ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 11, November 2016

A Survey on Mining Social Media Content for Analysing Employee's Working Experiences

Anjali Pawar[1], Rhiddhi Khartadkar[2], Sankalp Kale[3], Kiran Naik[4], Priyanka R. Rajmane[5]

Department of Computer Engineering, MES College of Engineering, Savitribai Phule Pune University

Pune, Maharashtra, India

ABSTRACT: The enterprise data mining applications involve complicated data such as multiple large heterogeneous data sources, user preferences, and business employment impact. A single process or one-step mining often limit the discovery of informative knowledge. It might also be time and space consuming, At times it is not possible to join relevant large data sources for mining patterns consisting of multiple aspects of information. It becomes important to develop effective approaches for mining patterns that combines necessary information from multiple relevant business lines that caters for real business settings and decision-making actions and not just providing a single line of patterns. In the recent years we have seen increasing efforts on mining more informative patterns, e.g., integrating frequent pattern mining with classifications to generate frequent pattern-based classifiers. This paper not only presents a specific algorithm but also builds on our existing works and proposes combined mining as a general approach to mining for informative patterns combining components from either multiple data sets or multiple features or by multiple methods on demand. We can summarize general frameworks, hadoop library, hadoop word count, Map reduce and we take employees database. In this paper first we take employee complex database for input then apply all the process then our application give result with good and bad rating

KEYWORDS : Prediction, Classification, FP Growth, Map Reduce, Naive Bayes.

I. INTRODUCTION

Automated prediction of trends and behaviors: Mining automates the process of finding predictive information in a large database. There are questions which require hands on analysis. Such questions can now be directly answered from the data. An example of a predictive problem is targeted marketing. Our aim is to achieve deeper and finer understanding of employee's experiences with respect to their learning-related issues and problems. In order to determine what employee problems a post indicates is a more complicated task than to determine the sentiment of a post. Therefore, our study requires a qualitative analysis, and it is not possible to do in a fully unsupervised way. [1] Sentiment analysis not applicable to our study. In our study, we implemented a multi-label classification model where we allowed one post to fall into multiple categories at the same time. Our work extends the scope of data-driven approaches in employment such as learning analytics and employment data mining. Traditionally, employment researchers have been using methods such as surveys, interviews, focus groups, company activities to collect data related to employees' learning experiences. Such methods are extremely time consuming and hence cannot be duplicated or repeated with high frequency. [2]

The emerging field of learning analytics and employment data mining has focused on analyzing structured data obtained from course management systems (CMS), company technology usage, or controlled online learning environments to inform employment decision making. However, to the best of our knowledge, there is no research found to directly mine and analyze employee- posted content from uncontrolled spaces on the social web with the clear goal of understanding employee' learning experiences. The drawbacks are In our study, through a qualitative content analysis, we found that employees are largely struggling with the heavy study load, and are not able to manage it successfully. [3] Heavy working load leads to many consequences including lack of social engagement, sleep problems, and other psychological and physical health problems. Our work is only the first step towards revealing actionable insights from employee-generated content on website in order to improve employment quality. We extend

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 11, November 2016

the proposed algorithm which analysis the employee's learning experiences by giving solutions to their problems. The higher employment conversation has shifted from simply ensuring access to on that focuses on success, supporting employees through completion and readiness for careers, citizenship and life. This system will help in improving transparency, accountability and equity in higher employment. We are making a system that will generate report of a company which includes company performance, principles,reviews or feedbacks given by the employee and rank that company on the basis of these parameters. [4]

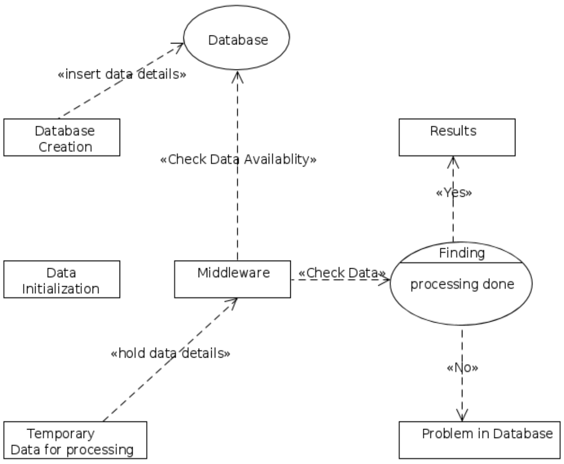

Figure 1: Basic Model

II. LITERATURE SURVEY

The research goals of this study are :

1)To demonstrate a workflow of website data sense-making for employment purposes, integrating both qualitative analysis and large-scale data mining techniques and

2)To explore employee' informal conversations on Website, in order to understand issues and problems employee encounter in their job experiences.

These studies have more emphasis on statistical models and algorithms. They cover a wide range of topics popularity prediction, event detection, topic discovery and post classification. Amongst these topics, post classification is most relevant to this study. Popular classification algorithms include Naïve Bayes, Decision Tree, Logistic Regression, Maximum Entropy, Boosting, and Support Vector Machines (SVM). [5]The existing studies found on post classification are binary classification on relevant and irrelevant content or multiclass classification on generic classes such as news, events opinions, deals and private messages. Analysis of sentiments can be very useful for mining customer opinions on products or companies through their reviews or online posts. It finds wide adoption in marketing and customer relationship management (CRM). We have chosen to focus on employees' posts on Website about problems in their job experiences mainly because:

1. Companies have been struggling with student recruitment and retention issues. Engineering graduates constitute a significant part of the nation's future workforce and have a direct impact on the nation's economic growth and global competency.

2. As per understanding of issues and problems in employee' life, the policymakers and educators can make more informed decisions on proper interventions and services that can help employee overcome difficulties in job.

III. MOTIVATION AND PROPOSED WORK

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 11, November 2016

We chose to focus on employees' posts on Website about problems in their company experiences mainly because:

1) Companies have been struggling with student recruitment and retention issues.

2) Engineering graduates constitute a significant part of the nation's future workforce and have a direct impact on the nation's economic growth and global competency

3)As per understanding of issues and problems in employee' life, the policymakers and educators can make more informed decisions on proper interventions and services that can help employee overcome barriers in learning.

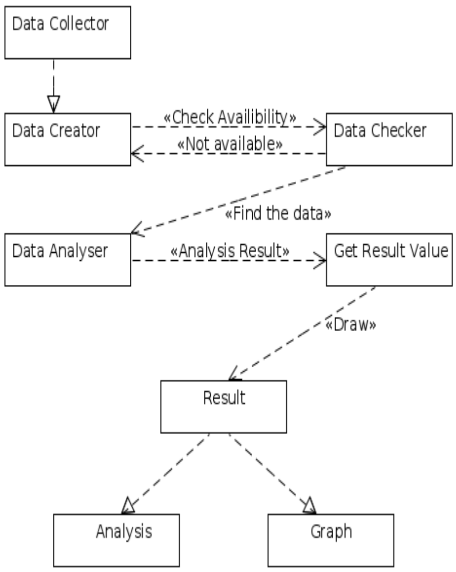

Figure 2: Working of the system

IV. MODULE DESCRIPTION

A. DATA COLLECTION

Collects all the information from the different employees posted their comments in website.s

B. DATA CLUSTERING

In this module, the raw data is clustered by using clustering algorithm. This algorithm starts with single cluster. Every point in a database is a cluster. Then it groups closest points into separate clusters, and continues until only one cluster remains. The computation of clusters calculated with help of distance matrix. The algorithm generates cluster feature tree while scanning the dataset. Each entry in the CF tree represents the cluster of objects and is characterized by triple (N, LS, SS).

C. DATA CLASSIFICATION

After clustering the data in different clusters based on the content, we use FP Growth classification algorithm. One popular way to implement multi-label classifier is to transform the multi-label classification problem into multiple

IJIRSET

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 11, November 2016

single-label classification problems. One simple transformation method is called one-versus-all or binary relevance. The basic concept is to assume independence among categories, and train a binary classifier for each category. All kinds of binary classifier can be transformed to multi-label classifier using the one-versus-all heuristic.

D. SUGGESTIONS AND FEEDBACK

After classification, finally we show results in good and bad format. Also we show graph.

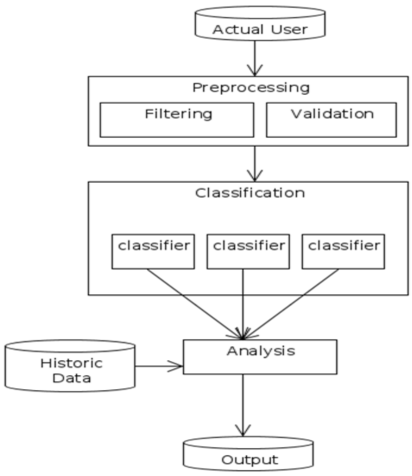

Figure 3: System Architecture

V. IMPLEMENTATION

A. DATA COLLECTION

From website we collect data and copied into excel sheet then we take that sheet as a input.

B. INDUCTIVE CONTENT ANALYSIS

Because website content like posts contain a large amount of informal language, sarcasm, acronyms, and misspellings, meaning is often ambiguous and subject to human interpretation. Rost et.al in their volume of edition stated that in large scale website data analysis, faulty assumptions are likely to arise if automatic algorithms are used without taking a qualitative look at the data. We concur with this argument, as we found no appropriate unsupervised algorithms could reveal in-depth meanings in our data.

There were no pre-defined categories of the data, so we needed to explore what employees were saying in the posts. Thus, we first conducted an inductive content analysis on the dataset. Inductive content analysis is a popular qualitative research method for manual analyzing of text content. Three researchers collaborated on the content analysis process.

C. DEVELOPMENT OF CATEGORIES

The lens we used in conducting the inductive content analysis was to identify what are the major worries, concerns, and issues that employees encounter in their job and life. Researcher A read a random sample of 2,000 posts from the 19,799 unique #employee Problems posts, and developed 13 initial categories including: curriculum problems, heavy work load, work difficulties, imbalanced life, future and carrier worries, lack of gender diversity, sleep problems, stress, lack of motivation, physical health problems, nerdy culture, identity crisis, and others. These were developed to

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 11, November 2016

identify as many issues as possible, without accounting for their relative significances. Researcher A wrote detailed descriptions and gave examples for each category and sent the codebook and the 2,000-posts sample to researchers B and C for review. Then, the three researchers discussed and collapsed the initial categories into five prominent themes, because they were themes with relatively large number of posts.

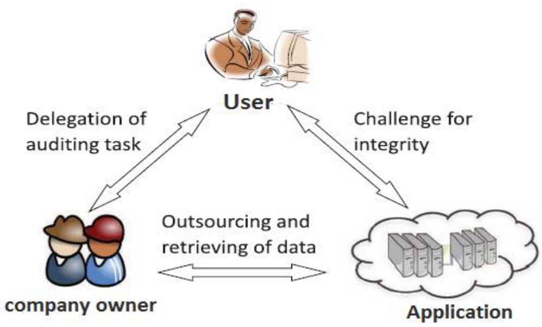

Figure 4: User and Application

We found that many posts could belong to more than one category. For example, "*This could very well turn into an all nighter...f*** you work report #nosleep*" falls into heavy work load, lack of sleep, and negative emotion at the same time. "*Why am I not in good company?? Hate being in company. Too much stuff. Way too complicated. No fun*" falls into heavy work load, and negative emotion at the same time. So one post can be labeled with multiple categories. This is a multi-label classification as opposed to a single-label classification where each post can only be labeled with one category. The categories one post belongs to are called this post's labels or label set.

D. RELEVANT MATHEMATICS ASSOCIATED WITH THE PROJECT

Input Set:-
For creating a new project P the set is represented as follows:-
P = L, UA1, UA2, UA3, UA4 , D
Attributes in input file = company infrastructure, salary, lab, project, work pressure, development enviroment, etc

Output set:-
PC c represents created, saved project in IDE.
Process/Functions:- Processes / Functions

Support:-The support supp(*X*) of an attribute *X* is defined as the proportion of transactions in the data set which contain the attribute.
supp(X)= no. of attributes which contain the comments X / total no. of transactions.
Confidence:-
conf(X->Y)= supp(X U Y)/supp(X).

Provide machine id for each products:

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 11, November 2016

Lets Make sets for each attributes:

Transaction ID	Computer Infrastructure	Lab	Salary	Projects
1	1	1	0	0
2	0	1	1	0
3	0	0	0	1
4	1	1	1	0
5	0	1	0	0

Table 1:Attributes

(infrastructure , 20)	(Lab , 25)	(salary , 22)	(projects , 33)
(infrastructure , 18)	(Lab , 27)	(salary , 32)	(projects , 37)
(infrastructure , 32)	(Lab , 20)	(salary , 33)	(projects , 38)
(infrastructure , 22)	(Lab , 19)	(salary , 20)	(projects , 31)
(infrastructure , 31)	(Lab , 24)	(salary , 19)	(projects , 30)

All five of these output streams will be fed into these reduce tasks, which combine the input results and output a sigle value producing a final result set as follows.

(infrastructure , 32) (Lab , 27) (salary , 33) (projects , 38)

E. CLASSIFICATION RESULTS

Assume From the inductive content analysis stage, we had a total of 2,785 #employees Problems posts annotated with multi categories. We got 70% of the 2,785 posts for good (1,950 posts), and 30% for testing (835 posts) for bad.

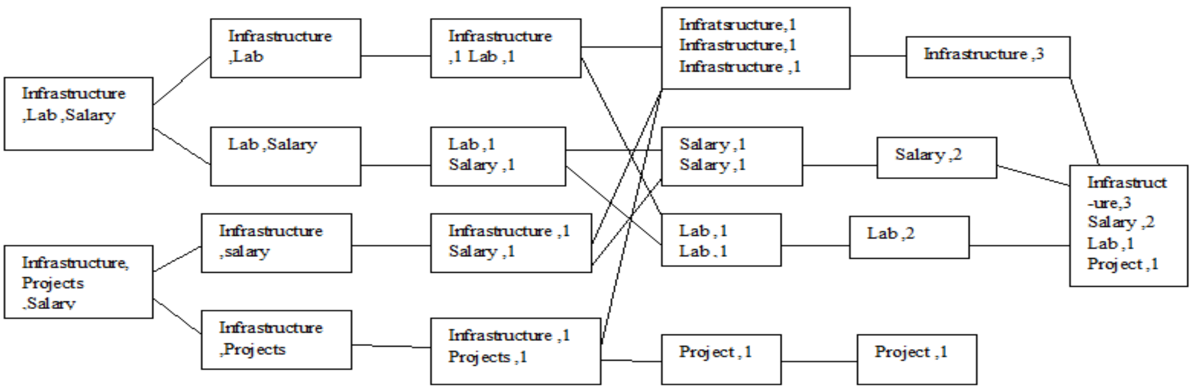

Figure 5: Mapping

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 11, November 2016

VI. CONCLUSION

Our study is beneficial to employees in company analytics, employment data mining, and learning technologies. It provides a workflow for analyzing website data for employee experience purposes that overcomes the major limitations of both manual qualitative analysis and large scale computational analysis of employee-generated textual content. Our study can inform employee administrators, practitioners and other relevant decision makers to gain further understanding of employee' company experiences. As an initial attempt to instrument the uncontrolled website space, we propose many possible directions for future work for researchers who are interested in this area, good job and services to them. In the future, which analysis display by graph.

REFERENCES

[1]G. Siemens and P. Long, "Penetrating the fog: Analytics in learning and education," *Educause Review*, vol. 46, no. 5, pp. 30–32, 2011.
[2]M. Rost, L. Barkhuus, H. Cramer, and B. Brown, "Representation and communication: challenges in interpreting large website datasets," in *Proceedings of the 2013 conference on Computer supported cooperative work*, 2013, pp. 357–362.
[3]M. Clark, S. Sheppard, C. Atman, L. Fleming, R. Miller, R. Stevens, R. Streveler, and K. Smith, "Academic pathways study: Processes and realities," in *Proceedings of the American Society for Engineering Education Annual Conference and Exposition* 2008.
[4]C. J. Atman, S. D. Sheppard, J. Turns, R. S. Adams, L. Fleming, R. Stevens, R. A. Streveler, K. Smith, R. Miller,
L. Leifer, K. Yasuhara, and D. Lund, "Enabling engineering student success: The final report for the Center for the Advancement of Engineering Education," Morgan & Claypool Publishers, Center for the Advancement of Engineering Education, 2010.
[5]R. Ferguson, "The state of learning analytics in 2012: A review and future challenges," *Knowledge Media Institute*, Bing Liu, Wynne [6]Hsu and Shu Chen, "Using General Impressions to Analyze Discovered Classification Rules", Proceedings of the Third International Conference on Knowledge Discovery and Data Mining (KDD-97), pp. 31-36, August 14-17, 1997, Newport Beach, California, USA.
[7] K. Wang and B. Liu, "Concurrent Discretization of Multiple Attributes", PRICAI 98, August 1998, Singapore
[8]Khaled Alsabti, Sanjay Ranka, and Vineet Singh, "CLOUDS: A Decision Tree Classifier for Large Datasets", Proc. of International Conference on Knowledge Discovery and Data Mining (KDD-98), August 1998, New York City.
[9]Bing Liu, Wynne Hsu, Yiming Ma, "Integrating Classification and Association Rule Mining." Proc. of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD-98), New York, USA, 1998.
[10]Mahesh V. Joshi, George Karypis and Vipin Kumar, "ScalParC: A New Scalable and Efficient Parallel Classification Algorithm for Mining Large Datasets", Proc. of 12th International Parallel Processing Symposium (IPPS/SPDP), April 1998, Orlando, USA.
[11]R. Agrawal, H. Mannila, R. Srikant, H. Toivonen and A. I. Verkamo, "Fast Discovery of Association Rules", Advances in Knowledge Discovery and Data Mining, Chapter 12, AAAI/MIT Press, 1995.
[12]R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules", Proc. of the 20th Int'l Conference on Very Large Databases, Santiago,
[13]R. Srikant, Q. Vu, R. Agrawal, "Mining Association Rules with Item Constraints", Proc. of the 3rd Int'l Conference on Knowledge Discovery in Databases and Data Mining, Newport Beach, California, August 1997.