Automated Evaluation of Students' Feedbacks using Text Mining Methods

Sartaj Ahmad, Ashutosh Gupta, Neeraj Kumar Gupta

Abstract: In this era of competition there is a culture of online reviews or feedbacks. These feedbacks may be about any product or service. However, major issues are their unstructured textual form and big number. It means every user gives feedback in own style. Study and analyzing of such unorganized big number of feedbacks that are growing every year becomes herculean task. This paper describes about mining of structured data (table) and unstructured data (text) both. An application from academic environment for structured and unstructured form of data is considered and discussed to enhance understanding and easiness of researcher. Stanford Parser plays a very useful role to understand the semantic of a sentence. It gives a base that how to separate data from the wellspring of information accessible in the literary structure like web based life, tweets, news, books and so on. It is also helpful to judge a teaching learning process in terms of teacher's performance and subject's weakness if any. This paper has five sections first about introduction, second about literature of text mining and its techniques, third about proposed work and result, fourth about future perspectives and finally fifth as a conclusion.

Keywords:* Content Mining, *Data Mining, Information Retrieval, Knowledge Extraction, Text Mining, Unstructured data mining.

I. INTRODUCTION

In academic many attributes are required to be a good teacher. Few are sound knowledge in the subject, good personality, hardworking and classroom skills etc. Still there may be some lacking in teaching methodology. This is the responsibility of the teacher itself that how to improve his teaching. It is also sure that every teacher wants to become subject wise popular among students. Therefore students' feedbacks play great role to improve teaching learning process. Generally, these feedbacks are collected at the end or during the semester in the form of two formats one is structured in nature means parameters (attributes) are given in the form of table and values are asked from the student. Second format is unstructured in nature means there is no field decided. In this student can give feedback in free style. Structured data is easy to handle and find knowledge. But major problem is with unstructured form and big number of these feedbacks that are increasing every year as number of students is increasing. Any feedback can have one line or more than one line. It becomes difficult to read and analyze all feedbacks manually. Therefore there should be an automatic technique to analyze the performance of teachers and students in terms of these feedbacks. In this paper one line feedbacks are considered as input to retrieve information. This paper presents study about digging of information and finding problems from such increasing rate of data due to maturity of academic organization. In this paper initially an art of state is provided on text mining and its techniques. Examples from academic environment are discussed as application of academic process. At last conclusion is presented.

II. LITERATURE REVIEW

Text can be available in any form of data like structured data (tables, list, tree), semi structured (HTML documents) and unstructured data (free text). But a recent study says that 70 to 80% information of any organization is kept in the text form. In this manner, it requires content mining which is substantially more intricate assignment than customary information mining. It extracts information from the text which is unstructured in nature [1]. Therefore it is multidisciplinary field that includes information retrieval, text analysis, text classification, representation, machine learning, database technology and information mining. It deals to find interesting pattern from textual documents while data mining deals with structured data. In other words text mining can be considered as an application of data mining [2]. However, there is problem with such text mining reason data does not pursue any data model. It is difficult how to extract concealed information from such sort of data. . Programming modules called wrapper can be utilized to separate information from sites. Yet, issue is that wrapper depends on manual procedures and composing such wrapper is troublesome procedure. Similarly, Opinion extraction from online sources [3] like reviews with respect to sold items or any service is additionally an issue. Once opinions are extracted these help the concerned users in their decision making. In the next paragraph different techniques for the extraction of Meta data from unstructured data are discussed.

A. Techniques for Meta data extraction from unstructured data

Bayesian Networks: In [4] it is depicted to get data from unstructured, ungrammatical and confused information sources from the web. It is joined with domain based ontology, probabilistic techniques and machine learning.

Text Analytics: In [5] distinctive arrangement of AI strategies like semantic, statistical which model and organize the data substance of textual hotspots for business insight and research are covered. Its sub tasks incorporate the followings.

Revised Manuscript Received on

***** Correspondence Author

Sartaj Ahmad*, IT Department, KIET Group of Institutions affiliated to AKTU Lucknow, Ghaziabad, India. Email: sartaj.ahmad@kiet.edu

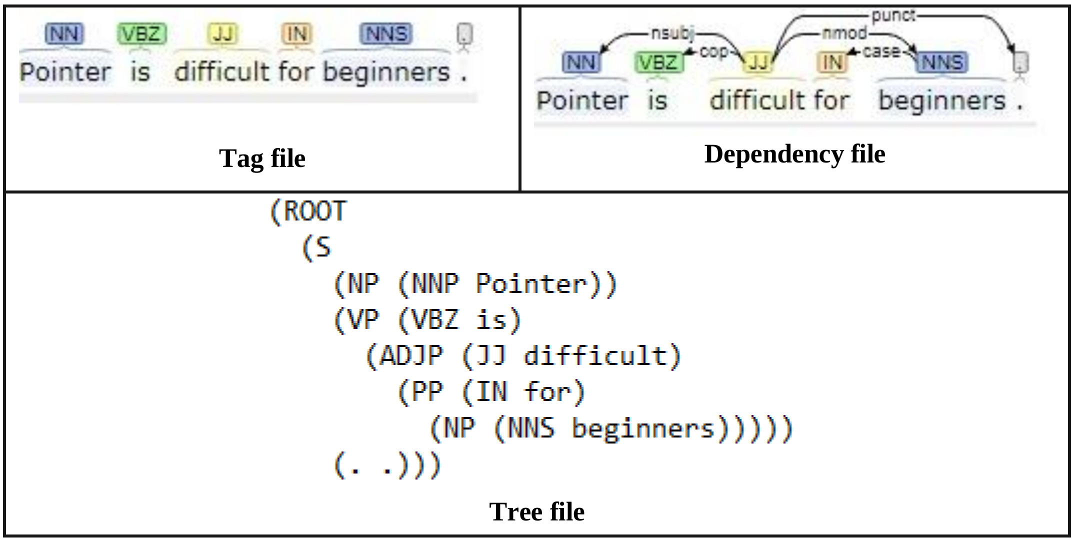
Ashutosh Gupta, School of Science in UPRTU, Allahabad, India, Email: ashutosh3333@gmail.com

Neeraj Kumar Gupta, EN Department**,** KIET Group of Institution affiliated to AKTU Lucknow, Ghaziabad, India, Email: neeraj.gupta@gmail.com

- Identification of collection of textual data from the Web.
- Named substance acknowledgment intends to distinguish named content highlights like individuals, associations, places, images, certain contractions, etc.
- Recognition of features of distinguishable objects called entities for example, Telephone numbers, email addresses, sex and so on.
- Identification of thing expressions (noun phrases), descriptive word (adjective) and verb modifier (adverb) that allude to a similar item
- Identification of hidden relationship in the text.

For this purpose there are numerous business, examination research and open source programming [6] choices. Few of them are MeshLabs, AeroText, Clarabridge, SAS, Sysomos etc.

Natural Language Processing (NLP): In [7], [8] Language can be understood by the human very easily because they have intelligence. But in case of computer this is difficult task. This ability can be produced in the computer through programs. These programs understand the use of NLP to locate some concealed examples from the textual content. They told how text can be processed for semantic analysis. Unstructured data are sent to our program which is taking help of Stanford Parser to break into tokens. After parsing some algorithms are deployed to consider tokens like (Noun (NN), Adjective (JJ), and Verb (VBZ)) for the handling further. Identified adjective's orientation is checked from the lists of words [9] for a noun. This process is explained as an application in the proposed works.


NN
VBZ
JJ
IN
NNS
Pointer is difficult for beginners .
nsubj
cop
punct
nmod
case
NN
VBZ
JJ
IN
NNS
Pointer is difficult for beginners .

Tag file

Dependency file

```
(ROOT
  (S
    (NP (NNP Pointer))
    (VP (VBZ is)
      (ADJP (JJ difficult)
        (PP (IN for)
          (NP (NNS beginners)))))
    (. .)))
```

Tree file

Fig. 2. Stanford Parser output

Support vector machine: It is a popular machine learning approach [10], [11] that can be used for classification and regression purposes both. In this approach best hyper plane is found which divides a data set into classes as shown below. It is used in different areas like detecting spam, sentiment analysis, image recognition etc.

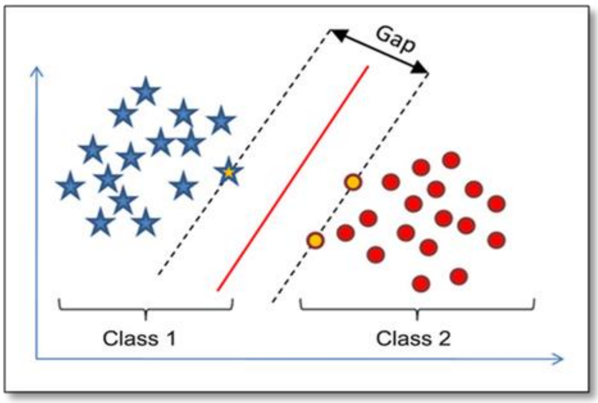


Fig. 1.Classification through Support vector machine

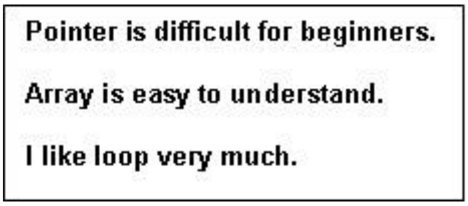
III. PROPOSED WORK

In this application subject's performance is measured in terms of two components one is based on opinions mining from the students' feedbacks (unstructured in nature) and second is based on classification of teachers' performances in terms of attributes values(structured data). These two components are presented as follows.

A. Opinions mining from the feedbacks of students (as an example of unstructured data mining)

Online reviews about subject (C programming) from under graduate students are collected. These reviews are written in free style means in unstructured mannered and send it back for further analysis. Manual Analysis of these reviews is very difficult due to unstructurdness. Therefore one approach is going to be discussed which analyze data and provide a summary. It helps teacher and student both. It enables teachers and students to find weakness and motivate themselves to increase their efforts to improve teaching learning process. Summary about the subject is obtained using the following steps.

Step 1: In this step we try to collect all reviews given by the different students. These reviews are broken into sentences, filtered and prepared for the parsing as shown in the following figure.



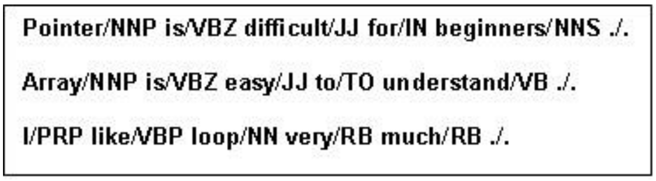
Pointer is difficult for beginners.

Array is easy to understand.

I like loop very much.

Fig. 3. Small Set of feedbacks in textual form

Step 2: In this step we pass filtered reviews to Stanford parser [12]. This parser gives three files as output one is Tag file second is dependency file and third is tree file shown in the figure 2 for one sentence. Many authors have used these files in their ways but out of these three files we are considering tag file for our processing. This tag file is having data as shown in the figure #4. We consider subjective sentences (sentences having noun and adjectives) from this files. This is done based on tag like NN/ NNP for noun while JJ for adjective.



Pointer/NNP is/VBZ difficult/JJ for/IN beginners/NNS ./.

Array/NNP is/VBZ easy/JJ to/TO understand/VB ./.

I/PRP like/VBP loop/NN very/RB much/RB ./.

Fig. 4.Tag file (Output of Stanford Parser)

Step 3: Now applying following algorithm on the above tag file to get topic wise opinion.

```
1. Read a file for each sentence and does
the step 2

2. Find words associated with NN/NNP

If word belongs to domain then
Find JJ and check its orientation using
lists of positive and negative words

   If orientation is positive then
   Increase the positive counter related
   to that word
    Otherwise
   Increase in the negative counter

3. Print list of topics along with
orientation
```

Using aforesaid algorithm a summary is produced. It shows feedback of the students about the subject's topics. For easiness C programming subject is discussed about 6 topics. These topics are considered depending on their popularity among the students. In this 500 reviews are considered for the experiment where 305 are extracted as positive, 150 as negative and rest are rejected because those were objective (not having noun and adjective phrases) reviews not subjective. Number of reviews for the experiment can be increased as per the requirement in future.

Table- I: Topic wise summary of feedbacks

S.No.	Topic	No. of +ve	No. of -ve
1	C_Looping	50	5
2	C_Array	105	10
3	C_Structure	25	10
4	C_Union	25	5
5	C_Pointer	50	100
6	C_ File Handling	50	20
	Total	305	150

Above table's summary can be represented graphically as follows

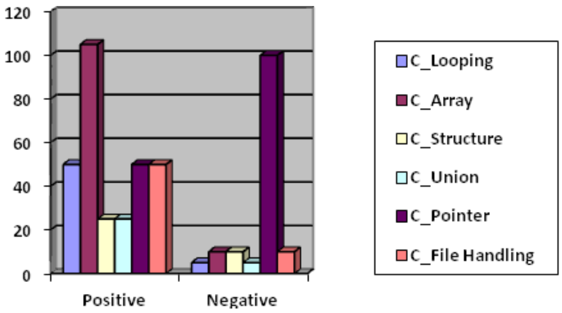


Fig. 5.Topic vs. opinions

Analysis of Summary

In the above figure positive side is overwhelming that shows positive reaction of all understudies about subject. Still there is a topic named pointer which is having progressively negative feelings. This may be due to the following reasons.

- Less comprehension of the understudies
- Not good instructing
- Less attention of the understudies

Therefore this topic requires more focus. One test of 10 marks is prepared to verify the above summary and subsequent outcome is obtained out of 500 students by the different teachers.

Table- II: Outcome of Test 1

Marks greater than equal to 8	Marks less than 8 and greater than equal to 5	Marks less than 5
125	175	200

Interpretation of result:

This outcome also supports reviews (feedbacks) output shown in table#1. Along these lines to conquer this issue some preventive moves are made.

Preventive action:

Organize remedial classes and provide more hands on practice. Teachers also rework on their teaching style and study material to get better students' understanding.

Corrective action:

Test number 2 is organized to confirm response of corrective/preventive deeds and acquire the outcome as shown in table III.

Table- III: Outcome of Test 2

Marks greater than equal to 8	Marks less than 8 and greater than equal to 5	Marks less than 5
150	300	50

Comparison between Test 1 & Test 2 outcomes:

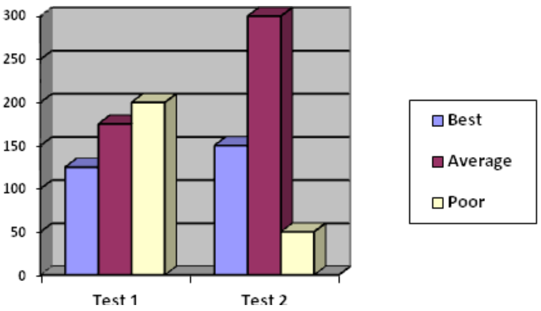


Fig. 6.Outcome of Test 1 vs. Outcome of Test 2

Analysis of figure 6:

It shows reduction in the number of weak understudies from test number 1 to test number 2 as far as C_Pointer topic is concerned.

Classification of teachers for the same subject based on their performance

In this performance of 5 teachers who were teaching same subjects are analyzed in terms of some attributes. These attributes are Teaching Experience, Teaching Methodology, Students Engagements in the class, Subject Knowledge, Presentation Skill with real life examples, Quality of study material, Quality of Tutorial / Assignment, Students' Query Satisfaction, Sincerity and Discipline and Students Motivation. Students are asked to give their feedbacks based on these attributes. These feedbacks are structured in nature while previous types were unstructured in nature. Therefore it increases our understanding in second dimension means about subject knowledge of a teacher. While previous approach is telling about students understanding and performance for the subject. But here source of information is students only. Therefore quality of data depends on the sincerity of students and inclusion of all types of students. Anyhow it is assumed that data is given sincerely by the students. Collected data can be seen in the following table which can be analyzed further.

Table- IV: Attributes wise average marks in the structured Format

S. No.	A	Max. Marks	T_1 Marks	T_2 Marks	T_3 Marks	T_4 Marks	T_5 Marks
1	A1	10	7	6	5	5	9
2	A2	10	8.9	9	7	6.5	9
3	A3	10	8.9	8	8	7	8.5
4	A4	10	9.3	8	6	6	9
5	A5	10	9.2	8.5	8	7.5	9
6	A6	10	9.3	7	7	8	8
7	A7	10	9.3	8	7	7.5	9
8	A8	10	9.1	8	7	6.5	8
9	A9	10	9	8.5	8	7.2	8.5
10	A10	10	9	9	8.5	6	8.6
Total		100	89	80	71.5	67.2	86.6

Note:
A means Attributes, **A1** means Teaching Experience, **A2** means Teaching Methodology, **A3** means Students Engagements in the class, **A4** means Subject Knowledge, **A5** means Presentation Skill with real life examples, **A6** means Quality of study material, **A7** means Quality of Tutorial / Assignment, **A8** means Students' Query Satisfaction, **A9** means Sincerity and Discipline, **A10** Students Motivation

After getting structured data as discussed earlier unstructured data (reviews in the form of text) is also collected about every teacher who are teaching same subject. These reviews also parsed using Stanford parser and analyzed them like subject's reviews discussed earlier. Opinion of each review is collected and analyzed either positive or negative. Corresponding to each teacher number of positive opinions and negative opinions are presented in the table number 5 and quantities like SPN (Sum of positive and negative sentiments), Positive ratio (P_Ratio) and Negative ratio (N_Ratio) are calculated as follows.

$$SPN = \text{Positive Sentiments}\,(P) + \text{Negative Sentiments}(N)$$

$$P\_Ratio = \frac{P}{SPN}$$

$$N\_Ratio = \frac{N}{SPN}$$

These quantities are made and applied to check the performance of the different parameters as shown in table number 5.

Table- V: Marks obtained through unstructured format (written in free style)

Teacher No.	P	N	SPN	P_Ratio	N_Ratio
T_1	50	40	90	0.55	0.44
T_2	45	30	85	0.52	0.33
T_3	38	25	63	0.60	0.39
T_4	30	42	72	0.41	0.58
T_5	53	25	78	0.68	0.32

To categorize the above data a formula is developed. Main thing with this formula is balancing between structured and unstructured data through equal weights. But weight can be changed according to the requirements.

Y = Φ_1(Score Obtained through Structured format) / Total + Φ_2 (Score Obtained through Unstructured format) / SPN

Where Φ_1 and Φ_2 are balancing factor such that

$0 < \Phi_1 < 1$

$\Phi_2 = 1 - \Phi_1$

Here values of Φ_1 and Φ_2 can be set as per the importance of structured and unstructured data. Following results are obtained using formula mentioned above.

Table- VI: Teacher vs. Classifying Attribute

Teacher No.	Classification Attribute(Y)	Class
T_1	0.72	A
T_2	0.69	A
T_3	0.66	A
T_4	0.54	B
T_5	0.77	A

Now depending on value of Y classification is made as per following rule.

If Y Greater than equal to 0.60Then
Class A
Otherwise
Class B

Note:
T_1, T_2, T_3, T_4, T_5 are teacher Number
Class A means good and Class B means need to improvement

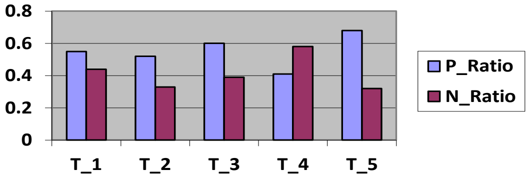


Fig. 7.Teacher versus P_Ratio & N_Ratio

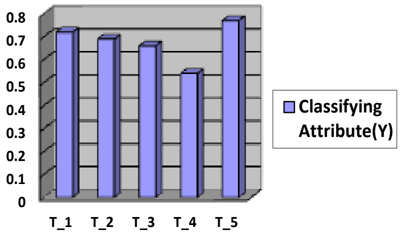


Fig. 8.Teacher Versus Classify attributes

Result Interpretation: Teacher number 4 needs to improve his/her performance.

IV. FUTURE PERSPECTIVES

Web and its use developing gradually which causes to create perpetually more content and usage data. This will continue expanding significance of Web mining and its procedures. Lots of literature is available in this regard. In [13] some research's future guidelines are discussed. In [14] authors also focused few future course of action that may be followed to ensure progress in the Web mining technologies.

A. **Semantic Web Mining:** Generally search engines search the content from the web based on keywords. These don't know the relationship hidden in the content or content is just human interpretable. In [15] authors discussed how to get insight of the meaning in the sentences and how to make content machine understandable and interpretable. In [16] authors focused on the result improvements by exploiting semantic in the web.

B. **Fraud and threat analysis:** There is considerable raise in attempted online frauds for example use of credit cards in an unauthorized manner after hacking into account database for blackmailing motives. eBay type site can also face auction frauds. In [17] authors described that text mining is the perfect analysis technique for detecting and preventing them. Much more research is needed to be focused on developing new and modified techniques to identify and observe such frauds.

C. **Customer reviews Analysis:** In [18] authors referenced a mainstream research zone where clients' feedbacks about an item or service are examined. This investigation gives some significant data to the client and dealer both. Client thinks about the prevalence and nature of item and trader additionally thinks about interest and shortcomings of the item.

In [19] authors discussed opinion mining throw comparative words. In [20] authors considered different research papers to provide vast critical survey about sentiment analysis. In [21] feature wise opinions are extracted and presented. In [22] authors discussed attitude of customers in terms of sentiment analysis. In [23] sentiment analysis based on gender is done. Tourism related reviews [24], [25] are analyzed through automated sentiment analysis.

D.**Web services optimization:** It is highly needed to make web’s services robust, scalable and efficient because its size is growing day by day. Web digging can be connected for the better understanding the conduct of these services and the knowledge extracted can be valuable for various types of advancements. Along these lines research is likewise required to create Web mining strategies to improve different parts of Web services. In [26] authors focused on query optimization which is very essential for quick response to speed up search in the document.

V. CONCLUSION

This paper introduces text mining, its techniques, application based on Natural Language Processing and Support vector machine and finally future perspectives. In this paper structured and unstructured form of data both are considered as examples to describe about knowledge extraction in a very simple manner. Performance of teachers and students both are analyzed through opinion mining. This application can help other researchers to analyze similar type of data in own way. New techniques and algorithms can be designed and developed for the extraction of more precise knowledge in future work.

REFERENCES

1. M. A. Hearst, “Text data mining: Issues, techniques, and the relationship to information access”. In Presentation notes for UW/MS workshop on data mining Vol. 1, 1997, pp. 99).
2. U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, “Advances in knowledge discovery and data mining”, 1996.
3. E. Cambria., B. Schuller, Y. Xia, andC. Havasi, “New avenues in opinion mining and sentiment analysis”, IEEE Intelligent systems, 28(2), 2013, pp. 15-21.
4. D. Grossman, and P. Domingos, “Learning Bayesian network classifiers by maximizing conditional likelihood”. In Proceedings of the twenty-first international conference on Machine learning ,2004, pp. 46.
5. V. Gupta, and G.S. Lehal, “A survey of text summarization extractive techniques”. Journal of emerging technologies in web intelligence, 2(3), 2010, pp. 258-268.
6. J. P. Callan, “Passage-level evidence in document retrieval”, In SIGIR’94, Springer, London, 1994, pp. 302-310.
7. A. Bharate, D. Gadekar, “Survey Paper on Natural Language Processing”, International Journal of Computer Engineering and Applications, 8(3), 2014, pp. 112-121.
8. S.Sun,C. Luo, and J. Chen, “ A review of natural language processing techniques for opinion mining systems”. Information fusion, 36, 2017, pp. 10-25.
9. Esuli and Sebastiani, SentiWordNet, Available: http://sentiwordnet.isti.cnr.it, accessed 09 Apr 2019.
10. M. Dragoni, S. Poria, and E. Cambria, “OntoSenticNet: A commonsense ontology for sentiment analysis”, IEEE Intelligent Systems, 33(3), 2018, pp. 77-85.
11. K. Xu, S. S. Liao, J. Li, and Y. Song,.” Mining comparative opinions from customer reviews for Competitive Intelligence. Decision support systems”, 50(4), 2011, pp. 43-754.
12. Stanford CoreNLP: Natural Language Software. Standford University, 2018
13. M.T. Ramakrishna, L.K. Gowdar, M.S. Havanur, and B. P. M. Swamy, “Web mining: Key accomplishments, applications and future directions”. In International Conference on Data Storage and Data Engineering,2010, pp. 187-191.
14. G. Stumme, A. Hotho, and B.Berendt. “Semantic web mining: State of the art and future directions”. Web semantics: Science, services and agents on the World Wide Web, 4(2), 2006, pp. 124-143.
15. K. Pol, N. Patil, S. Patankar, and C. Das, “A Survey on Web Content Mining and extraction of Structured and Semistructured data”. In First International Conference on Emerging Trends in Engineering and Technology, 2018 , pp. 543-546.
16. E. Cambria,D. Das, S. Bandyopadhyay, and A. Feraco, “ A practical guide to sentiment analysis”. Cham, Switzerland: Springer International Publishing, 2017.
17. R. Bhowmik, “Data Mining Techniques in Fraud Detection”, Journal of Digital Forensics Security and Law, 3(2),2008, pp. 35-53.
18. A. M. Popescu, and O. Etzioni, “ Extracting product features and opinions from reviews”. In Natural language processing and text mining, Springer, London, 2007, pp. 9-28.
19. M. Hu, and B. Liu, “Mining and summarizing customer reviews” In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining ,2004 , pp. 168-177, ACM.
20. W. Medhat, A. Hassan, and H. Korashy, “ Sentiment analysis algorithms and applications: A survey”. Ain Shams engineering journal, 5(4), 2014, pp. 1093-1113.
21. M. Nakayama, and Y.Wan, “Is culture of origin associated with more expressions? An analysis of Yelp reviews on Japanese restaurants”, Tourism Management, 66, 2018, pp. 329-338.
22. Hussein, D.M. El-Din Mohamed, “ A survey on sentiment analysis challenges”. Journal of King Saud University-Engineering Sciences, 30(4), 2018, pp. 330-338.
23. M. Thelwall, “Gender bias in sentiment analysis. Online Information Review”, 42(1),2018, pp. 45-57.
24. A.P. Kirilenko, S.O. Stepchenkova, H. Kim, and X. Li, “ Automated sentiment analysis in tourism: Comparison of approaches”, Journal of Travel Research, 57(8), 2018, pp. 1012-1025.
25. A. R. Alaei, S. Becken, and B. Stantic, “ Sentiment analysis in tourism: capitalizing on big data. Journal of Travel Research”, 58(2), 2019, pp. 175-191.
26. U. Srivastava, K. Munagala, J. Widom, and R. Motwani, “ Query optimization over web services”. In Proceedings of the 32nd international conference on Very large data bases (pp. 355-366). VLDB Endowment.

AUTHORS PROFILE

He did M.Tech.(Computer Science) from UPTU, Lucknow, presently known as AKTU, Lucknow. He is pursuing Ph.D.(CSE) from the same university. His area of interest includes Data Mining, Information Retrieval and Machine Learning. He has been working as Associate Professor in the department of Information Technology in KIET Group of Institutions affiliated to AKTU, University, Lucknow, UP., India.

He did Ph.D. from Department of Computer Science and Engineering, Motilal Nehru National Institute of Technology (MNNIT), Allahabad, India. His area of research includes Data Mining, Information retrieval and Data Compression. At present he is working as Director, School of Science, UP Rajarshi Tandon Open University, Allahabad, India.

He did Ph.D. from Jamia Millia Islamia, New Delhi, India. His area of interest includes Neuro Fuzzy System and Information Retrieval. At present he is working as Professor and Head of Department in Electrical Department of KIET Group of Institutions, Ghaziabad, UP, India.