ADV MATH SCI JOURNAL
Advances in Mathematics: Scientific Journal **9** (2020), no.6, 4047–4057
ISSN: 1857-8365 (printed); 1857-8438 (electronic)
https://doi.org/10.37418/amsj.9.6.86 Spec Issiue on ICAML-2020

TIME SERIES ANALYSIS OF PUBG AND TIKTOK APPLICATIONS USING SENTIMENTS OBTAINED FROM SOCIAL MEDIA-TWITTER

IRAM[1] AND HIMANSHU AGGARWAL

ABSTRACT. Social media is used as a platform to express thoughts and opinions on different things. The opinions contain positive, negative and neutral sentiments on a topic. The Twittersphere provides valuable information about events. This study focuses on analyzing the comparison of sentiments between two addictive applications that include PUBG and TikTok. The researcher explored the changes in user's tweets towards these addictive apps by utilizing the data analytics approaches including sentiment analysis and time series analysis to analyze the changes of tweet volume and sentiments. The algorithm used for sentiment analysis in this study is the Naive Bayes machine learning model and an Auto-Regressive Integrated Moving Average (ARIMA) model for time series analysis of tweets. The steps for this research work are data gathering, preprocessing data, sentiment analysis, and time series analysis. The sentiment analysis obtained in this research work shows that Twitter users produce more negative tweets in the case of PUBG and more neutral tweets in the case of TikTok.

1. INTRODUCTION

Addiction is the state of unable to control one's behavior. People express their opinions regarding various apps, events, political leaders, business, etc. on social media freely. This encourages researchers to use Twitter social media as a data source for sentiment analysis to obtain sentiments contained in an opinion

[1]*corresponding author*

2010 *Mathematics Subject Classification.* 94A16.

Key words and phrases. Time Series Analysis, PUBG, TikTok applications.

of users. In this research work, the researcher used PUBG and TikTok tweets as a data source. Player Unknown's Battlegrounds (PUBG) is an online video game in which more than one person can play. In March 2017, PUBG was developed for Microsoft Windows and then in 2018, it was developed for mobile. Almost there were 555 million PUBG players in 2019. But people have been started to addict to this game, as a result, many negative effects such as mental health disorder, lack of interest in the study and negative thoughts among players came into existence. Due to these negative impacts, cities of Gujarat which include Surat, Rajkot and Vadodara banned PUBG in March 2019. Also, in April 2019 Nepal and China banned PUBG. TikTok is an online video platform that provides facilities like uploading, converting, storing and playback videos to the users and owned by ByteDance. TikTok was launched in September 2017. TikTok has 500 million users in three years. This app includes both positive and negative effects. The positive effects are instant publicity and wholesome entertainment, which are less in comparison to negative effects such as addiction, wastage of time, comparison to get more likes, mental health effects. On 3rd April 2019, because of pornography, Madras High Court asked the Govt. of India to ban the app. As a result of this, both Google and Apple removed TikTok during that period. But the company had removed all inappropriate videos. Afterward, TikTok was made available. Due to the exploitation of PUBG and TikTok apps in the past encourages the study of these two addictive applications. Based on the tweets on Twitter, sentiment analysis can be carried out. Thus, the research questions includes-What are the volume of PUBG and TikTok users tweets on Twitter? What are the sentiment trends of PUBG and TikTok users tweets on Twitter? What will be the future sentiment trends using time series analysis?

2. Literature review

2.1. **Twitter for sentiment analysis and Time Series Analysis.** In recent years, a huge number of people have been attracted to social platforms. Most use social sites to express their emotions, beliefs or opinions about things, places or personalities. Twitter is a microblogging site. It was released on July 13, 2006 (Mostafa, 2013). Tweets consist of many characteristics [5]. The tweet has unique attributes like maximum character length of 280 characters, tweets

are fetched easily using Twitter API etc. The main purpose of sentiment analysis is to analyze opinions from a text to find out whether the opinion given by the user is positive, negative or neutral. The sentiment analysis has been carried out to monitor the brands to act during a sudden rise in negative sentiment for product safety [9]. Consumer security breaches can be detected in the early stages and prevent further destruction using sentiment analysis [6]. Phoong, (2018) demonstrated a negative sentiment score on employment in Malaysia [13]. The Facebook group has been used to determine recent trends and characteristics of people's food habits [2] and sentiment analysis of Social Networking Sites (SNS). Data using Machine Learning has been used to find the depression level of a person [7]. Tweets for the 2016 US Presidential Election has been used to calculate sentiment expressed and compare with polling data to see the correlation [10]. Mahtab(2018) explored sentiments of people expressed on Bangladesh Cricket [12]. The perception or influences of the phenomena has been measured by the sentiment analysis of Turkish Social media [11]. Time series are used to uncover the hidden trends and insights.Time series analysis has been used in different fields including political events, disasters, stock market predictions and economic forecasting. Tweets generated by users related to five leading UK retailers namely: Amazon, Argos, Asda, John Lewis, and Tesco during the Black Friday, Christmas, Boxing Day and New Year's Day in the United Kingdom have been used for sentiment analysis of custome's tweets and to understand what exactly drives changes in customer sentiments during these specific periods [8]. Time series analysis has been used to predict disease counts with structural trend changes [14] and also used to evaluate the temporal patterns of dengue incidence from 2001 to 2014 and forecast for 2015 in two Brazilian cities [3]. Li et al. (2016) used YouTube videos to discover the popularity of videos over time. (Connor, Balasubramanyan, Routledge, and Smith, n.d.) explored the relationship between public opinion and sentiment expressed through tweets regarding presidential polls.

3. Research methodology

3.1. **Model framework.** First, the dataset for research is the tweets that mentioned the user's opinions regarding addictive applications. The implementation of method and analysis begins with the pre-processing of tweets which includes

removing URL, username, hashtag, special characters etc. In this way, tweets are filtered. Then the sentiment analysis is performed on the cleaned tweets. It is performed to explore the volume in the tweet's sentiment. Then, it is followed by a time series analysis using an ARIMA model.

3.2. **Data collection and Pre-processing.** In this research work, 3750 tweets referring to PUBG and 2571 tweets referring to TikTok using Twitter API from May 2019 to September 2019 were gathered. The tweets were extracted with the help of keyword: "[application name] lang:en since:[start date] until:[end date]". The researcher filtered out non-English tweets during the pre-processing step. Tweets were further pre-processed by changing all the characters to lowercase and removing punctuations. Also, then the stopwords (such as 'is', 'are' etc.) were removed. Besides, the researcher omitted all meaningless words.

3.3. **Analyzing Tweet Sentiments and Time Series.** A 'TextBlob' approach is used to analyze tweet sentiment. It is a high-level library and uses a movie reviews dataset. Positive and negative features are extracted from each positive and negative review. The training data contains labeled positive and negative features. Naive Bayes Classifier is used to train data. The sentiment function returns two properties-polarity and subjectivity. Polarity is a float that lies in the range [-1, 1] where '1' means positive statement and '-1' means a negative statement. Subjective sentences describe personal opinion, emotion or judgment. Subjectivity is also a float value and lies in the range [0, 1]. In this study, the researcher classified the sentiment of each collected tweet as: positive, negative or neutral. This algorithm has been tested for accuracy and results in an accuracy of 86.43 as compared to other machine learning algorithms [4]. A time series is a set of observations taken at specified times at equal intervals. In its application, time series analysis takes input continuous data to analyze patterns, trends in a time series dataset over an observed time interval. Time series data has four components i.e. seasonality, trend, cycle, and irregularity component. The appearance of pattern in a regular interval during the same month, every year is called seasonality. A trend is a gradual shift or movement to relatively higher or lower values over a long period. It is a direction into which something is changing, developing. A trend may be upward, downward, horizontal or stationary. The wavelike pattern that occurs beyond a year is called Cycle. When

the above three patterns have been extracted out of the series the remaining part is the irregularity (noise) component.

3.4. **Time series analysis model.** The researcher used the Box-Jenkins procedure for fitting an ARIMA model. The ARIMA models are defined by three terms (p, d, q). The Box-Jenkins procedure consists of three steps. The first step is the identification of the model structure. If the series is non-stationary then it is made stationary by differencing. The second step is the estimation of parameters of an ARIMA model using Yule-Walker equations. The best model is selected by diagnostic checking i.e. the third step of the procedure. In this study, the researcher selected the model with the lowest AIC value [1] by fitting various ARIMA models.

3.4.1. *Stationarity and test to check stationarity of time series.* A series with constant mean and variance is called stationary series. The stationarity of series is important because if the series is non-stationary, then one can study its behavior only for the period under consideration. In a stationary time series, the data is independent on time. The Augmented Dickey-Fuller (ADF) test called a unit root test which includes the null hypothesis test. The null hypothesis states that the series is non-stationary which means that the time series has some time-dependent structure and includes a unit root. On the contrary, the alternate hypothesis states that the series is stationary and does not have a time-dependent structure. The result is interpreted by calculating the p-value. If a p-value is less than 5 percent then the null hypothesis is rejected; otherwise if a p-value greater than 5 percent then the null hypothesis is accepted. The non-stationary series is made stationary by differentiating. A series that is stationary after being differentiated d times is said to be integrated of order d, denoted by $I(d)$.

3.4.2. *Autoregressive Model (AR Model) and Moving Average Model (MA Model).* The representation of an AR model is the AR(p) model where it depends on p of its past values. The present value of a variable (Y_t) is determined by its past values $(Y_{t-1}, Y_{t-2}, Y_{t-3}, \ldots.)$ and Y_t (dependent variable Y at period t) is calculated by equation (3.1) as:

$$Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \cdots + \beta_p Y_{p-1} + \epsilon_1, \tag{3.1}$$

where a is the intercept term, Y_{t-1} is the first lagged value of the Y, β is the coefficient of lag 1 that the model estimates, ϵ_1 is the unexplained part (gap) of

actual data and fitted line by regression equation, called error. The MA model also estimates the present value of a variable (Y_t) where Y_t depends on the random error terms. The representation of an MA(q) model where it depends on q of its past values and is represented by equation (3.2) as:

$$Y_t = \alpha + \epsilon_t + \phi_1\epsilon_{t-1} + \phi_2\epsilon_{t-2} + \cdots + \phi_q\epsilon_{t-q}, \tag{3.2}$$

where the error terms are the errors of the autoregressive models of the respective lags and are assumed to be white noise processes with mean zero and constant variance. The errors ϵ_t and ϵ_{t-1} are the errors from the equations (3.3) and (3.4):

$$Y_t = \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \cdots + \beta_0 Y_0 + \epsilon_t, \tag{3.3}$$

$$Y_{t-1} = \beta_2 Y_{t-2} + \beta_3 Y_{t-3} + \cdots + \beta_0 Y_0 + \epsilon_{t-1}. \tag{3.4}$$

An ARIMA model is used for forecasting time series where the AR and the MA terms are combined. So the ARIMA model is represented by equation (3.5) as:

$$Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \cdots + \beta_p Y_{p-1} + \epsilon_1 + \phi_1\epsilon_{t-1} + \phi_2\epsilon_{t-2} + \ldots.. + \phi_q\epsilon_{t-q}. \tag{3.5}$$

3.4.3. *ACF (Autocorrelation Function) and PACF (Partial Autocorrelation Function) Plots for determining p, q parameters.* Correlation is used to determine the relationship between two variables. The correlation of the time series is calculated by using values of time series with the previous time is called autocorrelation. The ACF is the plot of the autocorrelation of time series and is used to find the p parameter of the time series. A partial autocorrelation is a relationship between the observations at previous time lags and the intervening eliminated observations. PACF is used to find the q parameter of the time series. The time series model is built with different combinations of ACF and PACF values. Akaike's Information Criterion (AIC) is used to compare the quality of a set of statistical models. The model with the lowest AIC value is selected.

4. Analysis and results

4.1. **Trends in Tweet volume and sentiment.** The researcher used the Naive Bayes algorithm for sentiment analysis and Auto-Regressive Integrated Moving Average (ARIMA) modeling to evaluate and forecast of temporal trends of two addictive apps i.e. PUBG and TikTok. During the study period, 3750 PUBG

tweets and 2571 TikTok tweets were downloaded. Out of 3750 PUBG tweets, the Naive Bayes model classified into 1273 positive, 1349 negative and 1128 neutral tweets. Similarly, 2571 TikTok tweets were classified into 850 positive, 546 negative and 1175 neutral tweets. The volume of PUBG tweets is much higher than the volume of TikTok tweets. The positive percentage of two apps during May(33%,38%), June(33%,28%), July(36%,28%), August (37%,43%), September(34%,31%). PUBG positive tweets remained in the range of 33%-37% but TikTok tweets show varying trends i.e. increased, decreased or remained constant during these five months. The negative percentage of two apps during May(40%,16%), June(37%,25%), July(31%,21%), August (36%,17%), September(38%,26%). The negative tweets of PUBG are at a greater height than TikTok tweets. The Tiktok negative tweets lie in range 15%-25% whereas PUBG tweets lie in the range 30%-40%, which means people's opinion towards PUBG is more negative. The neutral percentage of two apps during May(33%,46%), June(30%,46%), July(34%,49%), August (27%,38%), September(28%,42%). In this case, TikTok tweets are high as compared to PUBG tweets that imply people are neutral towards TikTok.

4.2. **Testing the accuracy of the model.** The researcher took negative tweets of PUBG for testing the accuracy of prediction done by an ARIMA model. For this, the researcher divided five month's tweets dataset into two parts, i.e. 75:25 training and testing data respectively. The PUBG negative tweets have 144 of days observations. The researcher took 109 days tweets as 75% of total data for training and 35 days tweets as 25% of total data for testing. In Figure 1, X-axis represents the day number for forecasting as it comprises 35 days and Y-axis represents the number of tweets on that day. The red line represents predicted values obtained from an ARIMA model and the blue line represents expected values for the testing data. As the predicted values appeared close to the expected values with a mean square error of 24.73, determine the accuracy of the ARIMA model. Figure 2 shows the forecast results for testing the accuracy of the model. The blue line represents the training data tweets, the orange line represents the actual tweets where the green line represents the forecast number of tweets and shaded area represents forecast tweets with 95% prediction interval means there is only a 5% chance that the real value will not be in that

range. The 95% prediction interval states that there is a high likelihood that the real observation will be within the range.

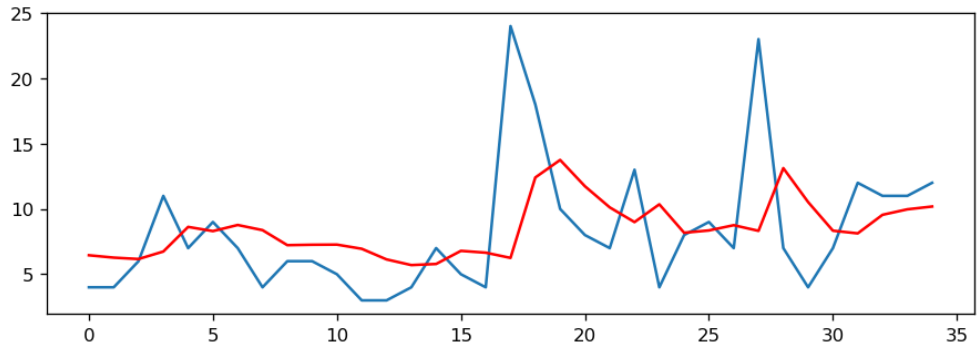

FIGURE 1. Predicted and expected values for testing data

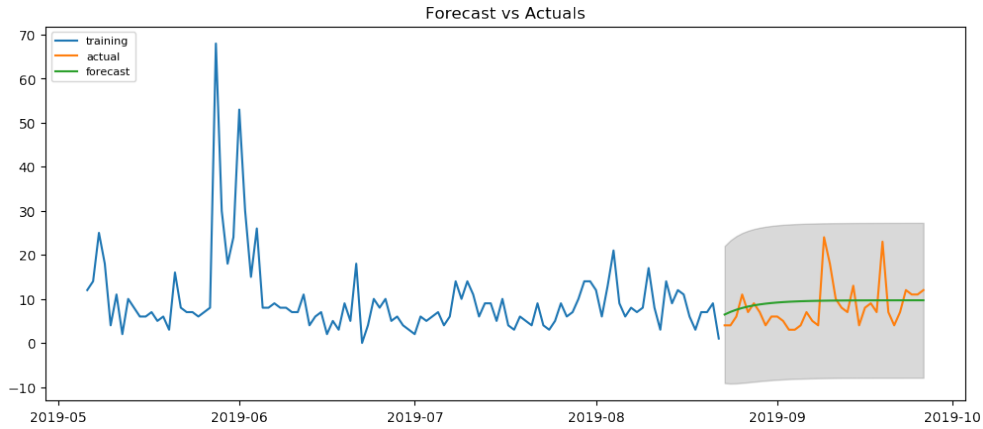

FIGURE 2. Forecast results for testing the accuracy of an ARIMA model

4.3. **Results.** Based on the obtained accuracy of an ARIMA model, the researcher used this model to forecast the number of tweets for the next months from October 2019 to January 2020 for both addictive applications. In the previous month's dataset from May 2019 to September 2019, PUBG has more negative tweets and TikTok has more neutral tweets. From the ADF test, the value of the p-value is less than 0.05 it states that the null hypothesis is rejected and the series is a stationary, necessary condition for forecasting using an ARIMA

TABLE 1. ARIMA models, coefficients and AIC for PUBG and TikTok

	ARIMA(p,d,q)	**AR1**	**MA1**	**AIC**
PUBG	(1,0,1)	0.8056	-0.5074	990.703
TikTok	(1,1,1)	0.2596	-0.8963	960.421

model. The AutoCorrelation Function (ACF) and Partial AutoCorrelation Function (PACF) suggested the best fit model was ARIMA (1, 0, 1) having the lowest AIC value. Figure 3 shows the forecast results for PUBG negative tweets.

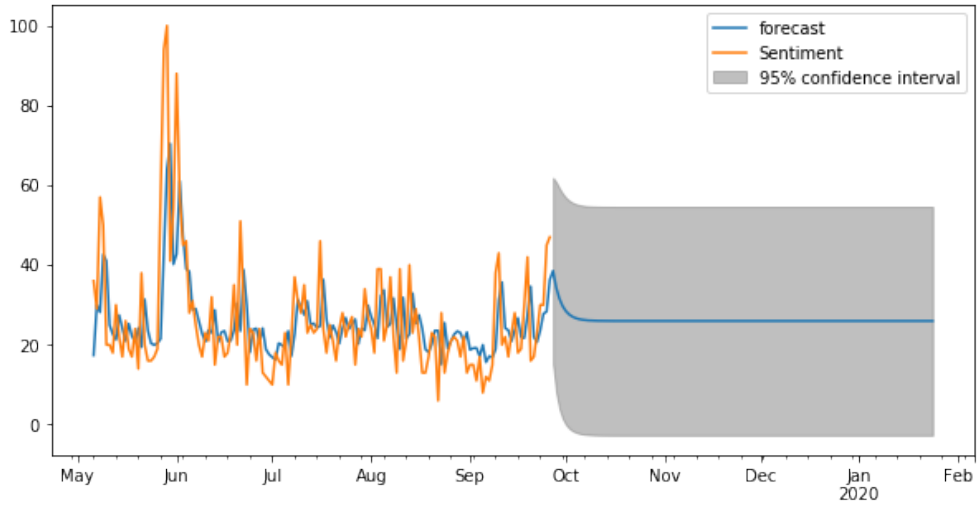

FIGURE 3. PUBG negative tweets forecast

The ADF test for TikTok has a p-value greater than 0.05. The null hypothesis is accepted means series is non-stationary. The series is made stationary by differencing the series. The AutoCorrelation Function (ACF) and Partial AutoCorrelation Function (PACF) suggested the best fit model was ARIMA (1, 1, 1). Figure 4 shows the forecast results for TikTok's neutral tweets. In Figure 3 the orange line represents the actual negative sentiments and the blue line represents the forecast negative sentiments using an ARIMA model. The average negative tweets for PUBG for one day is 25. In Figure 4 the orange line represents the actual neutral sentiments and the blue line represents the forecast neutral sentiments using an ARIMA model. The average neutral TikTok tweets for one day is 20.

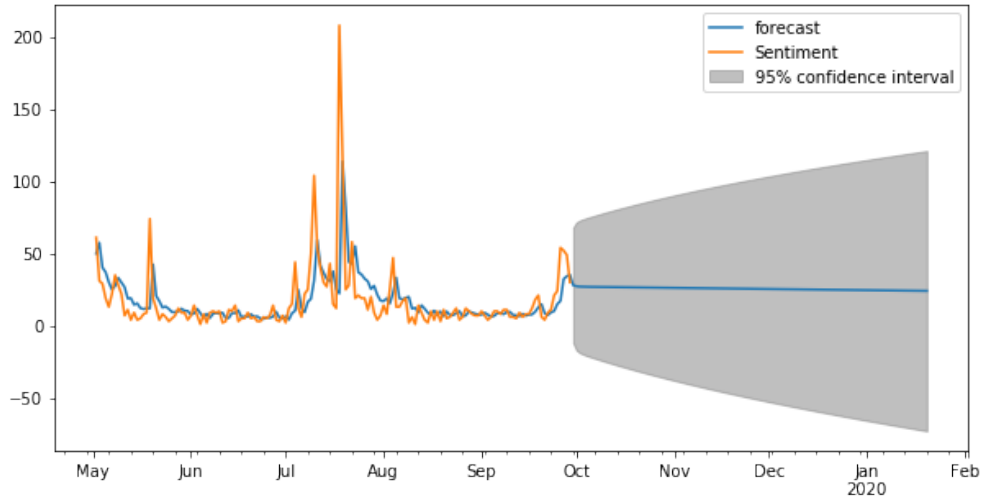

FIGURE 4. TikTok neutral tweets forecast

5. CONCLUSION

Based on the tweets obtained from the twitter about the two addictive applications - PUBG and TikTok, Twitter users are negative toward PUBG and neutral toward TikTok. From 3750 PUBG tweets, there are 1273 positive, 1349 negative and 1128 neutral tweets. Similarly, from 2571 TikTok tweets, there are 850 positive, 546 negative and 1175 neutral tweets. Further, this trend will remain the same in the future as forecast results shown by an ARIMA model with daily an average of 25 PUBG negative tweets, the researcher concludes that there is an again possibility of a ban on PUBG. As people's opinions are neutral in the case of TikTok and will remain the same with an average of 20 tweets per day. So there is no such chance of exploiting TikTok in the future.

REFERENCES

[1] H. AKAIKE: *A new look at the statistical model identification*, IEEE transactions on automatic control, **19**(6) (1974), 716–723.

[2] S. AKTER, M. T. AZIZ: *Sentiment analysis on facebook group using lexicon based approach*, 2016 3rd International Conference on Electrical Engineering and Information Communication Technology (ICEEICT), IEEE, (2016), 1–4.

[3] F. CORTES, C. T. MARTELLI: *Arraes de alencar ximenes r, montarroyos ur, siqueira junior jb, gonçalves cruz o, et al. time series analysis of dengue surveillance data in two brazilian cities*, Acta Trop, **182**(2018), 190–197.

[4] V. A. FITRI, R. ANDRESWARI, M. A. HASIBUAN: *Sentiment analysis of social media twitter with case of anti-lgbt campaign in indonesia using naïve bayes, decision tree, and random forest algorithm*, Procedia Computer Science, **161**(2019), 765–772.

[5] A. GO, R. BHAYANI, L. HUANG: *Twitter sentiment classification using distant supervision*, CS224N project report, Stanford, **1**(12) (2009), 1–6.

[6] J. HAO, H. DAI: *Social media content and sentiment analysis on consumer security breaches*, Journal of Financial Crime, **23**(4) (2016), 855–869.

[7] A. U. HASSAN, J. HUSSAIN, M. HUSSAIN, M. SADIQ, S. LEE: *Sentiment analysis of social networking sites (sns) data using machine learning approach for the measurement of depression*, 2017 International Conference on Information and Communication Technology Convergence (ICTC), IEEE, (2017), 138–140.

[8] N. F. IBRAHIM, X. WANG: *Decoding the sentiment dynamics of online retailing customers: Time series analysis of social media*, Computers in Human Behavior, **96**(2019), 32–45.

[9] H. ISAH, P. TRUNDLE, D. NEAGU: *Social media analysis for product safety using text mining and sentiment analysis*, 2014 14th UK workshop on computational intelligence (UKCI), IEEE, (2014), 1–7.

[10] B. JOYCE, J. DENG: *Sentiment analysis of tweets for the 2016 us presidential election*, 2017 IEEE MIT Undergraduate Research Technology Conference (URTC), IEEE, (2017), 1–4.

[11] H. KARAMOLLAOĞLU, İ. A. DOĞRU, M. DÖRTERLER, A. UTKU, O. YILDIZ: *Sentiment analysis on turkish social media shares through lexicon based approach*, 2018 3rd International Conference on Computer Science and Engineering (UBMK), IEEE, (2018), 45–49.

[12] S. A. MAHTAB, N. ISLAM, M. M. RAHAMAN: *Sentiment analysis on bangladesh cricket with support vector machine*, 2018 International Conference on Bangla Speech and Language Processing (ICBSLP) IEEE, (2018), 1–4.

[13] S. SHAYAA, P. S. WAI, Y. W. CHUNG, A. SULAIMAN, N. I. JAAFAR, S. B. ZAKARIA: *Social media sentiment analysis on employment in Malaysia*, the Proceedings of 8th Global Business and Finance Research Conference, Taipei, Taiwan, 2017.

[14] A. T. KHOEI, J. M. WILSON: *Using time-series analysis to predict disease counts with structural trend changes*, Information Processing & Management, **56**(3) (2019), 674–686.

DEPARTMENT OF COMPUTER SCIENCE
PUNJABI UNIVERSITY PATIALA
PUNJAB, INDIA
Email address: iram11.azhar@gmail.com

DEPARTMENT OF COMPUTER SCIENCE
PUNJABI UNIVERSITY PATIALA
PUNJAB, INDIA
Email address: himanshu.pup@gmail.com