution or Normal Distribution.

Finally, boxplots are made to understand few distributions of complex features. The boxplots are not given here due to the fact that the image becomes complex to interpret. A typical boxplot has 5 important features.

1. Outliers
2. Inter Quartile Range (IQR)
3. First Quartile
4. Second Quartile
5. Third Quartile

Outliers are the feature whose values are abnormally large compared to its average value. In order to understand the quartiles, consider a box with a horizontal line at its center and the box represents 100. The first quartile is the median of values between 0-50. Second quartile is nothing but the median. Third quartile is the median of values between 50-100 or top 75% values. The boxplot generated for this analysis returned with a lot of outliers and peculiar findings.

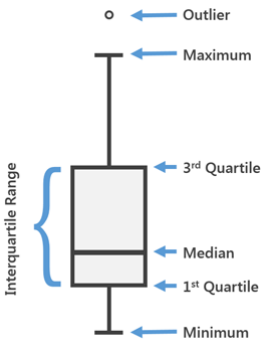

Fig 2: Boxplot [7]

III. Related Work

This paper is heavily inspired by visualization paper [4]. The base paper mainly talks about visualizations and fitting appropriate model that performs better. However, the scope of this paper is only to visualize the trends, this paper can also be extended to do the model generation too. It is important to visualize anything before we dive deep into it. This gives us an overall big picture as to how the data is distributed and how the features are correlated and stuff. The base paper uses an act of violence dataset which contains about 300 records. These records are cleaned for null values, and visualized. Once the visualization is completed and the dataset is clean, few models like Support Vector Machine are fitted onto the processed dataset and their accuracies are reported. However, in the base paper, the main problem is the way in which null values are handled and the size of dataset.

The null values are removed by removing the entire row which causes data loss. Since the dataset size is 300, there might be a huge bias towards one feature. In order to overcome this bias, feature selection is employed.

IV. RESULTS

The results are staggering as we have uncovered some pretty peculiar findings from the graphs. The results are as follows:

1. Friend Count and Usage fields have a lot of outliers which needs to be addressed.
2. As suspected the majority of users lie between 15 to 40 years of age with a few exceptions falling above 93. This might be due to the Facebook policy of not deleting accounts up to 90 days [6].
3. Usage is high between 0 to 2000 days, with majority of people using for 500 days.
4. Friend Count is slightly higher in case of female with value around 300.
5. Likes by female is skyrocketing around 0 to 3000 whereas males supposedly have less likes.

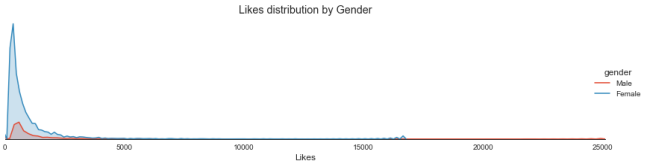

Fig 3: Likes Distribution

6. Similar trends can be seen in case of page likes as mentioned in point 5.
7. Surprisingly the usage with respect to age bins is having a linear relation, which is a bit suspicious as a person of age 93 will probably not use Facebook much logically. However, this can be a discrepancy in the dataset.
8. There are many correlated features.

V. FUTURE WORK

In future works, we can work on the result and enhance them even more. We can create a machine learning model, probably clustering model which will cluster according to age groups and give us better understanding and predict how many new users use the social media. We can also add more data regarding the topics the users are searching and probably use a technique called Topic Modelling to identify acts of violence.

REFERENCES

1. Sheena Batra, "Facebook Data" in Kaggle. Available: https://www.kaggle.com/sheenabatra/facebook-data
2. Shilin Zhao, Yan Guo, Quanhu Sheng and Yu Shyr "*Advanced Heat Map and Cluster Analysis*", *Hindawi 2014*.
3. Pieter Vermeesch, *On Visualization of Detrital Age Distributions"*. Elsevier.
4. Mahdi Hashemi and Margaret Hall, "*Visualization, Feature Selection, Machine Learning: Identifying the Responsible Group for Extreme Acts of Violence*", IEEE 2017
5. Alexandre Perrot, Romain Boruqui, Nicolas Hanusse and David Auber, "Heat Pipe: High Throughput, Low Latency Big Data Heatmap with Spark Streaming" *ICIV 2017*.
6. Facebook User Account delete policy. Available: https://www.facebook.com/help/359046244166395/
7. Boxplot. Available: https://pro.arcgis.com/en/pro-app/help/analysis/geoprocessing/charts/box-plot.htm

AUTHORS PROFILE

R. Sethuraman is working as an Assistant Professor and Data Scientist. He has 9 years of IT experience and responsible for the New Product/Technology Introduction, Service Oriented Architecture, Building Next Generation products and applications. Played the role as Technology Analyst includes designing software, web solution, project management and consulting and maintaining various projects.

Krishna Chaitanya Reddy V is a Computer Science student and IBM certified Data Science Professional with a demonstrated experience of 2 years. He has worked on many problems and real case scenarios involving data science and analysis. He has also conducted many seminars in the field of Machine Learning and Data Science.

M. Gautham Veer is a Computer Science student and IBM certified Data Science Professional. He has worked on many Machine Learning algorithms like SVM, Regression Models for the past 2 years. He has also done many projects on the topics like Data Science.

Dr. R. Subhashini is working as Professor in School of Computing at Sathyabama Institute of Science and Technology, Chennai. Her research interests are Machine Learning, Cloud Computing, Deep Learning, Data Analytics and Big Data. She is the project coordinator for DST-FIST Cloud Computing Lab. She has published more than 40 research publications in journals and conferences. She is the resource person in various guest talks and workshops. She has organized various FDP, workshops, National and International Conferences.