TIME SERIES FORECASTING OF BIG DATA: A STUDY ON THE FACEBOOK USER BASE

N. Anuradha, Roopini. J
[1]*Assistant Professor, Department of Computer Science, KRUPANIDHI DEGREE COLLEGE Bangalore, Karnataka, India*
[2]*Assistant Professor, Department of Computer Science, KRUPANIDHI DEGREE COLLEGE Bangalore, Karnataka, India*

Abstract: In the field of Information Technology and social media, big data analytics is stirring a big revolution. The role of big data analytics is steadily gaining ground in strategic decision making. Today, a huge repository of terabytes of data is generated each day from modern information systems such as IoT, cloud computing, and social media. Analysis of such massive data requires tremendous efforts at multiple levels to extract knowledge for decision making. Big data analytics is the process of examining such large data sets to identify underlying patterns.

This research paper makes an attempt to explore an aspect of big data analysis. This study is concentrated in the area of social media and aims to forecast the social media usage pattern by foraying into one aspect of big data analytics, namely Predictive Analytics and its real time implementation. The basic objective of this paper is to predict the future trend of social media usage pattern with the help of past observed values.

Keywords: *Social Media, Big Data Analytics, Predictive Analytics, Time Series Forecasting, Active users.*

I. INTRODUCTION

Analytics refers to the systematic computational analysis of data or statistics. Big data analytics scrutinizes enormous amount of data to reveal hidden patterns, relationships, and other insights. It allows organizations to harness the available data and apply it to identify new opportunities. Analysts working with big data basically wish to harness the knowledge that comes from analyzing the data (A. Banumathi and A. Aloysius, 2017). Big data offers four types of analysis, namely, Prescriptive Analytics, Diagnostic Analytics, Descriptive Analytics, and Predictive Analytics.

Predictive Analytics, which is used in this study, entails forecasting the future possibilities based on past data. It uses several techniques like machine learning, artificial intelligence, data mining, and statistics to predict future trends. In this paper, we are focusing on Time Series Forecasting (TSF) which is one of the techniques of Data Mining. TSF involves the use of a model to predict future values based on the past observed values in order to extract meaningful statistics. It is generally used for non-stationary data and in this paper it is used to study the number of active users of one of the social media sites i.e.; Facebook to predict a future trend.

Recent research in the field of Computational Social Science (Cioffi- Revilla, 2013; Conte et al., 2012; Lazer et al.,2009) has shown how data resulting from the widespread adoption and use of social media channels such as Facebook and Twitter can be used to predict outcomes such as Hollywood movie revenues (Asur and Huberman, 2010), Apple iPhone sales (Lassen, Madsen and Vatrapu, 2014), seasonal moods (Golder and Macy, 2011), and epidemic outbreaks (Chunara, Andrews, and Brownstein, 2012). The underlying assumptions for this research stream on predictive analytics with social media data (Evangelos et al., 2013) are that social media actions such as tweeting, liking, commenting, and rating are proxies for user/consumer's attention to a particular object/ product and that the shared digital artefact that is persistent can create social influence (Vatrapu et al., 2015).

SOCIAL MEDIA

Social media is a disputed subject in today's society with its multihued implications on society. The popularity of social media sites like Twitter, Instagram, Facebook, and YouTube is rising exponentially and such platforms play a crucial role in social networking across the globe. Social media is perceived by consumers as a more trustworthy source of information regarding products and services than corporate sponsored communications transmitted via the traditional elements of the promotion mix (W. Glynn Mangold and David J. Faulds 2009; Foux, 2006).

Facebook is a social media site founded by Mark Zuckerberg and his Harvard peers in February 2004. It was conceived as a forum for connecting people and facilitating social networking. As compared to telephones and e-mails, the usage of Facebook as an interactive online networking system has increased drastically. Hampton, Goulet, Rainie, and Purcell (2011) found that Facebook users reported greater levels of support (emotional, instrumental, and companionship) than non-Facebook users.

The empirical evidence suggests that individual differences were related to new media usage (Pornsakulvanich and Dumrongsiri, 2009). In addition, personality traits and skills were related to Facebook usage patterns (Parks-Leduc, Pattie, Pargas, and Eliason, 2014), and the amount of time spent on Facebook affected social relationships (Liu and Yu, 2013).

II. OBJECTIVES

- To analyze the number of active users on Facebook.
- To predict the number of Facebook users in the future years.

III. METHODOLOGY

To predict something in the future we use Time Series Forecast (TSF); as the name suggests, it is nothing but forecasting. In machine learning, we cannot predict the future value but using TSF we can predict future values. Forecasting delivers relevant and consistent information about past and present events and the likely future events. This is indispensable for sound planning and desired outcome (Neelam Peters, Aakanksha S. Choubey, 2016). A time series is a sequence of data being recorded at specific time intervals. When the outcome (independent variable) is dependent on time, in such scenarios we use TSF where data points (past values) are analyzed to forecast future outcomes.

The four components of TSF are:

- Trend: Trend is the increase or decrease in the series over a period of time. It persists over a long period of time.
- Seasonality: It is the regular pattern of up and down fluctuations. It is short term variations occurring due to seasonal factors.
- Cyclicity: It is the medium term variation caused by circumstances which repeat in irregular intervals.
- Irregularity: It refers to variations which occur due to unpredictable factors and also do not repeat in particular patterns.

In order to perform time series analysis data should be stationary because non-stationary time series has trend and seasonality components that will affect the forecasting.

The factors that determine stationary of time series are:

- Mean: Mean is constant with time
- Variance: The variance of the series should not be a function of time
- Covariance: The covariance of the i^{th} term and the $(i+m)^{th}$ term should not be a function of time

Here, we are using the moving average technique to get the overall idea of the trends in a data set; it is an average of any subset of numbers.

IV. DEMONSTRATION

The analysis and the prediction of data were carried out using Time Series Forecast (TSF). We started by collecting historical data of the active users (https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/). Here, we take into account the data for four years (2015 to 2018) and we use it to forecast the number of users in the next three years using the moving average method. We are considering primarily the three components those are trend, seasonality and irregularity. The following are the steps to analyze and predict the values:

1. Draw the table with the time code (t), year, quarter (Q) and the number of active users in billions (Y_t).
2. Calculate the moving average (MA) by taking the average of all the four quarters and place the value in the third quarter. Then, we take the average of next four values and place the value in the fourth quarter and so on. Since, these values are not centered we need to calculate the Centered Moving Average (CMA).
3. Calculate the CMA by taking the average of two numbers and place the value in the third quarter. Then we take the average of next two values and place in the fourth quarter and so on. With this calculation we smoothen out the data. We consider these values as baseline.
4. In order to come up with the prediction, we use multiplicative model, where, time series value at time is $S_t*I_t*T_t$.
5. So, we calculate seasonal and irregular component by $\mathbf{S_t, I_t=Y_t/CMA.}$
6. To find only the seasonal component, the logic is to average each seasonal irregular component combined for each quarter and that will help get rid of any irregularity, which means we take the average of all the first quarters and place it in the first quarter, then, take the average of all second quarters and place it in the second quarter and so on.

Quarter	St
1	1.002
2	1.000
3	1.000
4	0.999

7. Place these all the quarter values in the seasonal component column for all the years.
8. Now, the next step to do deseasonalize the data, which can be calculated using $\mathbf{Y_t / S_t}$
9. Now, we find the trend component at time t ($\mathbf{T_t}$). In order to this, we have to run simple linear regression (SLR) using the deseasonalize data and t. To do this, we need to perform the following actions in MS-Excel:
    a. Go to File – Options - Addins – go.
    b. Check Analysis2 pack – ok

c. Go to Data and we see Data Analysis in available.
d. Click on Data Analysis – select Regression – ok
e. Select Y variable as deseasonalize data and X variable as t data with titles.
f. Click on labels
g. Select the output range as source where we want to put the data.

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	1.309843197	1.30984	3617.12359	2.65748E-18
Residual	14	0.00506972	0.00036		
Total	15	1.314912917			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	1.358888198	0.009979168	136.172	2.9149E-23	1.337485012	1.380291384	1.337485012	1.380291384
t	0.062068349	0.001032021	60.1425	2.6575E-18	0.059854885	0.064281814	0.059854885	0.064281814

10. Here, we find the coefficients which are highlighted and these are Y (Intercept) and X (slope) variables. These values are used to calculate trend.
11. Now, we calculate trend by doing: intercept value + slope * time code (t).
12. Now, we forecast the values by using $S_t * T_t$.

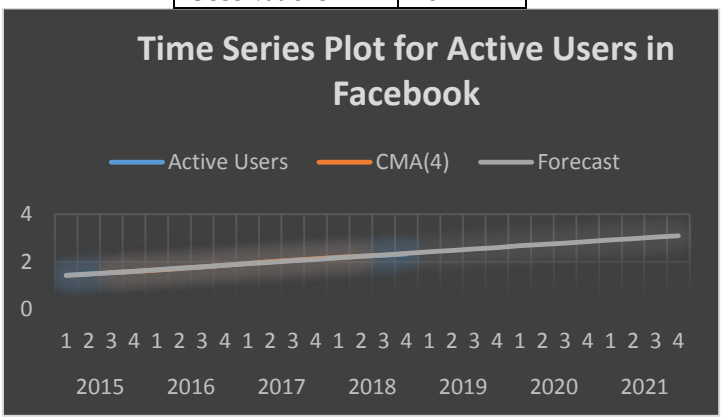
Quarterly Data for Facebook Active Users

			Y_t		Baseline	Y_t/CMA		Y_t / S_t	Trend	
t	Year	Quarter	Active Users (In Billions)	MA(4)	CMA(4)	S_t, I_t	S_t	Deseasonalize	T_t	Forecast
1	2015	1	1.44				1.002	1.437	1.421	1.424
2		2	1.49				1.000	1.490	1.483	1.483
3		3	1.54	1.52	1.54	0.999	1.000	1.540	1.545	1.545
4		4	1.59	1.57	1.60	0.997	0.999	1.592	1.607	1.606
5	2016	1	1.65	1.62	1.65	0.998	1.002	1.647	1.669	1.673
6		2	1.71	1.68	1.72	0.996	1.000	1.710	1.731	1.731
7		3	1.78	1.75	1.79	0.997	1.000	1.780	1.793	1.793
8		4	1.86	1.82	1.86	1.002	0.999	1.862	1.855	1.854
9	2017	1	1.93	1.89	1.93	1.001	1.002	1.926	1.918	1.921
10		2	2	1.97	2.00	1.001	1.000	2.000	1.980	1.980
11		3	2.07	2.03	2.06	1.004	1.000	2.070	2.042	2.042
12		4	2.12	2.10	2.12	0.998	0.999	2.122	2.104	2.102
13	2018	1	2.19	2.15	2.18	1.006	1.002	2.186	2.166	2.170
14		2	2.23	2.20	2.23	1.001	1.000	2.230	2.228	2.228
15		3	2.27	2.25			1.000	2.270	2.290	2.290
16		4	2.32				0.999	2.322	2.352	2.350
17	**2019**	**1**	**?**				**1.002**		**2.414**	**2.419**
18		**2**	**?**				**1.000**		**2.476**	**2.476**
19		**3**	**?**				**1.000**		**2.538**	**2.538**
20		**4**	**?**				**0.999**		**2.600**	**2.598**
21	**2020**	**1**	**?**				**1.002**		**2.662**	**2.668**
22		**2**	**?**				**1.000**		**2.724**	**2.724**
23		**3**	**?**				**1.000**		**2.786**	**2.786**
24		**4**	**?**				**0.999**		**2.849**	**2.846**
25	**2021**	**1**	**?**				**1.002**		**2.911**	**2.916**
26		**2**	**?**				**1.000**		**2.973**	**2.973**
27		**3**	**?**				**1.000**		**3.035**	**3.035**
28		**4**	**?**				**0.999**		**3.097**	**3.094**

13. To project the future data, extend the values of time code, year, quarter, S_t, T_t, and forecast components.

14. Plot the values in the graph.

Regression Statistics	
Multiple R	0.99807
R Square	0.996144
AdjustedR Square	0.995869
Standard Error	0.01903
Observations	16



V. CONCLUSION

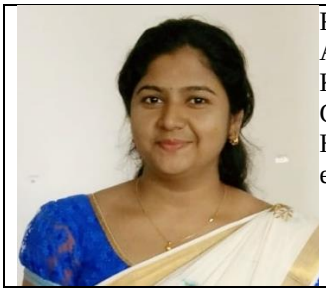
Using Predictive Analytics, organizations can focus on what strategies can be employed to emerge successful in a competitive scenario or even check/correct actions in the present. Our Time Series Analysis shows an upward trend in the active users of Facebook in the future years i.e.; 2019, 2020 and 2021 by analyzing the previous years' data and these findings can be useful for the researchers to understand the expected load on the internet traffic in the future years. The predicted data can also provide valuable insights to the organizations whose operations/marketing tactics are greatly interconnected with such social media platforms. Therefore, the Time Series Forecasting method can prove useful in gaining insights into spheres such as existing customer segments and potential customer base, and thereby help organizations in designing effective brand campaigns and marketing communication for the purposes of their respective businesses. An even closer look at such big data can provide interesting digital marketing insights to businesses.

VI. REFERENCES

[1]. Hampton, K., Goulet, L., Rainie, L., & Purcell, K. (2011).Social networkingsites and our lives.
[2]. Pornsakulvanich, V., & Dumrongsiri, N. (2009). Cultures and perceivedvalues influencing mobile phone use and satisfaction. University of Thai Chamber of Commerce Journal, 29, 1e20.
[3]. Parks-Leduc, L., Pattie, M. W., Pargas, F., & Eliason, R. G. (2014). Selfmonitoring as an aggregate construct: Relationships with personality and values. Personality and Individual Differences, 58, 3e8.

[4]. Liu, C. H., & Yu, C. H. (2013). Can Facebook induce well-being? Cyberpsychology, Behavior, and Social Networking, 16(9), 674e678.
[5]. A. Banumathi, A. Aloysius, (2017). Predictive Analytics Concepts in Big Data – A Survey.
[6]. Neelam Peters, Aakanksha S. Choubey (2016). A Survey on Data Classification and Machine Learning for Forecasting of Student Performance.
[7]. W.Glynn Mangold, David J Faulds (2009). Social Media: The new hybrid element of the promotion mix.
[8]. Foux G (2006). Consumer – Generated Media: Get your customers involved.
[9]. Cioffi-Revilla, C. (2013). Introduction to Computational Social Science: Principles and Applications: Springer Science & Business Media.
[10]. Conte, R., Gilbert, N., Bonelli, G., Cioffi-Revilla, C., Deffuant, G., Kertesz, J., Loreto, V., Moat, S., Nadal, J.P., Sanchez, A., Nowak, A & Helbing, D. (2012). Manifesto of computational social science.
[11]. Lazer, D., Pentland, A.S., Adamic, L., Aral, S., Barabasi, A.L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M. and Jebara, T. Computational Social Science. Science, 323(5915), 721–723. doi: 10.1126/science.1167742
[12]. Asur, S., & Huberman, B. A. (2010). Predicting the future with social media. Paper presented at the IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT).
[13]. Lassen, N., Madsen, R., & Vatrapu, R. (2014).Predicting iPhone Sales from iPhone Tweets. Proceedings of IEEE 18th International Enterprise Distributed Object Computing Conference (EDOC 2014), Ulm, Germany, 81–90, ISBN: 1541–7719/1514, doi: 1510.1109/EDOC.2014.1520.
[14]. Golder, S. A., & Macy, M. W. (2011). Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. Science, 333(6051), 1878–1881.
[15]. Chunara, R., Andrews, J. R., & Brownstein, J. S. (2012). Social and News Media Enable Estimation of Epidemiological Patterns Early in the 2010 Haitian Cholera Outbreak. American Journal of Tropical Medicine and Hygiene, 86(1), 39–45. doi: 10.4269/ajtmh.2012.11–0597
[16]. Evangelos K, Efthimios T and Konstantinos T. (2013) Understanding the predictive power of social media. Internet Research, 23(5), 544–559.
[17]. Vatrapu, R., Hussain, A., Lassen, N. B., Mukkamala, R., Flesch, B., & Madsen, R. (2015).
[18]. https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/.

N. Anuradha
Assistant Professor
Krupanidhi Degree College
Having 5 Years of Experience in Teaching



Roopini J
Assistant Professor
Krupanidhi Degree College
Having 1 year of experience in Teaching