Employee Turnover Prediction
Browsing history and HR data

H.I. Viran Pravinda, G.D.M.S. Sathyani, T.D.C. Kavinda, M.A.P.I. Perera, Dr. Malitha Wijesundara

viranpravinda32@gmail.com, savidya19@gmail.com, ckavindat@gmail.com, isoo.isurunath@gmail.com, malitha.w@sliit.lk

Abstract – Employees are the most valuable asset that a company can have without any doubt. Keeping a high valued and skillful employee within a company will be beneficial for the well-being of a company. Employees quitting the company will be the downfall of that company. Therefore, it is better if the higher management of a company can predict whether an employee is going to leave the company or not and try to keep him/her within the company. In this paper, we have briefly discussed how to achieve the above prediction using an employee's emails, call logs, web browsing history and HR data. Moreover, data mining techniques that we used to get the prediction from each and every component will be discussed.

Keywords – Web Browsing History, HR data

I. INTRODUCTION

In this competitive business world, Employees are the most valuable assets of a company. Even though a company could reach to its success by following standard procedures and techniques, if they do not have well mannered, effort driven employees, they won't be able to access the peak point of their success. Therefore, not only recruiting employees but also retaining good employees in the company, is also equally important. Even though there are many qualified employees out there ready to take a chance to work, loss of a current employee is irreplaceable. In this kind of situation, an organization will get effected in both financially and socially. Hiring a new employee will require an extra effort. Further, co-workers who worked with this particular employee will have to adapt to the new person and his new work patterns. The ripple effect of losing a great employee is tremendous and it goes well beyond what is easily quantified. [1] Turn off of a good employee will create uncertainty to both fellow employees and customers who dealt with that employee and organization might lose more employees or customers on that uncertainty.

In order to keep an employee interested, many things can be done. Identify their potentials and giving them opportunities to grow in the company is one of the major productive actions that can be try as a common procedure. But the better solution is, if we could focus on individual employees and predict if they are going to leave, we can take necessary specific actions suitable for each employee. Employee behavior in the working environment plays a big part in this.

Most of the researches on this topic are based on non-technical background, focusing on human behavioral patterns. But there are some researches actually focusing on technical background combined with social analysis theory to come up with a better solution. In one solution, they are using data mining techniques to analyze previous employees' records and build a model to predict future turnoffs. [2] Furthermore, there is a another application called "Workday" which will predict the turnoff patterns of an employee and suggesting possible career changes that will keep them in the company. [3]

In this paper, we are focusing on four main areas that we found by talking to with HR professionals of few companies (Ceylinco Life Insurance, Gajamuthu Food Products Pvt. Ltd.), Web Browsing History, Call Log, Emails, and HR data such as Attendance, Salary, Qualifications of the employee etc. These main criteria's will be evaluated and find patterns that will predict the turn over behavior.

II. RELATED WORK

Turn off of a good employee will create uncertainty to both fellow employees and customers who dealt with that employee and organization might lose more employees or customers on that uncertainty. [4] Therefore many researches have been focused on predicting turn over behavior of employees. There are many ways to predict the behavior of an employee in a working environment. Among those, social analytics and human

behavioral pattern recognition takes an important place since many researches have been conducted based on these theories. This method involves interviews, surveys with huge understanding of human behavior for generate a final result. There is a research conducted by Timothy and Peter [5] which involved about 100 managers to find out most common pre turnover behavioral changes of employees. Final outcome was outstanding but still consist with few drawbacks. These methods can be performed only by a person with good observations and skills. Furthermore, a manager who is willing to use these methods must observe each and every employee to notice some of the little changes. Still they won't be able to give much details on the turnover, just whether they are going to leave or not.

Another method is to use a computer generated method for prediction using collected variables. There are many researches on this method using different techniques and variable to maximize the accuracy level of the prediction. Summary of those researches is stated in Table 1.

III. METHODOLOGY

In this research paper, it is focused on predicting the employee turnover using two different sources namely web browsing history and HR data of an employee. Later, these data will be analyzed using different data mining techniques to get the final prediction.

The proposed work involves looking into the private and confidential information of employees. Some could think of this as a huge privacy issue while collecting these data from a particular company. But, there is a certain privacy policy that the companies have. According to that, the higher management of a company has all the privileges to look into employees' private data. The privacy policy that an employee should agree with the company while signing the contracts is as below.

Privacy policy

- [17] [18] An employer has the ability to monitor pretty much anything that an employee access on the company's computer system, even the personal email account of a particular employee.
- [19] [20] Employers can generally monitor, listen in and record employee phone calls on employer owned phones and phone systems.

According to the above mentioned policies, gathering data for this project is not an issue.

a) Employee turnover prediction using web browsing history

Browsing history is one of the major component that we are looking into while making the prediction of employee turnover. Previous researches have been mainly focused on age, gender, salary and many other HR related data but there are very few attempts to use web browsing history to predict the turnover behavior of an employee. But it is proven that the browsing history contains many details of a person. Here we are focusing on important pattern changes in the browsing history before employee leaving the company and based on those learnings, produce a prediction for current employees. Initially the main process has been divided into sub parts.

1. Data Collecting
2. Data Cleaning
3. Data Processing and Find Patterns
4. Making Prediction

Data Collecting

This process contains two major parts,

- Collecting data for finding patterns (involves previous employees)
- Collecting data for making the prediction (involves current employees)

In the process, privacy is a major issue when collecting data. As a solution for this problem, an encrypt mechanism is used when collection data. Since the real identity of a person is not required to make the prediction, all the identity related field have been omitted and the person has given a new code when taking the data set. And also none of these record won't be manually reviewed and the prediction system will only allow users to see final prediction result along with additional information about the turnover, not raw data. This will fix any privacy concerns of the employees. These data are used to find patterns.
Then to make the prediction, we have to collect browsing history for each and every employee currently working in the company. To this process, a third party

application called BrowsingHistoryView v2.05 is being used.

BrowsingHistoryView is a utility that reads the history data of 4 different Web browsers (Internet Explorer, Mozilla Firefox, Google Chrome, and Safari) and displays the browsing history of all these Web browsers in one table. The browsing history table includes the following information: Visited URL, Title, Visit Time, Visit Count, Web browser and User Profile. BrowsingHistoryView allows you to watch the browsing history of all user profiles in a running system, as well as to get the browsing history from external hard drive. You can also export the browsing history into csv/tab-delimited/html/xml file from the user interface, or from command line, without displaying any user interface. Also BrowsingHistoryView allows to do the recording history for remote computers in a network. This is a huge advantage of this application and also we can get the results for any time period we want. However, the records that generating through this application contains many unwanted fields which is not relevant to making the prediction. The final outcome will be save to csv (comma-separated value) file.

- Sample Record of the output of BrowsingHistoryView –

  URL : https://www.google.lk/?gws_rd=ssl
  Title : Google
  Visit Time : 8/2/2017 3:10:50 PM
  Visit Count : 142
  Visited From : http://www.google.lk/
  Visit Type : Auto Top Level
  Web Browser : Chrome
  User Profile : savidya
  Browser Profile : Default
  URL Length : 33
  Typed Count : 0

Data Cleaning

In this sub process, only the necessary fields will be considered for the next step. Here Typed Count, Visit Type, Web Browser, Browser Profile, URL Length and Type Count are not relevant when making the prediction. Further, the Visited URL must me categorize according to the Title.

- Categorizing URLs

In this process a third party API [21] has been used along with Python programming. The API is called Web Shrinker Category API. The Web Shrinker Category API gives developers the ability to lookup the categories that a particular URL, website, domain name, or IP address is categorized as.

1. URLs - Querying the categories for a URL will return the categories specific to that URL, not the domain name. This can be used to analyze the content present on a specific page of a website.
2. Websites / Domain Names - Queries for a domain name, like example.com, will return the main categories associated with that site and its content.
3. IP Addresses - Queries for IP addresses will return the most relevant categories for all of the content we've seen hosted on that IP address. This can be used in situations where you don't know which domain name to lookup but have an IP address.

API allows you to categorize URLs to following categories:

uncategorized, searchenginesandportals, newsandmedia, streamingmedia, entertainment, shopping, vehicles, gambling, informationtech, games, sports, economyandfinance, jobrelated, hacking, messageboardsandforums, socialnetworking, chatandmessaging, mediasharing, blogsandpersonal , health, adult, personals, religion, travel, abortion, education, drugs, alcoholandtobacco, business, advertising, humor, foodandrecipes, realestate, weapons, proxyandfilteravoidance, virtualreality, translators, parked, illegalcontent, contentserver

Since there many URLs which are much alike and falls into same category, we short listed the URLs by the Title. Here the URLs were grouped according to the title.

Ex:

http://www.google.lk/	Google	15
https://www.google.lk/?gws_rd=ssl	Google	20

From these records only the first record will be kept and the Visit Count of the omitting record will be added to the record that we are keeping.

http://www.google.lk/	Google	35

After categorizing the URLs, again these records will be saved to a csv file and send to next step.

b) Employee turnover prediction using HR data

In this part employee layoff prediction is analyzed from the side of human resources data. The data will be collected from both employees in the organization and also from the employees who left the organization. After that based on that data, organizational network will be build. Then using social network analysis and data mining techniques employee layoff patterns are observed. Through those patterns layoff prediction percentage will be given to the current employees.

As the first step data gathering part should be performed. For data gathering we should first discuss with organization and get the details from their human resource department. In this we should come to an agreement with organization to protect the privacy of employee data and use them only for research purpose. In data gathering we are mainly focused on details about employees who left the organization. Employees attributes like sex, age, education level, relationship status, position, department, salary, attendance etc. will be collected. These data will be enough for making our prediction model. But after making prediction model to apply it in current situation we should also collect the data of current employees in the organization. In collecting current employee details also we mainly focused on the employee attributes mentioned earlier.

After collecting the needed data, we should first arrange them in particular order before doing data analysis. From the employee attendance detail sheet as a final output average attendance percentage for each employee should be calculated. After that it should be merged with the employee detail sheet and make one single document. So finally these document will contain following attributes.

-Employee ID - Attendance
-Age -Department
-Sex - Relationship status
-Salary - performance (rank from 1-3)
-Leaved employee or not (0 - stay, 1- leave)
-Role

Then these data should be analyzed using data exploration and visualization techniques in R software before applying them into a model.

Role

In this data set we have mainly 5 roles as CEO, VPs, Directors, Managers, and Individual employees. In here CEO, VPs have different labor market than the other roles and as we don't have many CEOs and VPs including them in the model make no sense. So using our assumptions we can filter out those roles from our data set.

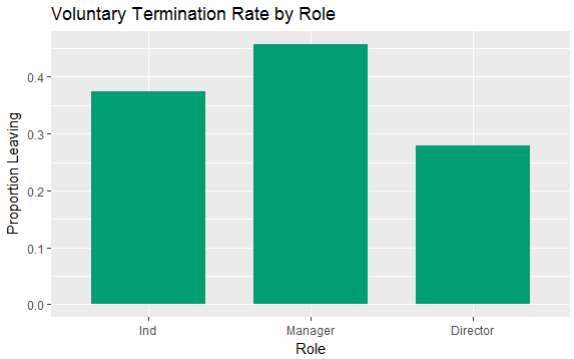

Figure 1 - Voluntary Termination Rate by Role

Performance

If we visualize employee performance with employee proportion of leaving we can see a clear relationship between those two variables. With the increment of employee performance, the proportion of leaving also have been increased. So this type of variables should be definitely should include in our model.

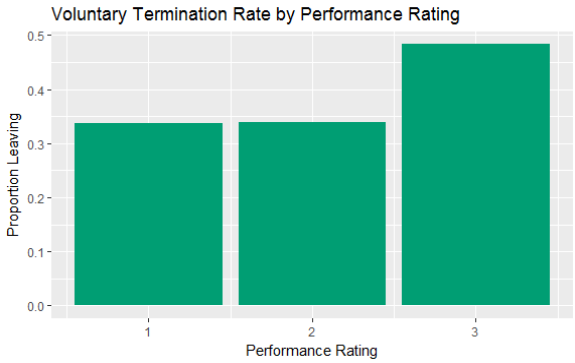

Figure 2 - Voluntary termination rate by performance rating

Sex

According to charts in sex variable also we can have clear relationship between sex and employee turnover. The number of female employee turnover is very higher than the turnover of male employees. So it should be also including in our model.

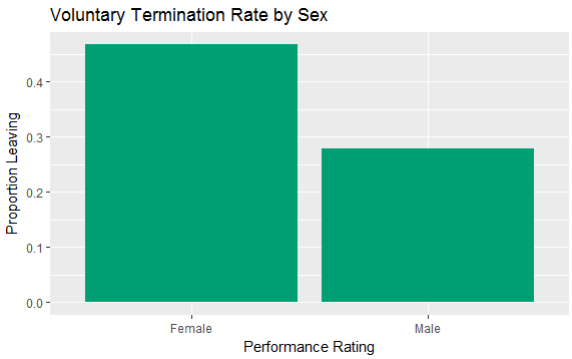

Figure 3 - Voluntary termination rate by sex

Department

In department wise according to chart we can see that employee turnover in sale department is higher than the other departments. So we can assume that there is a relationship between employee turnover and the departments.

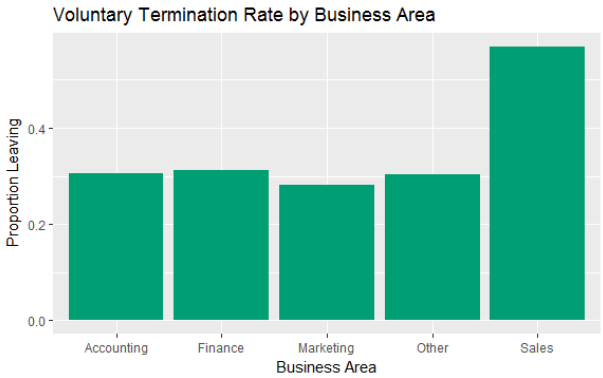

Figure 4 - Voluntary termination rate by department

Salary

We can compare salary with other variables and see the relationship. As an example we can compare it with employees' roles. By comparing with roles we can clearly see that there is a tight relationship between role and salary.

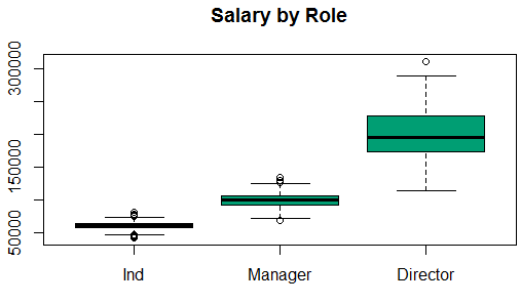

Figure 5 - Salary by role

In practical scenario employee will concern about the salaries of employees in same level. This suggests that looking at salary relative to the median of an employee's role makes more sense. So we can create new predictor variable from the data available. This is also called as "feature engineering".

Attendance

According to graphs attendance variable also have direct relationship with the employee turnover. With the decreasement in the average attendance of employee, the rate of employee turnover is also high.

Employee ID

This variable does not have anything to do with model. But we should include in the model in order to identify each employee uniquely.

After analyzing the data set and making the needed changes we have to split the data set into two as training set and test set. Then the training set is used to train our model and test set is used to test the accuracy of our model.

After that this data will use to make our predictive model. In here we have used the logistic regression modelling technique. Logistic regression model predicts the categorical output of employee staying or leaving. The formulation below shows that we are predicting the employee leave variable (vol_leave) (left side of formula) with the selected set of predictors (right side of formula).

```
Call:
glm(formula = vol_leave ~ perf + role + log_age + sex + area +
    sal_med_diff + sex * area, family = "binomial", data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4564  -0.9119  -0.6072   1.0910   3.0738

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)           8.596e-01  8.136e-01   1.057  0.29071
perf                  4.602e-01  4.392e-02  10.477  < 2e-16 ***
roleManager           6.892e-01  1.523e-01   4.526 6.01e-06 ***
roleDirector         -2.874e-01  4.277e-01  -0.672  0.50168
log_age              -7.119e-01  2.474e-01  -2.877  0.00401 **
sexMale              -1.154e+00  1.474e-01  -7.828 4.96e-15 ***
areaFinance          -9.511e-02  1.224e-01  -0.777  0.43702
areaMarketing        -6.155e-02  1.136e-01  -0.542  0.58809
areaOther            -1.942e-01  1.150e-01  -1.689  0.09124 .
areaSales             1.167e+00  1.073e-01  10.880  < 2e-16 ***
sal_med_diff         -6.418e-05  4.561e-06 -14.073  < 2e-16 ***
sexMale:areaFinance   2.896e-01  2.030e-01   1.426  0.15375
sexMale:areaMarketing 1.537e-01  1.908e-01   0.805  0.42054
sexMale:areaOther     3.521e-01  1.927e-01   1.828  0.06760 .
sexMale:areaSales     2.685e-01  1.727e-01   1.555  0.12001
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 9838.8  on 7399  degrees of freedom
Residual deviance: 8675.5  on 7385  degrees of freedom
AIC: 8705.5

Number of Fisher Scoring iterations: 4
```

Figure 6 - Used formulation

The variables having Pr value less than the .05 have significant impact on the employee leave prediction. So we should definitely include those variables in our model. And also according to above output we can remove some variables from the model as they have high Pr value. This will help to simplify our model.

After simplifying we can build our final regression model. And using it distribution of predicted leave probability can be calculated as follow

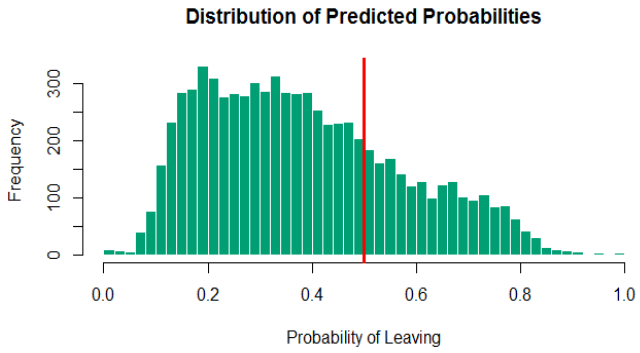

Figure 7 - Distribution of predicted probabilities

Measuring model accuracy using confusion matrix

After building the model we can check it accuracy using the test data set. By using test data set we get the probability of leaving graph as follows.

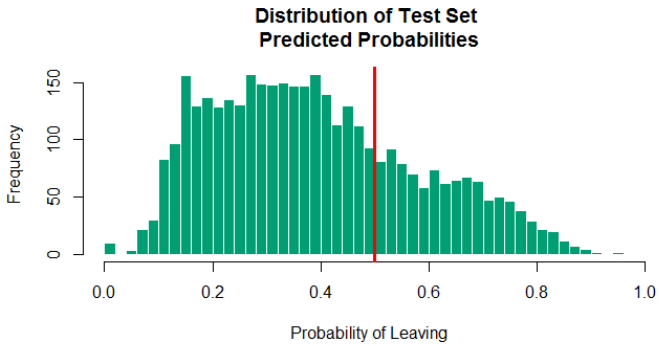

Figure 8 - Distribution of test set predicted probabilities

In here we used the cutoff as 0.5, probability above the 0.5 will consider as leave and probability below 0.5 will consider as stay. To get an accuracy as a percentage we have used the confusion matrix.

```
          0    1  Sum
FALSE  1919  788 2707
TRUE    369  624  993
Sum    2288 1412 3700
>
> sum(diag(accuracy))/sum(accuracy)
[1] 0.6872973
>
```

According to this we were able to get 68% accuracy level for current data set with our model. So using this technique for all the employee probability of leaving is calculated and given as the final output.

AUTHORS

First Author – H.I. Viran Pravinda, Student at Sri Lanka Institute of Information Technology (SLIIT)

Second Author – G.D.M.S. Sathyani, Student at Sri Lanka Institute of Information Technology (SLIIT)

Third Author – T.D.C. Kavinda, Student at Sri Lanka Institute of Information Technology (SLIIT)

Fourth Author – M.A.P.I. Perera, Student at Sri Lanka Institute of Information Technology (SLIIT)

Fifth Author – Dr. Malitha Wijesundara, Senior Lecturer at Sri Lanka Institute of Information Technology (SLIIT)

Sixth Author – Dr. Darshana Kasthurirathna, Senior Lecturer at Sri Lanka Institute of Information Technology (SLIIT)

CONCLUSION AND FUTURE WORK

Coming to conclusion, the main theme of this work is to identify the employees who have a higher risk of leaving the company. The paper gives an overview of different sources that the researchers have used to do the prediction of employee turnover and the final prediction has been made combining each and every source. Researchers worked on this topic and have come up with the most influential factors (emails, call logs, Web browsing history, HR data) that could have an impact on the prediction of employee turnover. They have ignored the factors that are not relevant to the prediction.

Future recommendations have multiple choices likewise predicting the company that the employee will move onto, the reason for leaving the current company (e.g.: Salary issue, not interested in the work) and the probable solutions to keep the employee within the current company. If someone can come up with the above factors within our proposed system, then the accuracy of the system will increase rapidly and it will be much more beneficial when it comes to the company side.

ACKNOWLEDGMENT

We would like to take this opportunity to express our profound gratitude and deep regard to Dr. Malitha Wijesundara (Supervisor) and Dr. Darshana Kasthurirathna (Co-Supervisor) for their guidance and invaluable advices that they gave us throughout the project.

REFERENCES

[1] CCG Resources. "***A company's most valuable asset is its employees.***" Internet: http://ccgresources.com/a-companys-most-valuable-asset-is-its-employees/, [2017-03-14]

[2] Huo-Tsan Chang, Hui-Ju Wu, I-Hsien Ting. "***Mining Organizational Networks for Layoff Prediction Model Construction***" 2009 International Conference on Advances in Social Network Analysis and Mining, 2009, pp. 411 - 416.

[3] Aaron Taube, "**How This Company Knows You're Going To Quit Your Job Before You Do**" Internet: http://www.businessinsider.com/workday-predicts-when-employees-will-quit-2014-11, 2014-11-19, [2017-03-14]

[4] Sergio López Bohle and P. Matthijs Bal and Paul G. W. Jansen and Pedro I. Leiva & Antonio Mladinic Alonso. ***"How mass layoffs are related to lower job performance and OCB among surviving employees in Chile: an investigation of the essential role of psychological contract"***. The International Journal of Human Resource Management [19 February 2016]

[5] Timothy M. Gardner, Peter W. Hom. ***"13 Signs That Someone Is About to Quit, According to Research"*** Internet: https://hbr.org/2016/10/13-signs-that-someone-is-about-to-quit-according-to-research, [2017-08-17]

[6] Chang Youzheng, G. M. ***"Data Mining to Improve Human Resource in Construction Company. International Seminar on Business and Information Management"*** [2008], 1(19), pp. 275 - 278.

[7] Sexton, R. S., McMurtrey, S., Michalopoulos, J.O., & Smith, A. M. ***"Employee turnover: a neuralnetwork solution. Computers & Operations Research"***. [2005], 32(10), pp. 2635-2651. doi: 10.1016/j.cor.2004.06.022.

[8] Valle, M. A., Varas, S., & Ruz, G. A. (2012). ***"Job performance prediction in a call center using a naive Bayes classifier. Expert Systems with Applications"*** [2012], 39(11), pp. 9939-9945. doi: 10.1016/j.eswa.2011.11.126.

[9] Fan, C.-Y., Fan, P.-S., Chan, T.-Y., & Chang, S.-H**.** ***"Using hybrid data mining and machine learning clustering analysis to predict the turnover rate for technology professionals. Expert Systems with***

Applications" [2012], 39(10), pp. 8844-8851. doi: 10.1016/j.eswa.2012.02.005

[10] Huo-Tsan Chang, Hui-Ju Wu, I-Hsien Ting. "***Mining Organizational Networks for Layoff Prediction Model Construction***" 2009 International Conference on Advances in Social Network Analysis and Mining, 2009, pp. 411 - 416.

[11] H. Jantan, A. R. Hamdan, and Z. A. Othman, ***"Towards Applying Data Mining Techniques for Talent Managements"***, 2009 International Conference on Computer Engineering and Applications, IPCSIT vol.2, Singapore, IACSIT Press, 2011.

[12] V. Nagadevara, V. Srinivasan, and R. Valk, ***"Establishing a link between employee turnover and withdrawal behaviours: Application of data mining techniques***", Research and Practice in Human Resource Management, 16(2), 81-97, 2008.

[13] W. C. Hong, S. Y. Wei, and Y. F. Chen, ***"A comparative test of two employee turnover prediction models"***, International Journal of Management, 24(4), 808, 2007.

[14] L. K. Marjorie, ***"Predictive Models of Employee Voluntary Turnover in a North American Professional Sales Force using Data-Mining Analysis"***, Texas, A&M University College of Education, 2007.

[15] D. Alao and A. B. Adeyemo, ***"Analyzing employee attrition using decision tree algorithms", Computing, Information Systems"***, Development Informatics and Allied Research Journal, 4, 2013.

[16] V. V. Saradhi and G. K. Palshikar, ***"Employee churn prediction"***, Expert Systems with Applications, 38(3), 1999-2006, 2011.

[17] Susan M. Heathfield, "**Internet and Email Policy Sample**" Internet: https://www.thebalance.com/internet-and-email-policy-sample-1918869, 2017-01-14, [2017-08-10]

[18] Kara Brandeisky, "5 Things You Didn't Know About Using Personal Email at Work", Internet: http://time.com/money/3729939/work-personal-email-hillary-clinton-byod/, 2015-03-03, [2017-08-10]

[19] Editorial Advisory Board, HR Examiner, Heather Bussing, "**Employee Privacy – What Can Employers Monitor?**", Internet: http://www.hrexaminer.com/employee-privacy-what-can-employers-monitor/, 2011-10-04, [2017-08-11]

[20] Rajesh K, "Monitoring Employee Mobile Phones: What Can Employers Do?", Internet: http://www.excitingip.com/5083/monitoring-employee-mobile-phones-what-can-employers-do/, 2015-09-11, [2017-08-11]

[21] [1]Web Shrinker. "***Website Category API Introduction",*** Internet: https://docs.webshrinker.com/v2/website-category-api.html?python#pre-signed-urls, [2017-08-14]