Predicting Service Outages using Tweets

Sunita A Yadwad , V. Valli Kumari

Abstract*— Every user of the internet has high aspirations on its reliability, efficiency, productivity and in many other aspects of the same. Providing an uninterrupted service is of prime importance .The amount of data along with enormous number of residual traces is increasing rapidly and significantly. As a result, analysis of log data has profoundly influenced many aspects of researcher's domains. Social media being integral part of the Internet, real time blogging services like Twitter are widely used due to their inherent nature of depicting social graph, propagating information and entire social dynamics. Content of tweets are of major interest to researchers as they reflect individuals experiences, real time events. Researchers have explored several applications of tweet analysis. One such application is detecting service outages through a myriad of messages posted by users regarding unavailability. Simple techniques are enough to extract key semantics from tweets as they are faster alerts for warning about service unavailability. Similarly, the outage mailing lists are text-based messages which are rich in semantic information about the underlying outages. Researchers find it a great challenge to automatically parse and process the data through NLP and text mining for service outage detection. An extensive study was conducted, aiming to explore the research directions and opportunities on log analysis, tweet analysis and outage mailing list analysis for the purpose of detecting and predicting service outages. A systematic- frame work is also articulated with a focus on all stages of analytics and we deliberately discussed potential research challenges & paths in the above said analysis.

We introduce three major data analysis methods for diagnosing the causes of service failures , detecting service failures prematurely and predicting them. We analyze Syslogs (contain log data generated by the system) for detecting the cause of a failure by automatically learning over millions of logs and analyze the data of a social networking service (namely, Twitter and outage mails) to detect possible service failures by extracting failure related tweets, which account for less than a percent of all tweet in real time with high accuracy. Paper is an effort not only to detect outages but also to forecast them using twitter analysis based on time series and neural network models. We further propose a log analysis model for the same.

Keywords—Tweets, service outage, time series analysis, log analysis

I. INTRODUCTION

Internet is an integral part of every individual's life. Increasing the availability of network is a vital issue, with user's tolerance for the network downtime decreasing. Any utility is expected to provide consistent and reliable service to customers and restore its services at the earliest in case of downtime. Prediction of service outages is an important study for proactive handling of unavailability [1].

Revised Manuscript Received on February 10, 2020.

Sunita A Yadwad , Department of CS &SE , Andhra University , Vishakapatnam , Andhra Pradesh , India .
Mail_id: sunitayadwad@gmail.com

Dr V. Valli Kumari , Department of CS &SE , Andhra University , Vishakapatnam , Andhra Pradesh , India . Mail_id: vallikumari@gmail.com

The major objective of this paper is prediction of unavailability of service (service outage) and the causes for the degradation in the due course.

A service outage can be defined as " An unplanned unavailability of either full or partial features belonging to the service which affect whole or a significant number of users in a way that the outage gets reported publicly" [2].

There are several methods for detecting service outages or downtime like log analysis, tweet analysis and email analysis.

Log analytics is automatic inference of meaningful patterns from voluminous and heterogeneous log data. The traces of each and every activity taking place in an application or device is recorded in a log file. However, the analyzing of log data is very difficult owing to the existence of varied types of logs listing messages with low and higher severity. Most research papers focus on mining of patterns occurring frequently in event logs to characterize what is normal behavior of a system. Some papers have proposed the that it is possible to mine the patterns from event logs with proper application of association rule mining. Classification and proper identification of log messages has many uses, not only for reporting and compliance, but also from the security and system maintenance point of view.

Twitter is a popular platform for discussing countless conversation topics, and the number of tweets now exceeds 400 million per day [3]. It is observed that the number of tweets that relate to network problems is very small in comparison to the tweets in total. We need proper method to extract only relevant tweets (first requirement). In addition, to detect the area where a service failure occurs, we need a way to determine the location of the tweeters (second requirement). The paper and the results emphasize on using the potentiality of twitter analysis using time series for service outage prediction as it has tremendous potential for capturing voluminous current events and sentiments as and when they happen. Exploring such applications that focus on Internet service outages detection through simple techniques can help identify important events [13].

Natural Language Processing (NLP) and Machine Learning technique can be used for analysis and categorization of keywords in the outage mailing list. This analysis plays a pivotal role in classification of the both the cause and effect of Internet outages. Most of the papers contribute in increasing the number of data-sets, using different ways in labeling methods and in predicting an on-going Internet outage to help Internet Service Providers and Internet maintenance. The main objective of our work is to harness firstly the benefits of log analysis for detection of service outage and the causes for the service degradation. Secondly use tweets for time series analysis and prediction of service outages.

Thirdly use email analysis of mails related to outages for listing the causes of service outages. Our paper emphasizes on only the twitter analysis for service outage prediction.

We use exponential smoothing technique for forecasting the future values and circumstances of Service Outages. Exponential smoothing methods are known for giving more weightage to the recent observations, and the lesser weights as we move further in exponential decreasing order [19]. The method partitions the values matching a predicate into separate bins. Exponentially Weighted Moving Average also known as EWMA is used in prediction of successive value in a time series with a significant deviations from normal behavior. Let $\mathbf{y_n}$ denote the number of tweets or logs expressing the required predicate at time interval n. One can then compute the EWMA value $\mathbf{M_n}$ at time n as in equation 1 follows:

$$\mathbf{M_n = \alpha * y_n + (1 - \alpha) * M_{n-1}} \quad \text{.....1}$$

Where α represents how much weight is to be given to the current value. We also compute a smoothed deviation $\boldsymbol{\sigma_n}$ **and** the deviation $\mathbf{D_n}$ in equations 3 and 2 to determine whether an anomaly is occurring at n.

$$\mathbf{D_n = y_n - M_{n-1}} \quad \text{.....2}$$

$$\mathbf{\sigma^2_n = \beta * D_n^2 + (1 - \beta) * \sigma^2_{n-1}} \quad \text{.....3}$$

Finally, **threshold** $\mathbf{T_n}$ can be computed in equation 4 as:

$$\mathbf{T_n = M_{n-1} + \epsilon * \sigma_{n-1}} \quad \text{.....4}$$

If $\mathbf{y_n}$ is greater than equal to $\mathbf{T_n}$ for two consecutive intervals a service outage is signaled.

One can wait for two consecutive threshold violations to determine an outage because declaring for the first violation yields a considerable number of false positives, and waiting for more produces too many false negatives.

II. RELATED WORKS

It is common perception that log files, especially error logs are most potential source of information for analysis post failure and proactively in lieu of upcoming failures. For better access of the information present logs, the data needs to be put into proper shape and we need to pick all the valuable pieces of information from the data deluge. Similarly Tweets and Outage related mails can be analyzed through NLP to detect various events related to service outages. There are several publications emphasizing the work on logs, tweets and mails for the detection of anomalies, failures and in availability. The related work can be further classified under three subcategories.

A. Usage of Outage Mailing Lists for Service Outage Detection

Outages mailing lists are unconventional dataset used to analyze reliability of internet. Through mailing list, network operators share insight experience and information about widespread outages. This dataset is helpful for researchers to perform longitudinal analysis of outages using machine learning, NLP and text mining. Authors [27, 28, 29] emphasize on integration of posts having subject of outages in outages mailing lists by extracting important data information pertaining to outages by filtering out unessential data. Using classification methods like NLTK and TF_IDF and classifier like DA and SVM they can predict an outage type for new thread. Papers [28, 29] have classified 13 outages types based on the causes like the problems in Routing, Power shuts, Natural calamities, Cyber attacks, Fiber Cuts, Domain Name Resolution problem, Device Failure, Congestion in networks, Internet Censorship, Maintenance of hardware and software , Server problems and Human errors[3]. In addition to these 13 outage types, they added one more unknown category because there are some messages having inadequate information to define which outage types they should be included. One can categorize Internet outage types from a big scope to a specific one through automatic mapping of each outage related e-mail thread in categories. Paper [29] emphasized on prediction algorithm which used collection of outages reports and repair logs written in free writing style, where as in the paper [31], authors emphasize on using social network contents like tweets and customer tickets for monitoring of network performance.

B. *Usage of Tweets for Service outage Detection*

Twitter is the most popular online social network service which allows the users to post tweets . Several researchers are using tweets to extract meaning from patterns of cloud data. Authors [23] used NLP methods on social media for indirect measures of network service status. They not only detected network attacks but also provided analysis of public service outages. Social tracking of data that can trigger cyber attack was developed by paper [23,24]. Papers [31] used social media to understand user experience of mobile network. In [13] Viewag et al showed micro-blogging services like twitter can impart situation awareness at times of natural hazards like floods and fire. Authors [35] examined twitter data to help minimize the future catastrophic blackout events. The method of first story detection helped in detecting fake and redundant tweets by checking similarity representation of tweets using locality sensitive hashing algorithm [32]. The authors D. Kergl et al. and Qiu et al. [33,31] were pioneers to evaluate the relationship between tweets from users and the CTT(customer care tickets) addressing the mobile network related experiences [27]. They found tweets which related to a problem were quicker than CTT to the tune of 10 min. Tweets could report a wide spectra of problems of vivid behavior. The paper[31] mapped both the problems reported by Twitter and the incidents already reported by the customer tickets. Not only could they identify the known incidents but could also identify newer short term problems unseen by the ticket system. This process makes mapping of grievances to the specific web services more feasible. This could achieve a major drill down of the Quality of Experience measurements to each and every domain, networks involved and the technologies applied [33].

Despite the popularity of exponential smoothing it is observed that the method dramatically fails in the presence of outliers, abundance of noise, or when the underlying time series observes a change. Paper [34] proposed a flexi model for time series analysis through exponential smoothing cells for overlapping time windows.

The approach detects and removes outliers, de-noise data, fills in the missing observations, and provides meaningful forecasts in every extreme situations. Recent innovations in the convex optimization of large-scale dynamic models make the approach of exponential smoothing tractable.

C. Usage of Log Analysis for Service outage Detection

As systems grow larger and more complex, the quantity and complexity of log files increases and the need for analysis of system log files. There has been previous success at using machine learning techniques on log data with procedures like clustering, principal component analysis, support vector machines and random indexing, genetic algorithms and inductive learning. Machine learning analysis of log files by textual classification tools for identifying events and failures is very common.

Paper [4] describes process used to preprocess error logs of a telecommunication system for a hidden Semi-Markov failure predictor through automatic assignment of IDs, clustering algorithm and filtering for reducing the noise sequences. Authors [5] reviewed numerous of widely and efficiently used solution in open source for collection of the log data from

IT systems. They not only contributed in terms of performance evaluations of the various solutions but also explained the advantages and weakness of each tools.

Papers [6, 20] made a comparative study of all open source tools for cloud computing and concluded that Open Stack developed by NASA is the best solution due to its huge architecture and community support.

The authors [7, 20] present correlation analysis approach between Open Stack logs using machine learning method. Specifically, they analyzed message based log to find the cause of an error by first collecting the event log from both system and application service level, conducting the normalization of collected messages for semantic context and better analysis accuracy, and clustering the data using unsupervised method like hierarchical clustering method and establishing correlation between message group using the Dynamic time warping.

Logreduce Python software was proposed in paper [8] which implements the process of log analysis transparently and assists other software for job failures.Authors [9] publish a method that reduces the effort required to analyze the log files and they claim of log size reduction up to 85% retaining 77% of useful information . The method highlights useful text in the failed log file, paving to debugging of the causes leading to failure.

Shweta Thakur etal.[10] proposed framework combining the distributed platforms of Hadoop and MapReduce for analysis of event logs. They sessionized users on basis of IP-address and timestamp . Many a times companies exploit the event log to analyze customer's behavior

A substantial research of log analysis focusing on three variant management problems like; (a) forensic analysis: post mortem of system logs, (b) fault detection: trap signs of critical failure, and (c) system failure prediction: done for console logs, a proactive approach for early symptoms and warning for potential failures was done in paper . [11]

Log based failure analysis [12] help in experimentation on heterogeneous logs when based on the text mining methods and the deep learning algorithms .

In the paper [14], authors present a work for automatic parsing of console logs (streamed) and detected early warning signals that predict system failure. The paper [15] stresses on experiments with dynamic programming approach for pattern recognition based on dynamic time warping technique (DTW) used.

Paper [16] proposed key ideas of dynamic syslog mining to represent syslog behavior using hidden markov models, a proper methodology for detection of failures and detection of sequence alarms patterns.

The authors [17] collected event logs from the IBM and GENE /L, the fastest supercomputers having 128k processor for better understanding failure in systems and to develop Adaptive Semantic Filtering (ASF) based fault tolerant strategies. They further went on to prove these strategies are accurate, easy in automation and more importantly light-weight. In the paper [18], authors propose an approach that they call SCAPE, which aims in analyzing the log messages during runtime and proactive prediction of problems in advance. The paper [22] conducted study of cloud outage in more than 32 popular services of internet. They analyzed numerous news headlines and public post outage reactions. The study emphasized on the reasons why an outage occurs despite the service agreements and redundant availability of infrastructure.

The authors [20] focus on content mining of data entries in web log to discover information pattern by using clustering algorithms which combine group of webs sites having common browsing contents. The authors [21] present log processing method for Blue gene /L and Cray XT4 systems comprising of three interrelated steps including event categorization, filtering of events and causality related filtering.

III. PROPOSED EXTENSIVE STUDY

The proposed study recommends use of twitter data, mailing lists and log files for the purpose of outage detection. The twitter analysis could help to detect a service outage based on the tweets which hint about it. Similarly there are customer mails indicating the presence of network unavailability. The error logs can also be analyzed to detect anomalies in the services. Emphasis in the paper is given to the implementation of Twitter analysis and the other two analysis are considered for future work.

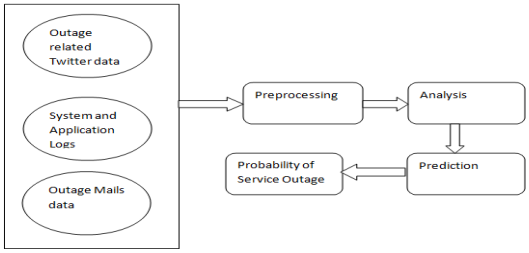

Fig. 1. Architectural model of proposed study

Log files are important source of information available as they contain the traces of execution, warnings, timestamps ,errors etc., which depict the status of individual system parts . They can help in not only analysis of the problem , its development , but also its propagation to other parts. Further, the analysis is usually done for root cause investigation as the system has experienced a problem and to prevent the problem from reoccurrence at runtime[30]. The datasets from outage email list can be considered at the level of threads. The goal is automatic categorization of each outage e-mail thread . However, because computers do not have the network knowledge, sometimes labeling task runs into ambiguity.

Each of the outage type are discrete so we can use document classification method to solve the problem and since the dataset comprises of large number of mail texts the bag of words method can be used to simplify the representation in natural language processing . Supervised and semi supervised learning can be used to label a part of data. [27,28]. This method fits our dataset well and produces considerable improvement in learning accuracy.

Twitter APIs can be used to access data available publicly and data relevant to the portrayal of service outage detection by selecting couple of keywords serving as good queries for the purpose in hand. Tweets must contain words indicating unavailability or performance related words like slow, drop or down. 'Man is the measure of all things' said Protagoras. Human can act as social sensors by identifying broader array failures be it about site being slow, messages being corrupt, messages not being delivered, time out or out of data. End users perception of outage is true measure of downtime.

IV. METHODOLOGY

A. Tweet Analysis

Twitter API can be used to retrieve publicly available data relevant to the task of service outage detection. We can start with tweets by selecting few important words that are deemed as keywords (good queries) for the purpose of service outage detection. Tweets must contain words indicating unavailability ,downtime or performance related words like slow, drop, fail or is-down.' Human can act as social sensors by identifying broader array failures be it about site being slow, messages being corrupt, messages not being delivered, time out or out of data. Twitter network can be considered as a heterogeneous sensor network, with each Twitter users serving as sensor nodes. The service outage signals of the base paper [26] comprised of two simple lexical features the first is the phrase namely "X is down" and the second is hash tag idiom "#Xfail". Here X stands for the name of an online service like Amazon, Facebook, Twitter, Whatsapp, Gmail etc.. This can be further supplemented by incorporating additional predicates into system like 'crashed', 'shut down' having similar syntax as 'is down'. Not just individually we also specify how many expressions of the IsDown or Fail predicates when combined infer occurrence of a service outage [26]. The proposed methodology considers the above two words for filtering the tweets.

The tweets are then split into time_frames in order of their time of origination. Exponential Smoothing is done on the data collected .All the expected values in each of the time frame are then compared with actual values, after which false positives are filtered out. Augustine and Cushing [25] implemented a very similar solution. They used the above said approach for monitoring outages and the network problems in NETFLIX which is a content based delivery network .They could evaluate the their system accuracy because they could fetch the outages list of web services that were being monitored. This explains the why usage of tweets is practically feasible for their use cases. This can be further extended by implementing a practical system which monitors the overall internet web Quality of Experience(QOE) using Twitter analysis. Not only outages in web services, but the level of degradation in web services is detectable. We can investigate whether root causes for the drops in Quality factor can be identified with additional information provided by tweets. Like for example analyzing geographical origin of complaints could help in tracing about regional problems.

Based on the analysis of related work we summarize the steps for Tweet analysis for service outage detection:

1. Pull the tweets from Twitter API for specified interval. The tweets in the result are for the year 2006-2018. A dataset representing tweets indicating the service down is computed based on the words of interest IsDown and #Fail and the bag of words method.
2. We filter out all tweets which do not pass the IsDown predicate. Filter the predicates with IsDown through Fail predicates to reduce the possibilities of considering Mis-tweets. Consider the retweets for the filtered Tweets .
3. We then split remainder tweets in time periods and count tweets in number for each time period. A dataset representing count of tweets per month indicating the service down is computed based.
4. Apply the Exponential smoothing techniques stl and ets through decomposition into trend ,seasonal and remainder value for each period.
5. Forecast the series to project the tweet count for period of next three years using Simple Exponential, Moving Average, Smoothening method and ARIMA model.
6. Compare the MAPE, MSE, ACF and RMSE
7. Perform neural network time series on the data by using ELP, MLP and neural network regression models
8. Apply the smoothing method to better the process and to get rid of the problems of outliers, noise and missing values.
9. Use the QOE method to find the causes for the degradation of service and methods to overcome through log analysis. So probability of occurrences of outages can be predicted.

B. Log Analysis

Through log analysis one can detect presence of abnormal logs The features we extract from the log data and learning algorithms like the unsupervised ones come to our aid.

There are several novel methods for the automated detection of abnormal logs. Logs can be categorized in several categories , e.g. server logs, event logs, dbase logs, error logs, application logs etc. The log contents can be both numerical and character data. For server logs, the numerical data represents features (e.g. CPU load, memory) . But for some purpose the non-numerical data (e.g. systems indicator) could also be interesting .The Time Series analysis is used for numerical data. We predict future values for time taken for server response, pages viewed by users, the success status for request and many more applications.

For the textual data we can apply natural language processing techniques like extracting N-gram frequencies from logs and by using TF-IDF. These methods not only help in determining frequency of a word with respect to a specific document , but also the word power in a document collection.

Several machine learning techniques have been deployed for automatic detection of abnormality in logs, such as text mining, anomaly detection methods. These techniques are bifurcated into supervised classification or unsupervised clustering techniques . Support Vector Machines can classify the system call sequences of solved problems. They fail in determining correlations between the system behaviors and known problems . Hence they cannot be applied for detection of unknown problems. Bayesian networks can also be used for classification technique in anomaly detection. PCA are used to separate the data space into two orthogonal subspaces..

Since it is impossible to be obtain clean data with normal values , many researches have had to rely on unlabelled methods like clustering techniques which show similarities among the data of a cluster but dissimilarity in objects of other clusters.

The proposed work is capable of system log monitoring, application logs monitoring, log files, the syslog data, and alerting whenever a log pattern gets detected . Implementation of an effective log file monitoring offers several benefits like increase in security, awareness of problems ,faster detection of failed services, increased service ,server and application availability

The steps followed in log analysis in the proposed work are as indicated in the diagram below :

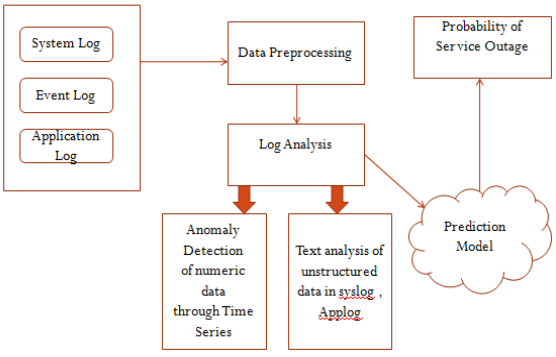

Fig. 2. Framework for log analysis model.

We analyze high volume structured and unstructured log data .We use Machine learning algorithms like PCA, Naïve Bayes, logistic regression, CNN and Deep learning algorithms for proper log analysis and prediction of service outages. Time series analysis furthers helps in forecasting. The proposed methodology helps in avoiding service interruptions, slowdowns and outages proactively .We can have a faster root-cause analysis of problem and recovery. We eventually enhance the system and application performance by improving end-user experience ,increasing operational efficiency and computing. The implementation of log analysis is an extension to the present work.

V. RESULTS

The objective of research can be attained through a model that forecasts the outages through the tweet counts by applying time series analysis method in an efficient way. Feasible and efficient forecasting methods were chosen for short term analysis. The proposed method to forecast the time series process can be is as shown in diagram below.

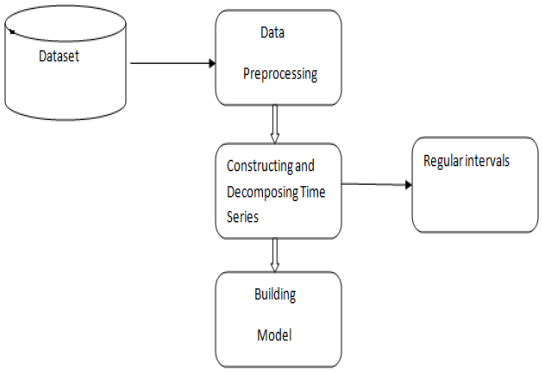

Fig. 3. The framework of steps for building the forecasting time series models

A. Dataset :

For a period of 12 years the count of tweets per month qualifying the predicates ISDOWN are tabulated through a series of analysis techniques with a combination of sentiment analysis, classification, and analysis of volume of traffic to detect if an outage is occurring. Tweets are pulled using Twitter APIs. Tweets qualifying the IsDown predicate are extracted and further filtered with an additional predicate called FAIL . The false positives number can be reduced by making the occurrence of a tweet be supplemented by the Fail predicate within a time period of 60 minutes of the detected start time. Retweets are also considered in order to depict the counts towards tweet analysis. The obtained tweets count against the time is maintained in the dataset. The values in the dataset are scaled for experimentation purpose to the scale 1:100

B. Data Preprocessing :

We use models from Time Series for forecasting values by analyzing all available historical data which are listed in the order of time. We demonstrate time series model in R for the tweets counts obtained against the time for a period of 12 years 2006-2017.

The sampling rate of data set in analysis is one month for a period of time from year 2006 to 2017. R and Rstudio are our building blocks for the model. The raw data is unfit model for construction of the forecasting model owing to missing values and the time frames recorded inappropriately. We apply the Exponential smoothing method to better the process and minimize the problems of outliers, noise and missing values in the preprocessing method. The dearth of information can reduce the prediction efficiency of the model. Methods for filling the lacking data could are vary from statistical functions like mean, median or filling previous value. We replaced the previous value for missing value assuming that the current data is more or less similar to the previous ones.

C. Regular intervals:

Intervals considered in research are on daily, weekly, monthly and yearly basis. But in our work we use monthly time interval.

D. Time Series and decomposing time series :.

This graph is representation of trend ,seasonality existing in our dataset.

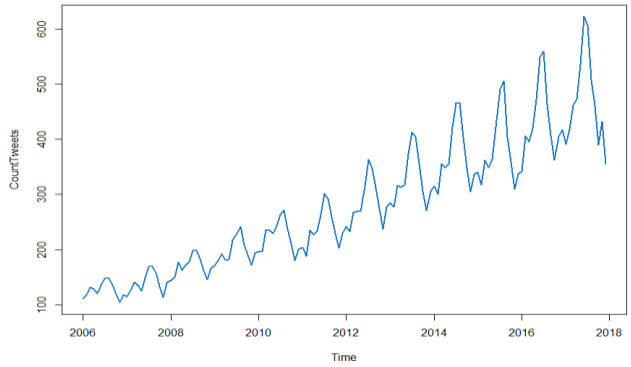

Fig. 4. Time series graph of count of tweets against time

In R the *aggregate* function is an average of the data in the previous year. Once aggregate is computed we plot the graph. Data is converted to log values for elimination of the trend ,seasonality. We forecast the values of tweet count for coming months .

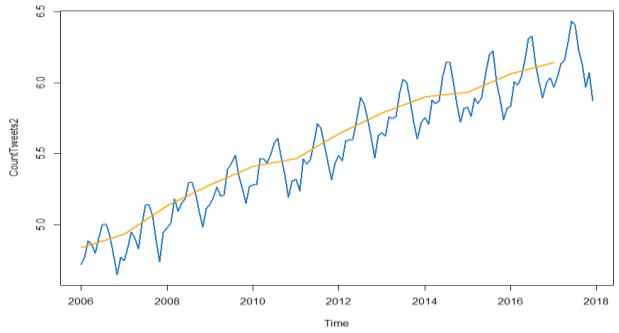

Fig. 5. Moving Average Plot for timeseries graph of tweet count

i. Time series construction : Time series is constructed by the function called the ts() function of R library with a frequency of 12 which indicates that a time series(ts) is a monthly series. The parameters frequency is set to 12 and start = c (2006, 12) for monthly series with start yearr 2006.

The code fragment in R is as given below:

CountTweets2 <- ts(time1$Count, start=c(year,1), end=c(year+i,12), frequency=12)

Here *time1* represent the dataset after log value are evaluated. *time1$Count* represents the count of tweets for the period mentioned from the start year to the end year. *CountTweets 2* computes the time series value

ii. Time series decomposition : This step decomposes time series into the components like the trend, the seasonal component , irregular and the cyclical components through the functions as below:

stl_pass2<- stl(CountTweets2,s.window="periodic")

accuracy (forecast(stl_pass2, h=36))

The function we use for decomposition is Season Trend Decomposition using loess (*stl)* . This will decompose the time series into components known as the trend, the seasonality and the remainder.

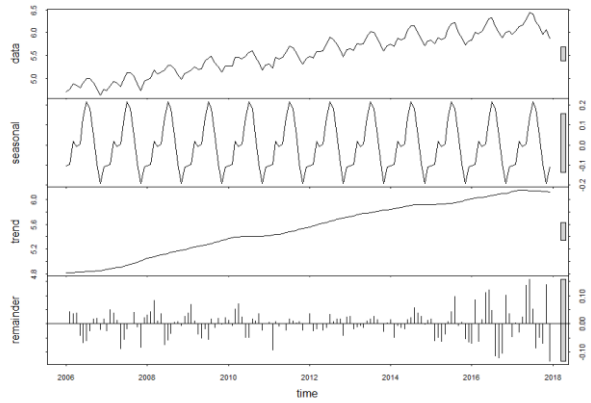

Fig. 6. Time Series Decomposition for trend analysis of tweet count

Box plots measure of how well a the tweet count data is distributed in the entire data set. Its purpose is division of data set into three measuring quartiles. Boxplots represents the lowest, the highest, the median and also the first and third quartile in the respective data set. It is used for comparison of the distribution of data across the entire data set .

boxplot(CountTweets ~ cycle(CountTweets)) # plot across months, a sense on seasonal effect

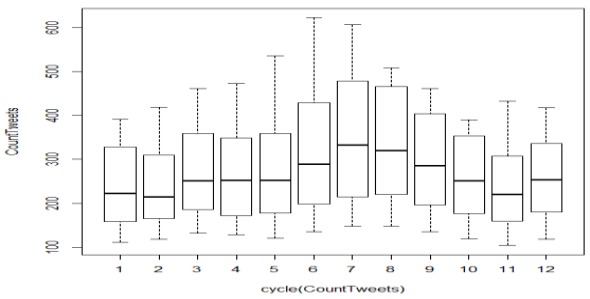

Fig. 7. Box plot of Tweet count distribution

E. Building Models

i. Seasonal Trend Decomposition ***and Exponential smoothing Model:***These methods are used for forecasting with better accuracy. Firstly accuracy of the model is computed using the Seasonal trend decomposition function , then we plot components of the STL with respect to each of the variable. We also plot the forecasted values for the next three years by using the weighted average function (ets). The values indicate the tweet counts forecasted for the next three consecutive years. The forecasted values help in predicting the status of service outages in the coming years.

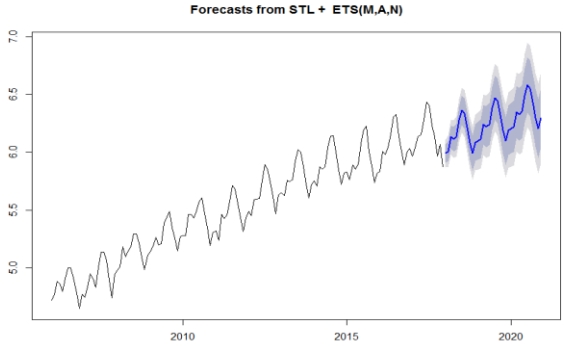

Fig 8. STL(M,A,N) for forecasting tweet count for next 3 years

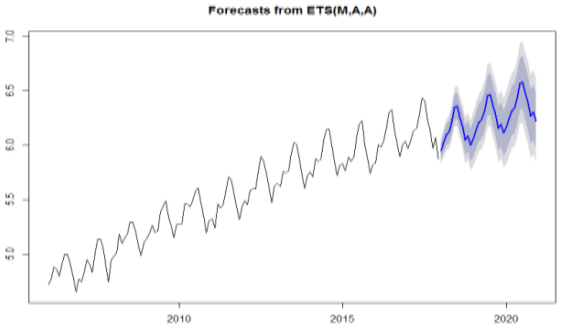

Fig. 9. ETS(M,A,A)) for forecasting tweet count for next 3 years

```
require(forecast)
forecast(stl_pass2, h=12)
ets_pass2 <- ets(CountTweets2)
 forecast(ets_pass2,h=36)
```

TABLE 1 Comparison of STL and ETS values for various criteria

	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
STL	0.0004448455	0.05717519	0.04155986	0.010 0.01047019	0.742468	0.3363071	0.1077069
ETS	0.0007972946	0.05377168	0.03485698	0.009547247	.009547247	0.2820666	0.02426403

The best suited forecasting method and the period are chosen with respect to the smallest value of both the Root Mean Square Error abbreviated as RMSE ,and Mean Average Error known as MAE , respectively. The most suitable forecast is ETS method in consideration of the above values .

*ii. ARIMA Model :*The Box and Jenkins method is used for forecasting and to build the Autoregressive Integrated Moving Average model (ARIMA) an automated way for forecasting. Note that in R, we use auto ARIMA and hence there is no need to specify the tuple order in terms of p,d,q .Here value p quantifies the number of Auto-Regressive terms , q represents number of Moving Average terms , d represents Number of non-seasonal differences like we do in Python. *ARIMA* is popular forecast function. The graphs for the forecasted values in ARIMA model are represented as :

```
arima_pass <- auto.arima(CountTweets2)
accuracy(arima_pass)
```

TABLE 2 : Coefficients

	ar1	ma1	sma1
	0.4538	-0.8575	-0.4097
s.e	0.2048	0.1539	0.0975

sigma^2 is estimated : 0.003253:
log likelihood=189.43 **AIC**=-370.85 **AICc**=-370.54 **BIC**=-359.35

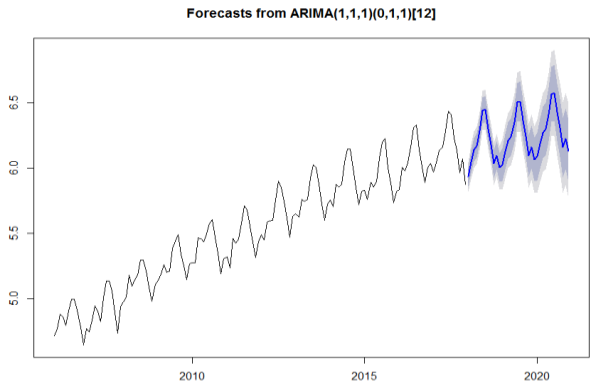

Fig. 10. ARIMA model

```
legend ( "topleft",  c(  " Actual", "Forecast", "Error Bounds (95% prediction interval)"),      col = c( 1, 2, 4 ),
lty=c(1, 1, 2))
res1 <- residuals(fit)  # getting the  fit  residuals
```

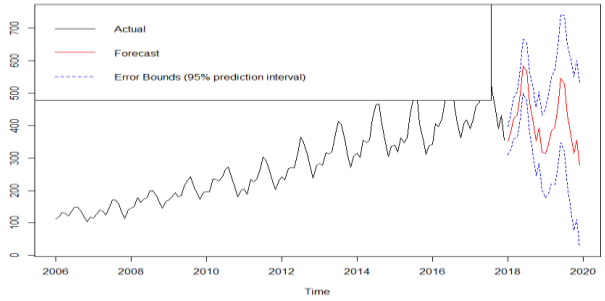

Fig. 11. Residuals Analysis

In the diagram above, the solid line(red) represents future values, and dotted lines (blue) are indication of error bounds with 95% confidence . The forecast values are higher for years 2018, 2019 as seen by the forecast diagrams. Years 2018 and 2019 saw some epic outages in reality.

iii. Neural Network Autoregression (nnetar): We see how the model performs with default/auto parameters. We can fit an MLP network to a time series. There is preprocessing of time series , specifying of autoregressive inputs. The pre-specified arguments trains 20 networks with a 5 node hidden layer to produce an ensemble forecast .

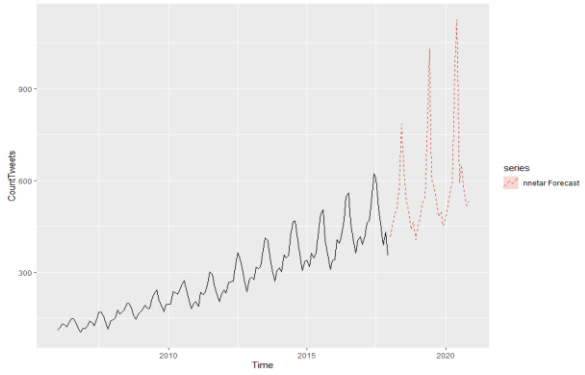

Fig .12. Neural Network Autoregression (nnetar)) for forecasting tweet count for next 3 years

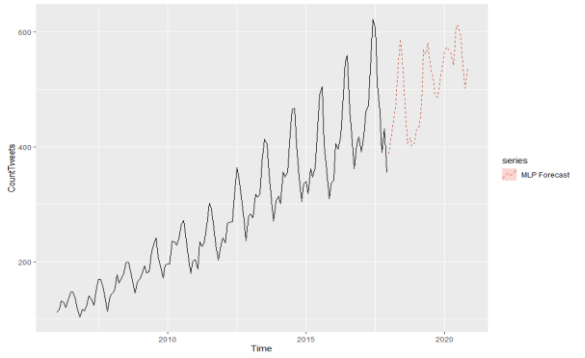

Fig. 13. nnfor - Multilayer Perceptrons (MLP)) for forecasting tweet count for next 3 years

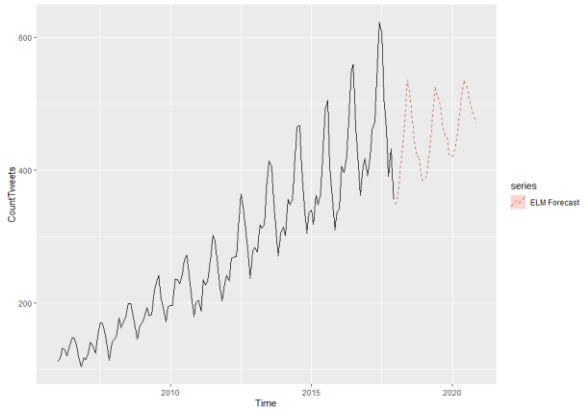

Fig . 14. nnfor - Extreme Learning Machine (ELM)) for forecasting tweet count for next 3 years

Mean absolute error: The mean absolute error (MAE), the mean difference between each model's predicted value and the test value is a precise measure of model performance. The MAE here helps us in concluding that the MLP model performed better for the twitter data with minimal MAE in comparison to nnetar and ELM models.

TABLE 3: Mean absolute error (MAE) measure of neural network model performance

MAE.nnetar	MAE.MLP	MAE.ELM
41.73539	19.37591	24.73438

We used Neural networks models as they serve as an alternative approach to the time series forecasting. The neural network methods forecast our series data nicely.

In the paper [26], the authors focused application of twitter analysis phenomenon for inferring Internet service availability to determine when some of the popular services experience downtime. Augustine et.al [25] used the approach [26] for outages monitoring and network problems monitoring in the NETFLIX network which is a well known content delivery network They evaluated the effectiveness of their method in comparison with other SPOONS system methods. They integrated the phrase "IsDown" and spike detection of Exponential Smoothing in the SPOONS system and named one as the trend monitor and other as keyword volume method. But their achievement was also restricted to outage detection and validation. The proposed work makes the following contribution:

1. It not only detects the outages through twitter but also forecasts the future downtime on similar lines. Time series analysis is used not only for exponential smoothening but through models like ETS, ARIMA and neural networks the forecasting of outage counts are done to show intensity of future downtimes.
2. A comparative study of the forecasting models is done to infer which is better suited for the prediction.

VI. CONCLUSION

The paper is a sincere effort to exploit the social network content and the logs for monitoring the network and detecting the presence of a service outage. Users of network provide their feedback about services and occurrence of an outage through either tweets or emails. These feedbacks are often complementary source for understanding the service outages and their impact on the users. By applying natural language processing techniques and Machine learning algorithms we can better understand tweets, mails and logs and use the analysis for service outage detection. Paper explores time series models and neural network models for outage prediction based on the count of tweets The work can further be extended to find the severity of the outages based on the scale of the responses and the sentiments in messages. The survey in the paper also conveys that service outages are inevitable however smart and successful the provider is. The work not only proposes means to detect service outages but also helps in forecasting the same using time series analysis. The main limitation here is the incompleteness of the Service outage data.

Majorly outages are a result of a selective few causes. So we focus on "80%" of the outages caused by the "20%" of the causes.

REFERENCES

1. Itron Analytics whitepaper on "Outage Detection" , 101531PO-02 07/17.
2. Haryadi S. Gunawi, Riza O. Suminto, Mingzhe Hao, Agung
3. Laksono, Jeffry Adityatama, Anang D. Satria, and Kurnia J. Eliazar.
4. " Why does the cloud stop computing? Lessons from hundreds of
5. service outages ". Proceedings of (SoCC'16) the 7th ACM Symposium on Cloud Computing .
6. Kieran Matherson. , *Supervisor*: Richard Nelson." Machine Learning Log File Analysis". Research Proposal. 13 March, 2015.
7. Salfner, F. and Tschirpke, S.: "Error log processing for accurate failure prediction". WASL'08, Proceedings of 1st USENIX conference on Analysis of system logs, pp. 4–4. USENIX Association, Berkeley, CA, USA (2008).
8. Risto Vaarandi, Paweł Nizi_ski, "Comparative Analysis of Open-Source Log Management Solutions for Security Monitoring and Network Forensics", NATO Cooperative Cyber Defence Centre of Excellence, Tallinn, Estonia.
9. "Openstack: Open source software for building private and public clouds." [Online]. Available: https://www.openstack.org
10. Ju-Won Park1 and Eunhye Kim2,Correlation Analysis of OpenStack "Log using Machine Learning Techniques" , International Conference Grid, Cloud, & Cluster Computing , GCC'17
11. Tristan de Cacqueray ,"Quiet log noise with Python and machine learning," 28 Sep 2018 , OpenStack Vulnerability Management Team (VMT) member working at Red Hat
12. T. Yang and V. Agrawal. " Log file anomaly detection". CS224d Fall 2016, 2016.
13. Shweta Thakur, Sampada Vishwas Massey, "Event Log Analysis: A Systematic Review ", International Journal of Scientific Research Engineering & Technology (IJSRET), ISSN 2278 – 0882 Volume 6, Issue 6, June 2017.
14. Zhang, Ke & Xu, Jianwu & Renqiang Min, Martin & Jiang, Guofei & Pelechrinis, Konstantinos & Zhang, Hui. (2016). "Automated IT system failure prediction: A deep learning approach". 1291-1300
15. A. Jiao and I. Daulagala. "Logfile failure prediction using recurrent and quasi recurrent neural networks". CS 224N, Winter 2016, 2016.
16. Vieweg, Sarah & Hughes, Amanda & Starbird, Kate & Palen, Leysia. (2010). " Microblogging during two natural hazards events: What Twitter may contribute to situational awareness". Conference on Human Factors in Computing Systems - Proceedings. 2. 1079-1088. 10.1145/1753326.1753486.
17. Ke Zhang, Jianwu Xu, Martin Renqiang Min, Guofei Jiang, Konstantinos Pelechrinis, Hui Zhang: "Automated IT system failure prediction: A deep learning approach" . BigData 2016: 1291-1300
18. Berndt D, Clifford J (1994) , "Using dynamic time warping to find patterns in time series". AAAI-94 workshop on knowledge discovery in databases, pp 229–248
19. Yamanishi, Kenji & Maruyama, Yuko. (2005). "Dynamic syslog mining for network failure monitoring." 499-508. 10.1145/1081870.1081927.
20. Liang, Yinglung & Zhang, Yanyong & Xiong, Hui & Sahoo, Ramendra. (2007)." An Adaptive Semantic Filter for Blue Gene/L Failure Log Analysis" . 1-8. 10.1109/IPDPS.2007.370635.
21. Pitakrat, J. Grunert, A. van Hoorn ,O. Kabierschke, F .Keller, "A framework for system event classification and prediction by means of machine learning **" In** Proc. 8th Int. Conf. on Performance Evaluation Methodol ogies and Tools (VALUETOOLS '14)(2014)
22. Vazquez, C. & Krishnan, R. & John, E.. (2015). "Time series forecasting of cloud data center workloads for dynamic resource provisioning". 6. 87-110
23. Asadi, Tawfiq & Obaid, Ahmed. "An efficient web usage mining algorithm based on log file data ". 92. 215-224. (2016).
24. Zheng, Ziming & Lan, Zhiling & H. Park, Byung & Geist, Al. (2009). "System Log Pre-p rocessing to Improve Failure Prediction". 572 - 577. 10.1109/DSN.2009.5270289.
25. Zheng Li, Mingfei Liang, Liam O'Brien, He Zhang:The Cloud's Cloudy Moment: "A Systematic Survey of Public Cloud Service Outage", *International Journal of Cloud Computing and Services Science*, 2 (5):1-15, 2013. (doi:10.11591/closer.v2i5.5125).
26. Kumar, Sumeet & Carley, Kathleen. (2016). "Understanding DDoS cyber-attacks using social media analytics". 231-236. 10.1109/ISI.2016.7745480."
27. Hernandez, Aldo & Sanchez-Perez, Gabriel & Toscano-Medina, Karina & Martinez-Hernandez, Victor & Perez-Meana, Hector & Olivares Mercado, Jesus & Sanchez, Victor. (2018)." Social Sentiment Sensor in Twitter for Predicting Cyber-Attacks Using ℓ1 Regularization. Sensors". 18. 1380. 10.3390/s18051380
28. Eriq Augustine , Cailin Cushing , Alex Dekhtyar , Kevin McEntee , Kimberly Paterson , Matt Tognetti, "Outage detection via real-time social stream analysis: leveraging the power of online complaints", Proceedings of the 21st international conference companion on World Wide Web, April 16-20, 2012, Lyon, France [doi>10.1145/2187980.2187983
29. K. Levchenko, B. Meeder, M. Motoyama, S. Savage,and G. M. Voelker. "Measuring online service availability using twitte"r. In Proc. of the 3rd Workshop on Online Social Networks (WOSN 2010), 2010
30. Guanyu Zhu ,Wei-Ting Lin,Zhaowei Sun ,Network Outages Analysis and Real-Time Prediction".
31. Ritwik Banerjee, , Luis Chiang, Abbas Razaghpanah, Yejin Choi, Akassh Mishra, Vyaas Sekar and Phillipa Gill."Internet Outages, the Eyewitness Accounts: Analysis of the Outages Mailing List"
32. Jaech, Aaron & Zhang, Baosen & Ostendorf, Mari & Kirschen, D.s. (2018). "Real-Time Prediction of the Duration of Distribution System Outages". IEEE Transactions on Power Systems. PP. 10.1109/TPWRS.2018.2860904 . .In Proceedings of the 16th International Conference on Passive and Active Network Measurement, PAM 2015 – Vol. 8995, pp. 206 – 217. Springer, 2015
33. "Predicting Service Outage Using Machine Learning Techniques" ,HPE Innovation Center.
34. Qiu, Tongqing & Feng, Junlan & Ge, Zihui & Wang, Jia & Xu, Jun & Yates, Jennifer. (2010). " Listen to me if you can: Tracking user experience of mobile network on social media ". Proceedings of the ACM SIGCOMM Internet Measurement Conference, IMC. 288-293. 10.1145/1879141.1879178.
35. Huddar Mahesh ,Ramannavar Manjula , Sidnal Nandini. (2015). "Scalable distributed first story detection using storm for twitter data". ICAETR 2014, International Conference on Advances in Engineering and Technology Research, 2014, 10.1109/ ICAETR.2014.7012915.
36. Kergl D. , Roedler R,Rodosek G.D (2017) "Towards Internet Scale Quality-of-Experience Measurement with Twitter "In Tuncer D., Koch R.,Badonnel R, Stiller B. (eds) Security of Networks and Services in an All-Connected World. AIIMS 2017. Lecture Notes in Computer Science ,vol 10356. Springer,Cham.
37. Abrami, Avner & Aravkin, Aleksandr & Kim, Younghun. (2017). "Time Series Using Exponential Smoothing Cells".
38. K. Lee, J.-y. Shin, R. Zadeh, "Twitters effectiveness on blackout detection during hurricane sandy", 2013

AUTHORS PROFILE

Ms. Sunita A Yadwad, is a Research Scholar at Andhra University with over 22 years of teaching experience.. An MTech in Computer Science and Engineering and a BE in CSE her research interests mainly focus on Data Analytics and Machine learning.

Dr Valli Kumari Vatsavayi, a Professor at Andhra University with over 27 years of experience in teaching and research, holds a PhD degree with Gold medal for best research. An MTech in Computer Science and Technology and a BE in ECE she researches on Block Chain, Computer Vision, Data Analytics, Security, Privacy, Forensics and Threat Intelligence. She is the Principal Co-Investigator of the Cyber Security and Data Analytics Centre sponsored by ISEA Project, MEITY, GOI.