International Journal of Information Technology and Management

ISSN online: 1741-5179 - ISSN print: 1461-4111
https://www.inderscience.com/ijitm

Scrutinising medical practitioners' Twitter feeds: an analysis

Arushi Jain, Vishal Bhatnagar, Nilanjan Dey, Amira S. Ashour, Fuqian Shi

DOI: 10.1504/IJITM.2023.10055155

Article History:

Received:	28 October 2017
Last revised:	08 July 2018
Accepted:	05 May 2019
Published online:	05 April 2023

Scrutinising medical practitioners' Twitter feeds: an analysis

Arushi Jain*

Indian Institute of Technology,
Dhanbad, Jharkhand, India
Email: arushijain1391@gmail.com
*Corresponding author

Vishal Bhatnagar

Ambedkar Institute of Advanced Communication Technologies and Research,
Geeta Colony, Delhi, India
Email: vishalbhatnagar@yahoo.com

Nilanjan Dey

Techno India College of Technology,
Kolkata, West Bengal, India
Email: neelanjan.dey@gmail.com

Amira S. Ashour

Tanta University,
Gharbia Governorate, Egypt
Email: amirasashour@yahoo.com

Fuqian Shi

Wenzhou Medical University,
Wenzhou City, 325035, China
Email: shifuqian@gmail.com

Abstract: Mining of social media data has found widespread applications in recent times. Twitter feeds and Facebook posts are being used to device product marketing strategies, sentiment analysis, financial predictions and forebode alarming situations. Twitter feeds analysis can be applied for analysing the behaviour and experiences of medical practitioners. Doctors' informal conversations on Twitter can provide deep insights about their work experiences, their concerns about the profession, their feelings – pathos or excitement they feel – and the affecting conditions. In the present work, Twitter feed of doctors along with the Twitter hashtags are used to collect data from tweets with hashtags such as #DoctorProblems. Afterward, data analysis was performed to determine the major problems faced by the members of the medical fraternity. These problems were categorised into five main categories.

The multi-label naïve-Bayes classification algorithm and MapReduce paradigm was then implemented. The experimental results were evaluated and compared. It was found that the naïve Bayes multi-label classifier is superior to the MapReduce method for Twitter mining in terms of the time complexity. The obtained results could help employers provide better working environments to their doctors and make better informed decisions in order to help the concerned parties.

Keywords: big data; Hadoop distributed file system; HDFS; MapReduce; naïve Bayes multi-label classifier; Tweets; evaluation-based measure; label-based measure.

Reference to this paper should be made as follows: Jain, A., Bhatnagar, V., Dey, N., Ashour, A.S. and Shi, F. (2023) 'Scrutinising medical practitioners' Twitter feeds: an analysis', *Int. J. Information Technology and Management*, Vol. 22, Nos. 1/2, pp.127–139.

Biographical notes: Arushi Jain is doing her PhD in Computer and Engineering from IIT Dhanbad, India. Her research works include big data, data mining, data science and machine learning.

Vishal Bhatnagar holds BTech, MTech and PhD in the engineering field. He has more than 20 years of teaching experience in various technical institutions. He is currently working as Professor in Computer Science in Engineering Department at Ambedkar Institute of Advanced Communication Technologies and Research (Government of Delhi), GGSIPU, Delhi, India. His research interests include database, advance database, data warehouse, data-mining, social network analysis and big data analytics. He has to his credit more than 100 research papers in various international/national journals and conferences.

Nilanjan Dey is an Assistant Professor in Department of Information Technology at Techno India College of Technology, Kolkata, India. He is a Visiting Fellow of the University of Reading, UK. He is the Editor-in-Chief of *International Journal of Ambient Computing and Intelligence*, IGI Global, Associated Editor of *IEEE Access* and *International Journal of Information Technology*, Springer. He is the series Co-Editor of *Springer Tracts in Nature-Inspired Computing*, *Springer Nature*, series Co-Editor of *Advances in Ubiquitous Sensing Applications for Healthcare*, Elsevier, series Editor of *Computational Intelligence in Engineering Problem Solving* and *Intelligent Signal Processing and Data Analysis*, CRC.

Amira S. Ashour is the Vice Chair of CS Department, CIT College, Taif University, KSA for five years. She is in the Electronics and Electrical Communications Engineering, Faculty of Engineering, Tanta University, Egypt. She received her PhD in the Smart Antenna (2005) from the Electronics and Electrical Communications Engineering, Tanta University, Egypt. Her research interests include: image processing, medical imaging, machine learning, biomedical systems, pattern recognition, signal/image/video processing, image analysis, computer vision, and optimisation.

Fuqian Shi is an Associate Professor at College of Information and Engineering, Wenzhou Medical University. His research interests include fuzzy inference system, artificial neuro networks, and biomechanical engineering He also serve as Associate Editor of *International Journal of Ambient Computing and Intelligence (IJACI)*, *International Journal of Rough Sets and Data Analysis (IJRSDA)*, and special issue editor of fuzzy engineering and intelligent transportation in *INFORMATION: An International Interdisciplinary Journal*. He published over 40 journal papers and conference proceedings.

1 Introduction

In recent era, social media has become a computer-intermediated tool that facilitates interaction and ideas among different individuals. The data thus collected includes feeling and emotions of the different people. This can be used to gain knowledge and other perspective of human feelings. Twitter mining have become prominent among researchers for discovering what people are talking about. The tone of the hashtag indicates whether people have positive, negative or neutral perspective regarding the trending topic. However, analysing the social media data has its several challenges due to the volume, velocity, value, veracity, the data variety, which changed rapidly and referred to as the 5 Vs. Such massive data is considered to be big data due to its humongous amount (Zikopoulos and Eaton, 2011).

One of the most efficient open source software frame work for large scale processing and storage of big data is the Apache Hadoop platform (González-Vélez and Kontagora, 2011; Jarrah et al., 2015) which is a cluster of commodity hardware. The Apache Hadoop provides scalable and pragmatic infrastructure. Major components include Hadoop distributed file system (HDFS) and MapReduce paradigm. MapReduce performs parallel task and in a synchronised fashion. It runs the user defined processing logic at the machines where the data lives rather than bringing all the data to the same place using networks. This guarantee superior performance of the Hadoop compared to other distributed file systems.

Through the internet, a growth of the social media platforms provides opportunities to the medical sector (Aye et al., 2015). Social media are used by patients, health care institutions, physicians, and public health, which result in significant impacts for the care of the individual patient. Additionally, the interaction through Twitter, Facebook and such social media sources allow the development of online groups to improve the disease knowledge and outcomes (Terry, 2009). Furthermore, easy medical information sharing is achieved through the social networking sites between doctor-patient related to diagnosis, treatment, and testing. Consequently, this work is interested with the Twitter feed of doctors based on Twitter hashtags collecting data from tweets with hashtags such as #DoctorProblems. Afterward, the collected data was analysed to extract the major problems faced by the members of the medical fraternity.

2 Related research work and motivation

Social media has vastly transformed the communication/interaction between people and the searching (Banu and Tripathi, 2016). It encompasses a wide variety of websites including social networking communities, Facebook, Twitter,YouTube blogs, online web journals, and educational sites. Therefore, researchers become interested with developing new data analysis methods to handle the growing demands and the increased information capacity.

Utilising structured and unstructured data from the hospital, Chen et al. (2017) recommended a new multimodal disease risk prediction (CNN-MDRP) algorithm-based on convolutional neural network. In relation to numerous distinct prediction algorithms, the forecasting precision of algorithm attains at 94.8% with a convergence speed that is quicker than that of the CNN-based unimodal disease algorithm.

A particle swarm optimisation-based approach by Chatterjee et al. (2016a, 2016b) to train the NN (NN-PSO), this classifier was able to predict structure failure of multistoried building. Chatterjee et al. (2015) proposed a model to forecast the training quality and other relatable mistakes in order to avoid any severe effect on athlete's performance.

A novel application of particle swarm optimisation (PSO) was implemented by Chatterjee et al. (2017) in order to separate the patients who are having dengue fever and recovering fast than those who do not have dengue fever.

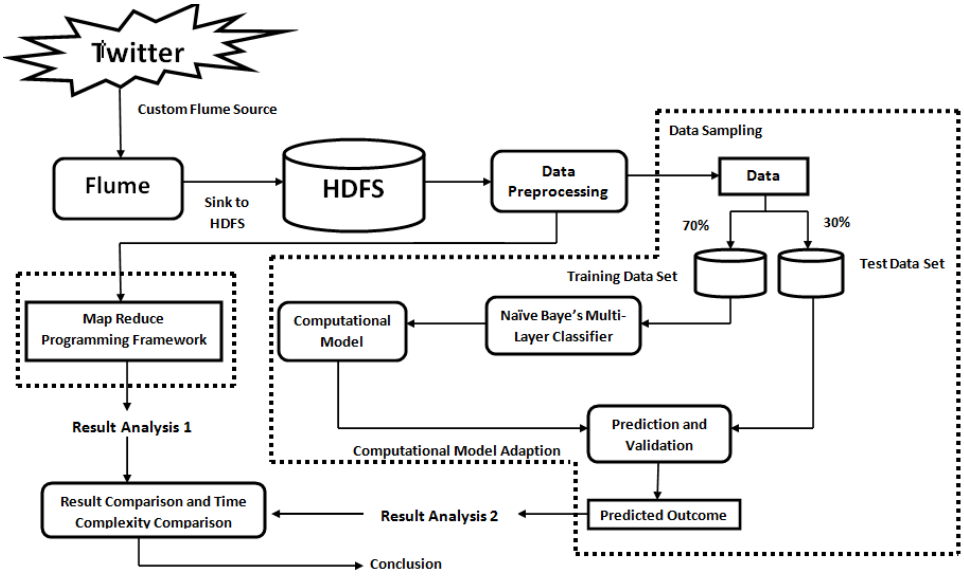
Hore et al. (2017) proposed a novel algorithm by combining neural network along with genetic algorithm to solve the problems of recognition of ISL gestures.

3 Methodology

From the preceding literatures (Carr and Hayes, 2015), it is established that several studies were concerned with big data in several applications for relations and information extraction. Consequently, this work proposed an approach for Twitter feed of doctors based on Twitter hashtags such as #DoctorProblems for data collecting from tweets.

These data are further analysed to identify the main problems that face the medical community members. The analysed data is categorised into five classes. The methodology is illustrated in Figure 1.

Figure 1 Proposed system methodology



As illustrated in Figure 1, the Twitter feed of doctors based on the Twitter hashtags were loaded in the first step into the HDFS through flume. Afterward, a preprocessing for the data before training the classifier is performed by:

1 all the hashtags were removed, only keeping the text

2 emoticons, punctuations, http links and other non-letter symbols were removed

3 replace the two similar letters for example 'hoooot' was replaced by 'hot'.

This was implemented using MapRedude program, using bifferedfile reader function performing cleaning operation on each line.

Data analysis was performed in the second step to find the major problems that the medical fraternity members can face. The notable topics were bad working environment, social isolation, insensitivity in patient doctor interaction, gloomy eloquence, sleep deprivation on shift work, and others. The most relevant 15 words from the each group are depicted in Table 1.

Table 1 The five major groups and the relevant words in each prominent theme

Group	*Significant words*
Unforgiving work environment	Fast paced, ultra-demanding, hard work, inconsistent policies, favouritism, badly treated, dysfunctional relationship, communication issue, long term grudges, back-biting, criticising, insufficient funding, egocentrism, expectations, too much
Social Isolation	Miss, tense, troubled, work hard, excel, workload, increase, drastic, responsibility, party, doubts, struggling, constant, shortfalls, pressure
Insensitive in patient doctor interaction	Coldness, excessive emotionality, unwillingness, lack of empathy, aggression, intimidation, lack of morality, bullying, hypocritical, pervasive, discourtesy, abruptness, disrespect, unrealistic, intolerance
Gloomy Eloquence	Sad, depressed, weakness, evaluation, disappointment, blue, down, hopeless, black, dark, bad, miserable, dejected, discouraged, fed up
Sleep Deprivation on shift work	Weekend, limit, wake, awake, time, available, call, maximum, training, continue, week, long delay, fatigue, malpractice, schedule

Further, the multi-label naïve-Bayes classification algorithm and MapReduce paradigm was implemented based on these categories. The dataset used in the current work consists of 1.5 lakh tweets. Since, the classification process has mainly training and testing phases. Thus, 70% of the 1.5 lakh tweets were used for training, and rest 30% were used for testing. The obtained results could help employers provide better working environments to their doctors and make better informed decisions in order to assist the concerned parties.

3.1 Mapreduce programming framework

The Twitter data was processed using MapReduce programming framework. In the current work MapReduce programming paradigm was used to calculate the percentage of tweet in each prominent theme.

Mapper Algorithm

Assume N is the number of tweets

String x [6][], $x[i][]$ = array of strings in i^{th} *group*

$y[6]$ = total number of tweets in group $i = y[i]$

$z[6]$ = number of tweets in group $i = z[i]$

for $i = 1$ to N

for $p = 1$ to 6

if(tweet[i] contains any word in $x[p][]$

$y[p]$++

End for

End for

Reducer Algorithm

for i = 1 to 6

percentage of tweets belonging to category i

$$z[i] = \left(\frac{y[i]}{N} * 100 \right)\%$$

End for

The previous procedures are used and tested by the collected dataset.

3.2 Naïve Bayes multi-label classifier

The naïve Bayes multi-label classifier is used to process the Twitter data. In the current work, a naïve Bayes mutli-label classifier is implemented to classify the tweets from the dataset based on the six groups (Chen et al., 2014). Assume p number of words in the collection of training document $D = (\beta_1, \beta_2, \beta_3, \ldots \beta_P)$, total number Q groups $G = (g_1, g_2, g_3, \ldots, g_Q)$. If a word β_p appears in group g for $r\beta_{pg}$ times, then the probability of this word falling in a specific group g (Chen et al., 2014) is:

$$p(\beta_p | g) = \frac{r_{\beta_p} g}{\sum_{p=1}^{p} r_{\beta_p} g} \tag{1}$$

However, the probability of this word in the different groups other than this group (Chen et al., 2014) is:

$$p(\beta_p | g) = \frac{r_{\beta_p} g'}{\sum_{p=1}^{p} r_{\beta_p} g'} \tag{2}$$

Group g probability is given by (Chen et al., 2014):

$$p(g) = \frac{G}{R} \tag{3}$$

In addition, the probability of other group g' is calculated by (Chen et al., 2014):

$$p(g') = \frac{R - G}{R} \tag{4}$$

For a document d_i in the testing set, assume M words of $\beta_{di} = \{\beta_{d1}, \beta_{d2}, \beta_{d3} \ldots \beta_{iM}\}$, where β_{di} is a subset of β. According to Bayes' theorem, the probability that this belongs to group g is calculated by (Chen et al., 2014):

$$p(g | d_i) = \frac{p(d_i | g) \cdot p(g)}{p(d_i)} \alpha \prod_{m=1}^{M} p(\beta_{im} | g) \cdot p(g) \tag{5}$$

The probability that d_i is (Chen et al., 2014):

$$p(g'|d_i) = \frac{p(d_i|g') \cdot p(g')}{p(d_i)} \alpha \prod_{m=1}^{M} p(\beta_{im}|g') \cdot p(g') \quad (6)$$

where $p(g|d_i) + p(g'|d_i) = 1$.

3.2.1 Comparative analysis with other algorithms

Maximum entropy

Assume p number of words in the collection of training document $D = (\beta_1, \beta_2, \beta_3, \ldots, \beta_P)$, total number Q groups $G = (g_1, g_2, g_3, \ldots, g_Q)$.

Probability distribution has an exponential form which can be represented as,

$$P(G|T) = \frac{1}{Z(T)} \exp\left(\sum_i \lambda_{(i,g)} F_{(i,g)}(t, g)\right)$$

t single tweet

λ vector weight

$Z(T)$ normalisation function = $(\sum_i \lambda_{(i,g)} F_{(i,g)})$ $(t, g))$

$F_{i,g}$ indicator function for group g which is defined as.

$$F_{i,g}(t, g) = \begin{cases} 1, & \text{if } g' = G \text{ and } t \text{ contains } \beta p \\ 0, & \text{otherwise} \end{cases}$$

The indicator function will return 1 when the group of a particular document is G and the document contain the word β_P.

Pseudo Code:

1 Initialise all weight parameter $\lambda_i = 0$

2 Repeat till convergence

3 Calculate the probability distribution

4 For each parameter λ_i calculate $\Delta\lambda_i$, which will satisfy the following equation:

$\sum_g P(g, t) \cdot f_i(t, g) \cdot \exp(\Delta\lambda_i f(t, g)) = \sum_t f_i(t, g)$

5 Update the vector weight value:

6 $\lambda_i = \lambda_i + \Delta\lambda_i$

Support vector machine

1 Assume p number of words in the collection of training document $D = (\beta_1, \beta_2, \beta_3, \ldots, \beta_P)$, total number Q groups $G = (g_1, g_2, g_3, \ldots, g_Q)$.

2 Notation to formally define a hyper plane:

$$h(X) = \lambda^x \phi(x) + \theta$$

X feature vector

x training examples closest to the hyperplane

λ vector weights

ϕ nonlinear mapping function

θ bias vector.

3 We chosen a hyper plane in such a way that:

$$\lambda^x \varphi(x) + \theta = 1.$$

4 The distance between the plane (λ, θ):

$$\text{Distance} = \frac{\left(\left|\lambda^x \phi(x) + \theta\right|\right)}{\|\lambda\|}$$

5 In canonical hyperplane, the numerator is equal to 1 and distance to the support vector is given by:

$$\text{Distance}_{\text{support vector}} = \frac{\left(\left|\lambda^x \phi(x) + \theta\right|\right)}{\|\lambda\|} = \frac{1}{\|\lambda\|}.$$

6 Margin can be presented as $\frac{2}{\|2\|}$.

7 Problem of maximising M is equivalent to minimising a function $L(\lambda)$ subject to constraints

$$\min_{(\lambda,\theta)} L(\lambda) = \frac{1}{2}\|\lambda\|^2 \text{ subject to } y_i\left(\lambda^x \phi(x) + \theta\right) >= 1 \,\forall_i,$$

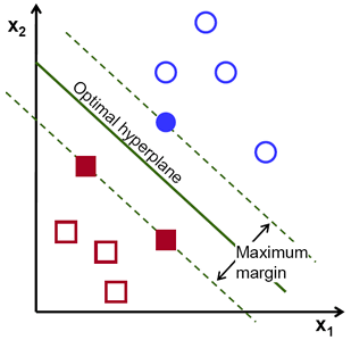
y_i represents labels of the training examples.

Note: in the above algorithm we used kernel function for classification, which can be represented as:

$$\text{function } f = \text{kernel}\left(x_1, x_2\right)$$

$$f = \exp\left\{\frac{-\|x_1 - x_2\|^2}{2\lambda^2}\right\}$$

Figure 2 Optimal Hyperplane using the SVM algorithm (see online version for colours)

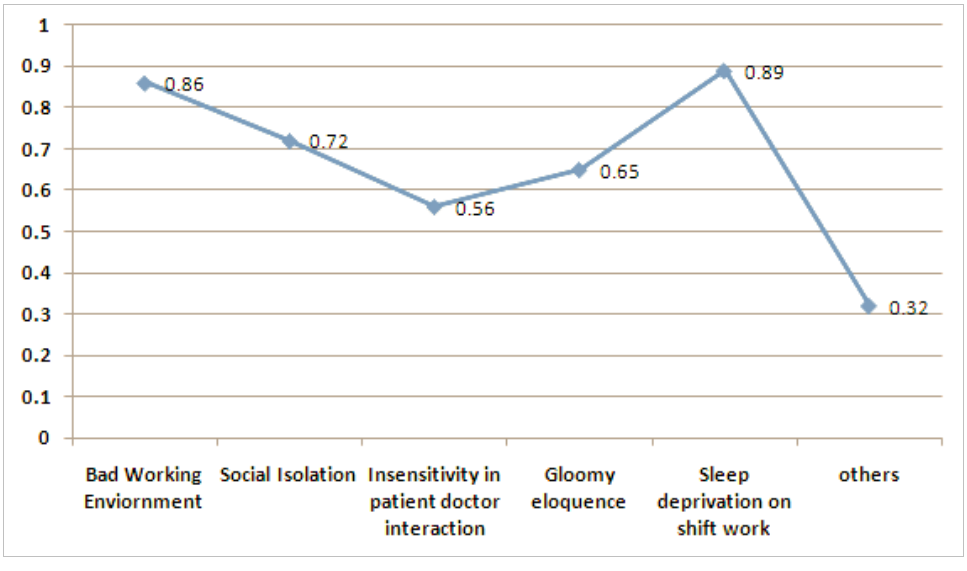


4 Results and implications

4.1 MapReduce results

Figure 3 illustrated the map reduce results for the different groups mentioned in Table 1.

Figure 3 MapReduce results (see online version for colours)



In Figure 3, x-axis represents the different groups while y-axis represents the percentage of tweets belong to each groups. From the MapReduce program, 0.89% of the tweets belongs to the group sleep deprivation during various shifts, followed by 0.86% of the tweets belongs to bad working environment, while least number of tweets for the insensitivity patient doctor interaction. These results can assist the employers to provide better working environments to their doctors and make better informed decisions that support he concerned parties.

4.2 Evaluation measures for the naïve Bayes multi-label classifier

Table 2 represents the six evaluation measures of naïve Bayes multi label for different threshold from 0 to 1.

Table 3 represents the best performance comes when the threshold value is 0.5 compared to other threshold values.

In Table 4 for each group precision is calculated.

In Figure 4, x-axis depicts the different groups while y-axis depicts the specific tweets percentage to the total tweets belonged to the different groups. Figure 4 establishes that large number of doctors has an issue regarding sleep deprivation during various shift works followed by bad working environment. The same results are comparable to the MapReduce results.

Table 2 Six evaluation measures with naïve Bayes multi label with segment 0 to 1

T	*Accuracy (a)*	*Precision (pr)*	*Recall (re)*	*Harmonic average (F_1)*	F_{micro}	F_{macro}
0	0.1522	0.1411	0.9142	0.0407	0.0402	0.0308
0.1	0.5210	0.5866	0.6213	0.6034	0.5643	0.4819
0.2	0.6543	0.6873	0.7166	0.7016	0.6900	0.5880
0.3	0.6829	0.6923	0.7523	0.7210	0.7084	0.6932
0.4	0.7182	0.7123	0.7418	0.7267	0.7012	0.6001
0.5	0.7234	0.7777	0.8234	0.7998	0.7824	0.6821
0.6	0.7268	0.7331	0.7444	0.7387	0.7342	0.6213
0.7	0.7312	0.7388	0.7481	0.7434	0.7221	0.6222
0.8	0.7324	0.7456	0.7569	0.7512	0.7361	0.6289
0.9	0.7362	0.7338	0.7413	0.7375	0.7001	0.6221
1	0.7431	0.7551	0.7682	0.7615	0.7448	0.6411
Random	0.0568	0.0570	0.0575	0.0572	0.0551	0.0432

Table 3 Label-based accuracy and F_1

Group	*Label accuracy*	*Label F_1*	*Rand accuracy*	*Rand F_1*
Unforgiving work environment	0.9012	0.5478	0.0662	0.0023
Social isolation	0.9432	0.5932	0.0716	0.0045
Insensitive in patient-doctor interaction	0.9654	0.6372	0.0749	0.0052
Gloomy eloquence	0.9739	0.6751	0.0756	0.0061
Sleep deprivation on shift work	0.9843	0.7529	0.0803	0.0074
Others	0.8264	0.8832	0.0487	0.0561

Figure 4 Output of naive Bayes multi label classifier (see online version for colours)

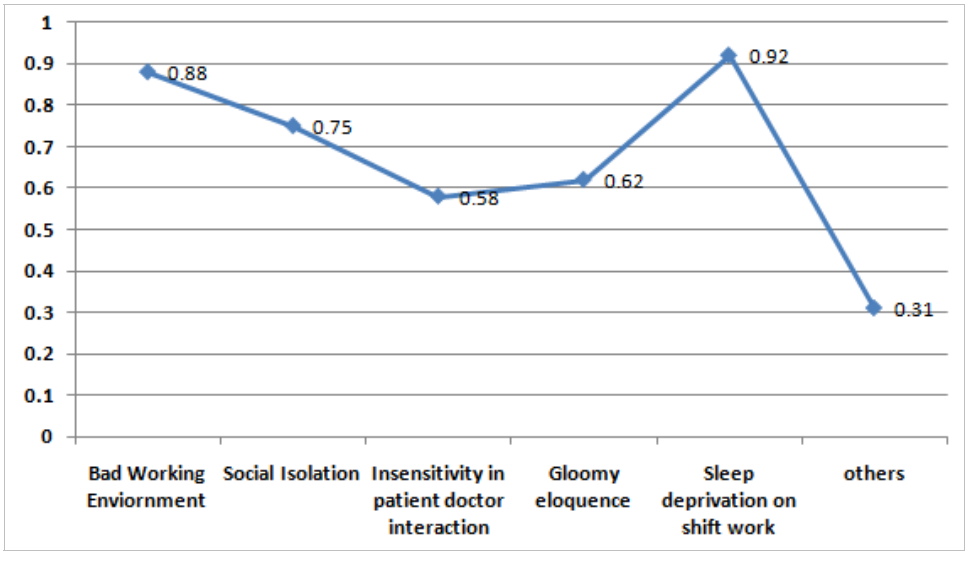


Table 4 Example-based measurement and the precision for each group

Ex. accuracy(a)	*Ex. precision (pr)*	*Ex. recall (re)*	*Ex. F_1*	*F_{micro}*	*F_{macro}*
0.7234	0.7777	0.8234	0.7998	0.7824	0.6821

4.2.1 Performance measure

A matrix can be created between true positive, true negative, false negative and false positive.

Table 5 Contingency table per group

	True g	*True not g*
Predicted *g*	True positive(t_p)	False positive (f_p)
False Prediction (not *g*)	False negative (f_n)	True negative(t_n)

where t_p is true positive classification, f_p is false positive classification, f_n is false negative classification, and t_n is true negative classification.

$$a = \frac{t_p + t_n}{t_p + t_n + f_p + f_n}$$

$$pr = \frac{t_p}{t_p + f_p}$$

$$re = \frac{t_p}{t_p + f_n}$$

$$F_1 = \frac{2.t}{2.t_p + f_p + f_n}$$

Table 6 Precision and recall for the techniques

Group	*Precision*			*Recall*		
	Maximum entropy	*SVM*	*Naïve Bayes multi-label classifier*	*Maximum entropy*	*SVM*	*Naïve Bayes*
Unforgiving work environment	0.78	0.82	0.88	0.73	0.80	0.88
Social isolation	0.62	0.66	0.72	0.59	0.64	0.74
Insensitive in patient doctor interaction	0.41	0.49	0.52	0.42	0.48	0.50
Gloomy eloquence	0.57	0.49	0.61	0.53	0.45	0.62
Sleep deprivation on shift work	0.72	0.73	0.92	0.68	0.70	0.90
Others	0.22	0.21	0.38	0.21	0.19	0.35

5 Time complexity comparison

Table 7 reports a detailed comparative study of various used techniques. In first technique, a MapReduce program was implemented, while in the second technique naïve Bayes multi-label classifier was implemented long with support vector machine (SVM) and maximum entropy.

Table 7 Results comparison for the techniques

Group	*Map reduce results*	*Naïve Bayes multi-label classifier results*	*Maximum entropy*	*SVM*
Unforgiving work environment	0.86	0.88	0.78	0.82
Social isolation	0.72	0.75	0.62	0.66
Insensitive in patient doctor interaction	0.56	0.52	0.41	0.49
Gloomy eloquence	0.65	0.61	0.57	0.49
Sleep deprivation on shift work	0.89	0.92	0.72	0.73
Others	0.32	0.38	0.22	0.21

Table 4 deploys that percentage of tweets belonging to different groups are comparable to each other in both the cases. Furthermore, the time complexity of the MapReduce programming framework is O (n*c*w), while the time complexity of naïve Bayes multi-label classifier is linearly proportional to the time to read all the data, thus it is of order $O(n)$. Where, n is the number of tweets, c is the number of categories, and w is the average number of words present in each category.

From the exceeding results, it is established that the naïve Bayes multi-label classifier is the most efficient method for the Twitter mining as the time complexity is linearly proportional to the time to read all data. Maximum tweets belong to the group sleep deprivation during various shifts, followed by the tweets belonging to bad working environment. The results can support senior authorities to provide better working environments to their doctors and better cope mechanisms can be developed to help in alleviating their major issues.

Another method is suggested to analyse the use of emoticons, which can be used to gauge the sentiment of the author of the tweet. However, for the sake of the current work, knowing only the sentiment of the doctor who is tweeting is not sufficient as it does not provide much actionable knowledge of on mechanisms to improve the quality of environment for doctors so as to solve their problems. The aim is to understand the day to day experiences of the medical practitioners, what leads to their problems and eventually provide better remedies.

6 Conclusions

Doctors ‘informal conversations on Twitter can provide deep insights into their work experiences, their concerns about the profession, their feelings- pathos or excitement they feel and the conditions that affect. The current work deployed that the naïve Bayes multi-label classifier is the most efficient method for the Twitter mining as time complexity is linearly proportional to the time to read all data. The experimental results

depicted that maximum tweets belong to the group sleep deprivation during various shifts, followed by the tweets belonging to bad working environment. The results could help senior authorities to provide better working environments to their doctors and better cope mechanisms can be developed thus help in alleviating their major issues.

References

Al-Jarrah, O.Y., Yoo, P.D., Muhaidat, S., Karagiannidis, G.K. and Taha, K. (2015) 'Efficient machine learning for big data: a review', *Big Data Research*, Vol. 2, No. 3, pp.87–93.

Aye, K.N. and Thein, T. (2015) 'A platform for big data analytics on distributed scale-out storage system', *International Journal of Big Data Intelligence*, Vol. 2, No. 2, pp.127–141.

Bhanu, S.K. and Tripathy, B.K. (2016) 'Rough set based similarity measures for data analytics in spatial epidemiology', *International Journal of Rough Seta and Data Analysis*, Vol. 3, No. 1, pp.114–123.

Carr, C.T. and Hayes, R.A. (2015) 'Social media: defining, developing, and divining', *Atlantic Journal of Communication*, Vol. 23, No. 1, pp.46–65.

Chatterjee, S., Ghosh, S., Dawn, S., Hore, S. and Dey, N. (2015) 'A quality prediction method for weight lifting activity', *Michael Faraday IET International Summit*.

Chatterjee, S., Ghosh, S., Dawn, S., Hore, S. and Dey, N. (2016a) 'Forest type classification: a hybrid NN-GA model based approach', *Michael Faraday IET International Summit*.

Chatterjee, S., Hore, S., Dey, N., Chakraborty, S. and Ashour, A.S. (2017) 'Dengue fever classification using gene expression data: a pso based artificial neural network approach', in Satapathy, S., Bhateja, V., Udgata, S. and Pattnaik, P. (Eds.): *Proceedings of the 5th International Conference on Frontiers in Intelligent Computing: Theory and Applications. Advances in Intelligent Systems and Computing*, Vol. 516, Springer, Singapore.

Chatterjee, S., Sarkar S., Hore, S., Dey, N., Ashour, A. and Balas, V.E. (2016b) 'Particle swarm optimization trained neural network for structural failure prediction of multistoried RC buildings', *Neural Computing and Applications*, Vol. 28, No. 8, pp.2005–2016.

Chen, M., Hao, Y., Hwang, K., Wang, L. and Wang, L. (2017) 'Disease prediction by machine learning over big data from healthcare communities', *IEEE Access*, Vol. 5, No. 5, pp.8869–8879.

Chen, X., Vorvoreanu, M. and Madhavan, K. (2014) 'Mining social media data for understanding student's learning experience', *IEEE Transaction on Learning Technologies*, Vol. 7, No. 3, pp.246–259.

Deepak, D. and John, S.J. (2016) 'Information systems on hesitant fuzzy sets', *International Journal of Rough Seta and Data Analysis*, Vol. 3, No. 2, pp.55–70.

Fedoryszak, M., Tkaczyk, D. and Bolikowski, L. (2013) 'Large scale citation matching using apache Hadoop', *Research and Advanced Technology for Digital Libraries Lecture Notes in Computer Science*, Vol. 8092, pp.362–365, Springer, Egypt.

González-Vélez, H. and Kontagora, M. (2011) 'Performance evaluation of MapReduce using full virtualisation on a departmental cloud', *International Journal of Applied Mathematics and Computer Science*, Vol. 21, No. 2, pp.275–284.

Hore, S. et al. (2017) 'Indian sign language recognition using optimized neural networks', in Balas, V., Jain, L. and Zhao, X. (Eds.): *Information Technology and Intelligent Transportation Systems. Advances in Intelligent Systems and Computing*, Vol. 455, No. 1, Springer.

Terry, M. (2009) 'Twittering healthcare: social media and medicine', *Telemedicine and e-Health*, Vol. 15, No. 6, pp.507–510.

Zikopoulos, P. and Eaton, C. (2011) *Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data*, McGraw-Hill Osborne Media, Canda.