Communication behavior analysis to understand employee attrition

Abdel-Rahmen Korichi, Hamamache Kheddouci, Taha Tehseen

Communication behavior analysis to understand employee attrition

Abdel-Rahmen Korichi[1,2], Hamamache Kheddouci[1] and Taha Tehseen[2]

***Abstract*— In this paper, we study the behavioral communication patterns of employees in the last few months before they quit their company. Our study is based on Slack communication metadata from two technology companies. We analyse the communication patterns of terminated employees in the last 15 months before they left the organisation. We use different metrics such as the volume of messages sent, the activity in different channels, the average range of communication, and the number of collaborators over time. We also compare the behavior of the employees who are about to quit versus the company average in the same time period. We discover that there is a clear communication pattern for employees who quit with a pronounced decrease of engagement in the last 5 to 6 months before they left. We then build a machine learning model to predict if an employee is at risk of leaving the company within the next 6 months.**

I. INTRODUCTION

One of the most common ways for organizations to measure their employees' engagement is through surveys. However, the rise of digitalization and remote working, accelerated by the COVID-19 crisis, has created an environment where millions of users from thousands of organizations now use tools like Slack or Microsoft Teams for their daily communication. This offers new ways to measure employee engagement and preempt employee attrition.

In this research, we aimed to leverage all this untapped information and get meaningful insights about employee engagement. We explored Slack channel data from two of Panalyt's clients and tried to deduce conclusions about employee engagement, especially for those who left the organization, in the last 15 months before they quit.

II. RELATED RESEARCH

Employees are a company's biggest asset and losing these valuable contributors can have serious ramifications for it. According to research by PwC (an international global services firm), each leaver costs a company 1 to 1.2 times their annual salary. The cost to an economy is even higher with around $27 billion lost in the US economy because of inefficient hiring practices [1]. Traditional research on the reasons behind employee attrition has focused on a number of different factors. For some employees their age and education are important contributing factors. Younger, relatively inexperienced and highly educated employees tend to have lower satisfaction about their work. They harbor lower commitment to the organization and these negative attitudes are associated with turnover intentions. Studies have found that females tend to have a higher attrition rate than males. This can be a result of traditional social expectation of women to give birth and also take care of the family [2].

Similarly, Firth et al (2007) [3] add that 'job stressors' such as overtime hours, or job ambiguity have a direct psychological impact on employees which can be a precursor for their attrition. Keeping track of overtime hours, as well as the working relationships between managers and employees can lead to a decrease in dissatisfaction with their employment. In addition, it increases motivation in employees and reduces odds of attrition.

Another increasingly relevant factor contributing to employee attrition is the communication pattern within a company. In a study conducted by Gloor et al. [4], they examined the correlation between managers' communication patterns and their likelihood of leaving the organization. The findings revealed that managers who are approaching their departure exhibit initial high engagement and centrality in organizational communication. They are also responsive and require fewer follow-up messages to elicit a response. However, in the months leading up to their termination, their responsiveness decreases, and multiple reminders are necessary to elicit a reply.

Feeley et al [5] found that close relationships developed at work were significant predictors of turnover. Their results showed that, among a population of fast-food restaurant employees, the one with more "out-degree" links with friends, i.e strong relationship with co-workers outside of the workplace, had a lower chance of leaving.

While there are some useful metrics devised by these earlier studies to measure risk of attrition, we find that some of these metrics are often unactionable because of the complexity behind their calculation. We need to be mindful that the main intended audience for these metrics would be HR managers and other people in leadership positions. Since many HR decisions need to be taken quickly, the simpler and reproducible a metric is, the easier and more efficient it is to adopt.

In terms of attrition prediction related to employee communication metadata, the closest study that we found is from Patil et al [6]. They conclude that employee communication behavior changes in the last 3 weeks before they quit, with fewer email communication to selected individuals. They trained a classifier to predict attrition with a moderate accuracy of 60-65% on a large company dataset.

Although the study is insightful, we argue that the employees' decision to quit is on average much earlier than 3 weeks before the termination date, and we will prove that

[1]Université Claude Bernard, Lyon 1
[2]Panalyt Pte. Ltd.

in our study, with higher performances. Also, we will only use actionable and easily understandable managers metrics to evaluate those trends, so that relevant measures can be taken to mitigate the risk of attrition and help maintain their workforce's size and satisfaction.

III. DATA COLLECTION AND PROCESSING

A. Population

The population under study is from two Japanese tech companies. Dataset A is comprised of the Slack communication data of 479 employees, from January 1st 2019 to March 31st 2022, including 50 employees who quit during that period. Dataset B is comprised of the Slack communication data of 587 employees from January 1st 2020 to October 31st 2022, including 94 employees who quit during that period. This data includes their public and private channels, but does not include any direct messages. We do not include direct messages as the company's Slack subscription does not allow access to direct messages for extraction due to privacy reasons. For the purpose of this study, we look at three populations:

1) *Leavers*:
   These are employees that have a valid termination date in the time frame we are studying. In addition, we only choose leavers that have a minimum tenure of 12 months *i.e.* they stayed employed for at least 12 months before their termination. This is to make sure that the employees have had time getting used to the company and we have enough data points to see their behavior change over time.
2) *The entire population*:
   The total population, including the leavers we have defined above. This population helps to identify the company wide communication trends over time.
3) *The active population*:
   The total population, excluding the leavers we have defined above. This population serves as a benchmark to compare leavers versus active employees.

B. Description of the communication data

The data is obtained directly from Slack in the form of a series of files for each channel. Each file has the channel name as its filename and contains a series of text files for every day of communication in the channel. Each of these text files contains a list of messages sent where each message has the following fields:

- type: To indicate that the record is a message.
- user: An alphanumeric ID of the employee that sent the message.
- text: The content of the message.
- ts: UNIX timestamp for when the message was sent.
- channelID: the ID of the channel where the message has been posted.

As explained in the Results Section, we do not include any text analysis on the message content. Thus, all content is redacted from the data before using it. For this purpose, we created an anonymizer application that reads in each file and iteratively removes any content from it. This includes the body of the message, as well as any links, URLs or descriptions included within. The application then exports the redacted files which we concatenate together and transform into a single database table using a Python script.

With the data imported, we begin our analysis on the outcome. We define the following groups of metrics:

1) Volume of messages sent:
   The metric we measure here is the total number of unique messages sent by an employee in a month. This is the most basic way of measuring communication volume with more connected employees tending to have more messages sent.
2) Working hours:
   This group of metrics measures the timing of communication. In particular we capture 3 metrics:
   a) When employees are starting their communication. This is the time they send their first message of the day.
   b) When employees are ending their communication. This is the time they send their last message of the day.
   c) The range of communication. This is the total span of their communication and is the difference in hours and minutes of the previous two metrics.

   This set of metrics is another angle at looking to look at engagement levels among employees. We wanted to observe if communication spans changed as employees neared termination with our intuition suggesting that they would shrink for employees on the way out.
3) Channels Activity:
   The metric we measure here is how many channels is the employee actively a part of. An employee being active in a channel means that they sent at least 5 messages in it in the same month. This threshold is set to avoid situations where an employee sends a one-off message in a channel without the intention of actively contributing to it. We experimented with various thresholds, such as 10 messages and 50 messages, and determined that 5 was the optimal number for capturing clearer trends over time. This metric is a good measure for diversity of communication and how involved the employee is in separate communication groups.
4) Number of collaborations:
   The metric we measure here is how many frequent collaborators does the employee have. This is essentially a measure of the communication sphere of an employee. We define two employees being active collaborators if they are both active in at least one channel, i.e. if they both have sent at least 5 messages in the same channel.

   For each defined metric, we perform calculations using two different approaches. Firstly, we calculate the

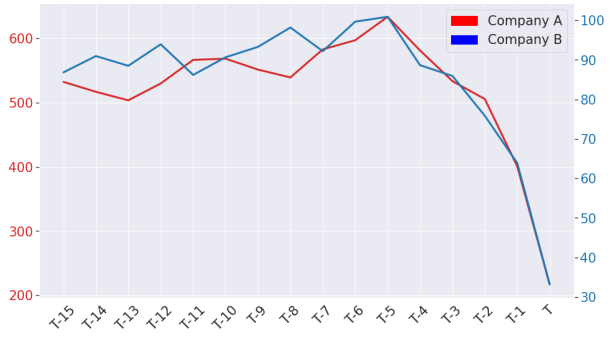
metric for employees who leave the company on a monthly basis during the 15-month period leading up to their termination. We chose a 15-month time frame as it is substantial enough to reveal meaningful trends over a relatively long period, while also ensuring that we include an adequate number of employees with varying total tenures in the study, without excluding those with shorter durations of employment.

Secondly, we calculate a benchmark as follows: we take the calculated metric for each terminated employee and subtract the same metric calculated for the active population in that particular month. This helps us observe each metric for leavers against the active population at that point in time.

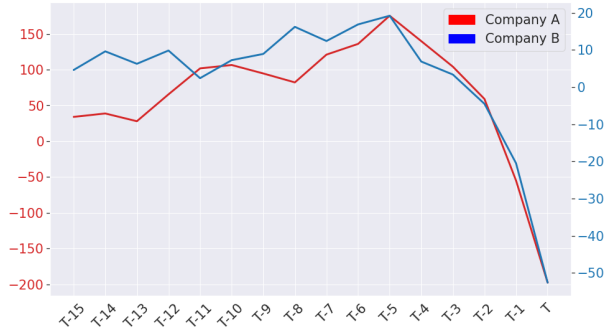
IV. DESCRIPTIVE ANALYSIS

A. Volume of messages sent

We start by studying the volume of messages sent, which we calculate by looking at the average count of messages sent by month for each employee.



(a) Average count of messages sent by leavers for each month before their eventual termination date.



(b) Average count of messages sent by leavers for each month before their eventual termination date benchmarked against average messages sent by the active population for that month.

Fig. 1: Analysis of volume of messages sent

In Figure 1a, we focus only on the terminated employee and especially the last 15 months before they quit. We see that the average messages sent increases slightly for both companies. For Company A, it increases from 550 messages sent 15 months before their eventual termination to 625 messages sent 5 months before their eventual termination. Similarly for Company B, it increases from 85 messages sent 15 months before termination to 95 messages sent 6 months before termination. At this point, the average number of messages sent constantly decreases at a sharp rate as they get closer to their termination date. This indicates a gradual disengagement from 5 to 6 months before employees quit.

When benchmarked against the active population (Figure 1b), we see that the leavers sent a higher number of messages from 15 to 5 months before their eventual termination compared to their counterparts. However, similar to the previous graph, the average number of messages benchmarked against the active population starts rising at 8 months before termination, reaching a peak at 5 months before termination. At this point the leavers have sent an average of around 160 messages more than the company average in Company A and 20 more in Company B. After this point we see a sharp decreasing trend, similar to the previous figure, and the average for leavers dips below the average for the active population eventually decreasing to 200 lesser messages for Company A and 50 lesser messages for Company B at the time of termination. What we understand from this graph is that, on average, leavers tend to be more active than the active population and the gap widens until it reaches its peak 5 months before their termination. There is a similar inflection point in Figure 1a and Figure 1b that could indicate a pattern of overworking individuals.

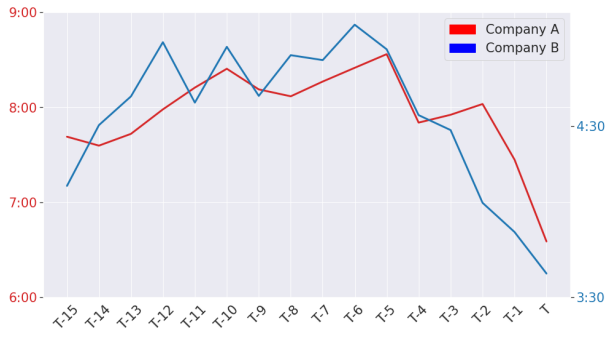
B. Working hours

In the followings graphs, we study the range of communication times for messages sent in the active population and for messages sent by leavers.

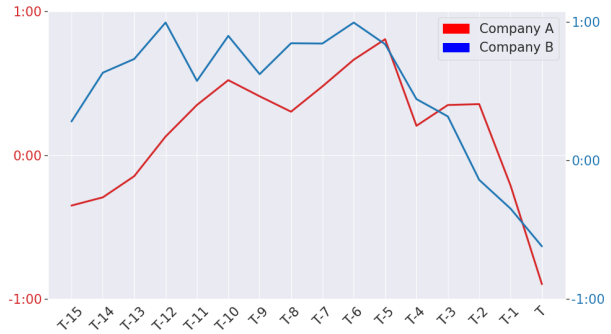
The average total communication (Figure 2a) for employees who leave their positions exhibits a steady increase starting from 15 months before termination until approximately 5 months prior to their termination date. Subsequently, there is a decreasing trend in the average total communication during the last 5 months, with a minor spike occurring around 2-3 months before termination. Finally, the average total communication reaches its lowest point towards the end of their employment, specifically 6 hours and 30 minutes for Company A and 3 hours and 45 minutes for Company B.

We can observe that on average, before employees quit, they start communicating more until they reach a point around 5 months before termination where the trend starts to reverse.

In Figure 2b, we observe that leavers are communicating lesser overall 15 months before their termination. Both these averages increase compared to the benchmark, reaching highs in the window 5-7 months before termination. The highest average positive gap in total communication between leavers and the active population is 5 months before termination and is about 40 minutes. The highest positive difference in last message sent is 7 months before termination with leavers on average sending their last message 15 minutes later than the active population average. Both these averages then decrease to fall below the active population average. Leavers were on average communicating an hour lesser than the active population average.

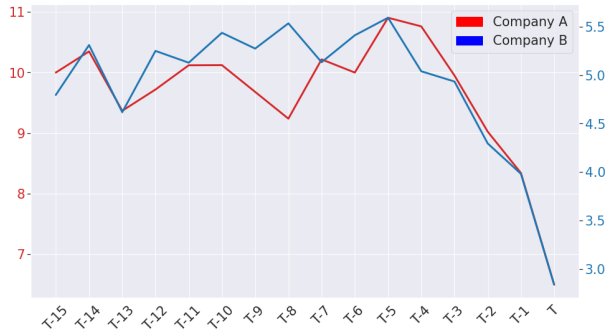


(a) Average total hours of communication for each terminated employee for each month leading up to their eventual termination date.

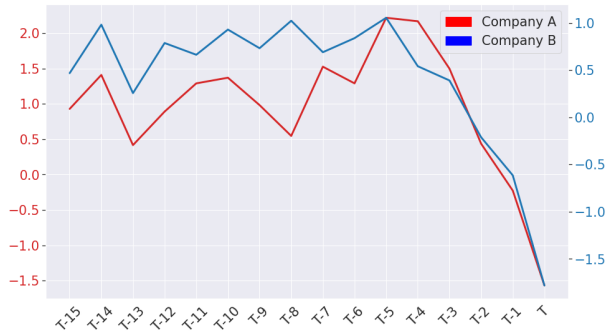


(b) Average total hours of communication for terminated employee benchmarked against the averages for the active population for that month for each month leading up to their eventual termination date.

Fig. 2: Analysis of communication times



(a) Average count of active channels used by leavers for each month leading up to their eventual termination date.



(b) Average count of active channels used by leavers benchmarked against the average for the active population for that month for each month leading up to their eventual termination date.

Fig. 3: Analysis of active channels

We find again a pattern of employees overworking compared to the average employee, especially 5 to 10 months before their termination.

C. Channels activity

Here we study the number of channels in which employees are active.

The average count of active channels per leaver (Figure 3a) starts from around 10 for Company A and 9.5 for Company B, increasing to till 7 months before termination where it stabilizes at around 10 for Company A and 5 for Company B. It then starts a downward trend through the rest of the time period reaching the lowest count at the time of termination. The average count of channels increases to under 7 for Company A and under 3 for Company B.

The average count of channels actively used by leavers benchmarked against the active population average (Figure 3b) starts at 0.5-1 more channels 15 months before termination and starts widening slightly until 5 months before termination. At this point the gap begins shrinking rapidly and eventually we see leavers being involved in lesser channels. At the time of termination, the leavers are involved in 1.5 channels lesser than the active population.

Employees being active in more channels indicates that not only are they communicating more but there are also likely to be involved in more projects for the organization. This suggests again that employees who quit tend to overwork, or are over tasked compared to the rest of the company, and this is apparent a few months before they leave.

D. Number of collaborations

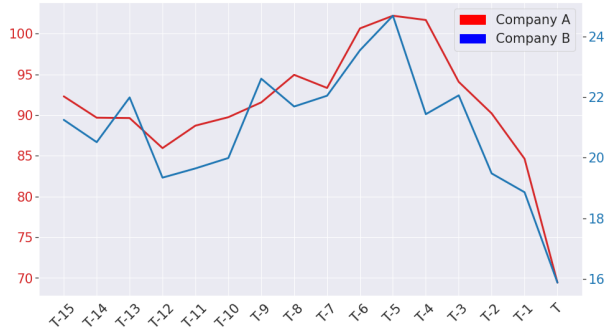
In this last section, we analyze the average number of collaborations for the employees in the organization.

The average count of close collaborators for each leaver (Figure 4a) shows an increasing trend until 5 months before termination. From there the average goes on a rapid decreasing reaching the lowers count at the time of termination.

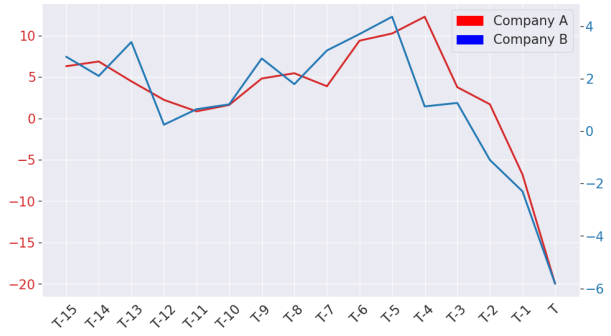
For leavers, the average count of collaborators benchmarked against the active population starts with a positive difference of 5 for Company A and 3 for Company B 15 months before termination (Figure 4b). It increases to 10 for Company A and 4 for Company B at 5 months before termination. For the remaining months the number of collaborators decreases rapidly and each terminated employee is involved in active conversation with 20 lesser collaborators for Company A and 6 lesser collaborators for Company B. We see again an inflection point at 4 months before the termination date.

E. Conclusion of the analysis

The analysis of the four metrics - volume of messages sent, working hours, channels activity and collaboration - all



(a) Average count of active collaborators for each terminated employee for each month leading up to their eventual termination date.



(b) Average count of active collaborators for each terminated employee benchmarked against the average for the active population for that month for each month leading up to their eventual termination date.

Fig. 4: Analysis of collaborations

show that there is a clear trend between 5 to 6 months before the termination and the termination date.

The average trend is monotonic in most cases (volume of messages sent, channels activity and collaboration). In the last 5 to 6 months, on average, employees send less messages (monotonic downward trend), see their communication range reducing, are active in less channels (monotonic downward trend), and communicate to less people (monotonic downward trend). What we notice is that people disengaging with their organization tend to have a pattern of isolation in their last months before they leave.

A second observation is that we can see an inflection point for all metrics. The inflection point is 5 months before termination for volume of messages sent, working hours and collaboration, and 7 months before termination for channels activity. This suggests that those employees tend to disengage after a high peak of activity, and this might be part of the reason why they want to leave.

Finally, when compared to the active population, we see that employees about to quit tend to be on average more active and engaged in more channels. This could indicate that those employees are hard workers and/or are involved in more projects for the company. This insight along with the behavior mentioned in the earlier paragraphs might be a pattern of gradual overwork for employees about to quit, followed by a strong reduction of engagement in the company. As a business leader, these insights can be particularly useful in preventing or reducing attrition by making sure employees are not being overworked either with too many collaborators and projects or that they are not working too many extra hours.

Following this analysis, a question that arises is: is it possible to identify disengaging employees and about to quit, to potentially reduce attrition in a company by identifying unhealthy communication patterns and take preemptive actions?

V. PREDICTIVE ANALYSIS

A. *Features and labels*

Our method to answer that problem is through building a classification model.

The features used to build the classification model are the same as those described in the previous sections. For each employee, in the last 15 months preceding his termination, we compute:

- Average volume of messages sent
- Average time of the first message sent
- Average time of the last message sent
- Average communication range
- Average number of channels where the employee is active
- Average number of collaborators

In addition to those features and as used during the prior analysis, we add the same features but we subtract from it the corresponding active employee average for each month. At this stage, we have the original 6 features defined above plus 6 other features that we call benchmarked features.

Finally, we add lag features: these are values at prior time steps. We had the best results when adding two and four months lag. So for each of the twelve features already computed, we now have the same feature as of two and four months before, which makes 36 features.

For the labels, we define *Employees at risk* a discrete variable. Records where the number of months before the termination date are less than 6 are labelled as 1 (at risk), and records where the number of months before the termination date are greater than 6 are labelled as 0 (not at risk):

$$\begin{cases} 1 \; if \; \# \; months \; before \; the \; termination \; date <= 6 \\ 0 \; if \; \# \; months \; before \; the \; termination \; date > 6 \end{cases}$$

B. *Results*

For company A, we end up with 675 monthly records, from which 254 are labelled at risk and 421 are labelled not at risk. For company B, we have 1195 monthly records, from which 495 are labelled at risk and 700 are labelled not at risk.

For each company, we apply a 10-fold cross-validation on different classifiers to provide a robust estimate of the performance of our model: XGBoost, Linear regression (LR), Random forest (RF) and Gaussian Naive Bayes (GNB).

To avoid any data leakage, we make sure that different records from the same employee are not used in both the training datasets and the testing datasets.

Here are the results for each company A and B:

TABLE I: Predictions for company A

Company	Classifier	Accuracy (std)	Precision (std)	Recall (std)
A	XGBoost	72% (0.07)	63% (0.11)	68% (0.08)
	LR	55% (0.07)	57% (0.07)	59% (0.13)
	RF	54% (0.04)	57% (0.04)	63% (0.12)
	GNB	55% (0.06)	56% (0.06)	73% (0.09)

With 72% accuracy, 63% precision and 68% recall, the model with the highest performance for company A is the XGBoost classifier. The Gaussian Naive Bayes classifier has a higher recall but a much lower precision.

TABLE II: Predictions for company B

Company	Classifier	Accuracy (std)	Precision (std)	Recall (std)
B	XGBoost	71% (0.05)	63% (0.07)	74% (0.06)
	LR	66% (0.03)	70% (0.04)	71% (0.04)
	RF	58% (0.02)	61% (0.03)	69% (0.05)
	GNB	57% (0.03)	59% (0.02)	81% (0.08)

For company B, XGBoost has the highest accuracy and recall (71% and 74% respectively), but the logistic regression classifier is not far behind and has a higher precision (70% for the logistic regression versus 63% for XGBoost), which makes it slightly more reliable to identify people at risk.

Overall, we can conclude that for two different companies, there is a correlation between the people leaving and their communication behavior.

In addition to that, by using a libary like SHAP (SHapley Additive exPlanations) or LIME - both popular for model explainability - it is even possible to explain the output of our machine learning model, and understand what are the top features that drive the output at risk, either at the individual level, or at the group level.

For example, in the case of company A, the primary factor influencing attrition is the average time of the last message sent. This is followed by the average time of the last message sent four months prior.

VI. CONCLUSIONS AND FUTURE WORKS

This study helps deduce very insightful conclusions with a limited amount of metrics. We proved using two different and independent companies' communication real datasets that, on average, there is an inflexion point and a clear trends from 5 to 6 months before an employee is about to quit to his termination date.

We also built a machine learning classifier and we have shown that it is possible to predict very early if an employee is about to leave.

As we proved in this study that communication metadata play an important role in predicting employee attrition, we are confident that by combining communication metadata with HR data (age, tenure, manager, salary, performance, teams, departments, job title, etc.), we could understand employee's network better - e.g. managers versus peers - and build a model with higher performances.

From a psychological point of view, it would be interesting to compare when employees take a conscious decision of quitting versus when they actually start disengaging in their interactions.

Another useful information to further develop the study would be to know the termination reason of employees who quit. This piece of information is often recorded by organizations, and we might extract more nuanced patterns from terminated employees' communications.

There are nonetheless questions that arise about the ethics of such practices, although those questions are mitigated by the fact that we protect the privacy of employees to an extent by never looking at their private content and what the employees actually say. We recommend interested organizations to be truly transparent with their employees about what their are doing and the reasons behind. They should also keep the insights deduced from such studies at a macro and not individual level. One could find for example that for one department people are overworking and have a high level of attrition compared to other departments instead of looking for such trends among specific individuals.

REFERENCES

[1] Marsden, T. (2016), "What is the true cost of attrition?", Strategic HR Review, Vol. 15 No. 4, pp. 189-190. https://doi.org/10.1108/SHR-05-2016-0039

[2] Zhang, Y. (2016) A Review of Employee Turnover Influence Factor and Countermeasure. Journal of Human Resource and Sustainability Studies, 4, 85-91. doi: 10.4236/jhrss.2016.42010.

[3] Firth, Lucy & Mellor, David & Moore, Kathleen (Kate) & Loquet, Claude. (2004). How Can Managers Reduce Employee Intention to Quit?. Journal of Managerial Psychology. 19. 170-187. 10.1108/02683940410526127.

[4] Gloor, P. A., Fronzetti Colladon, A., Grippa, F., & Giacomelli, G. (2017). Forecasting Managerial Turnover through E-Mail Based Social Network Analysis. Computers in Human Behavior, 71, 343-352. http://dx.doi.org/10.1016/j.chb.2017.02.017

[5] Thomas Hugh Feeley, Jennie Hwang & George A. Barnett (2008) Predicting Employee Turnover from Friendship Networks, Journal of Applied Communication Research, 36:1, 56-73, DOI: 10.1080/00909880701799790

[6] Patil, Akshay & Liu, Juan & Shen, Jianqiang & Brdiczka, Oliver & Gao, Jie & Hanley, John. (2013). Modeling Attrition in Organizations From Email Communication. Proceedings - SocialCom/PASSAT/BigData/EconCom/BioMedCom 2013. 331-338. 10.1109/SocialCom.2013.52.