A REVIEW: DATA PREPROCESSING AND TECHNIQUES OF TEXT MINING

Neeta Yadav and Dr. Neelendra Badal
M.Tech Scholar, Computer Science & Engineering, KNIT, India
Professor, Computer Science & Engineering, KNIT, India

ABSTRACT: The huge amount of data continuously generated in the world every day and it is very difficult task of extracting meaningful information from large amount of data, fetching meaningful information from raw data is challenging task. Data mining is done basically for fetching valuable information from the large amount of data. There are numerous research fields related to data mining are text mining, web mining, sequential pattern mining, medical mining, multimedia mining, structural mining, and graph mining. Text mining nowadays very emerging research area, it is the process of mining meaningful and valuable information from a textual document. Preprocessing is crucial step in text mining or data mining. Preprocessing of raw data includes removal of missing values and discretization. Missing values can be handled by replacing missing value technique and by deletion method. Discretization responsible for converting continuous attributes into intervals and relating each interval with some specific data value and it is handled by the binning method like equal width binning, equal frequency binning and by entropy-based discretization method. It may be supervised or unsupervised but doing discretization sometimes responsible for the loss of information.

Keyword- Discretization, Missing Value, Stemming, Stop Words, TF/IDF, Tokenization

1. INTRODUCTION

In this paper, the review is done on techniques of text mining that is tokenization, stemming, stop word removal and TF/IDF. Further the review is done on missing values and methods of handling missing values like replace the missing value and deletion method and it gives the overall comparison about all missing value handling strategy. And in last of the paper, a review is done on discretization and methods of discretization like equal width binning, equal frequency binning and entropy-based discretization and a comparison is also established between all these.

2. TEXT MINING TECHNIQUES

It is basically the process of fetching high quality of information or deriving high quality of information from text. It is also referred as text data mining or text analytics. Nowadays text mining is emerging field which gives the better result and team efficiency. Text mining is just like data mining actually text mining uses data mining methods to uncovering meaningful or valuable patterns from unstructured data. Data mining concerned with structured data and text mining deals with unstructured data or semi-structured datasets. Its main aim is to discover hidden facts.

A. Text Mining Process
B. Text Preprocessing
C. Feature Generation
D. Feature Selection
E. Text/Data Mining
F. Analyze Results

Preprocessing is very crucial step in text mining processes it plays an essential role in text mining and in its application, it is time taking process of TDM. It involves techniques that transform raw data into an understandable pattern. Its Key steps are tokenization, Stemming, Stop Word Removal and TD/IDF.

(Anjali Ganesh, 2009) discussed the purpose of removal or reduction of stemming. It's main work to reduce its inflectional form and also into its derivational form to its base form. In this paper also discusses the methods of stemming and their comparison and also its limitations. In this paper she also focuses on the basic differences between stemming and lemmatization.

(Ramasubramaniam & Ramya, 2013) contribute his work for improvement in stemming techniques which are used text preprocessing of text mining. This paper focuses on one of the disadvantages of the porters' algorithm. In this paper, Spell check is introduced to overcome the wrong matches and for increasing efficiency. So it saves time from processing for the misspelled.

(Vishal & Gurpreet, 2013) analyses the effectiveness and performance of stemmers in applications like spelling checker regarding languages. This paper gives detailed information regarding several stemming techniques and existing stemmers for Hindi languages.

Hassan Saif focused in his paper whether removal of stop words helpful in adding effectiveness of Twitter sentiment classification methods, for achieving this he has applied six different approaches to twitter data from six distinct datasets. The outcome shows that using the precompiled list of stop words degrades the performance of Twitter sentiment classification approaches. On the other side, the dynamic generation of stop words lists by removing those infrequent terms. " (K.K Agbele , 2012) discussed the methods for the development of pervasive computing application that is adaptable for users, in this paper context-aware stemming algorithm (CAS) is proposed, which is a modified version of the porters stemmer, considered only generated meaningful stemming words as its output, the result shows that modified algorithm reduces the error rate of porters algorithm".

2.1 PREPROCESSING METHODS

2.1.1 Tokenization

Tokenization is the very first step of morphological analyses. Tokenization means basically the breaking of the sentence of text into words, symbols, phrases etc. called tokens. Tokenization provides consistency in the documents. Tokenization is not always fruitful because sometimes it slows down the whole information retrieval process. Below in figure 1, it is described by an example that is – this girl belongs to Varanasi.

Example:-

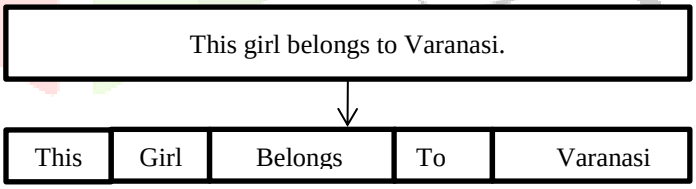

Figure1: It depicts the tokenization of sentence, sentence-this girl belongs to Varanasi. Tokens-this, girl, belongs, to, Varanasi.

2.1.2 Stop Words Removal

Stop words are those words which are used very frequently and it is considered as a useless in information retrieval and text mining. Stop words are like articles, pronouns, prepositions, and conjunctions. In preprocessing the first step is to remove stop words present in the document. The main objective of removing stop words is to make text heavier. See figure 2, stop word removal described below with suitable example.

Example:-

Sentences	After stop word removal
I like reading, so I read	Like, reading, read

Figure 2- It depicts the stop words removal process.

Various methods for stop words removal are as the following:-

1. Classic Methods
2. Methods based on Zipf's law (Z-methods)
3. The mutual information method (MI)
4. Term-based random sampling (TBRS)

2.1.3 Stemming

In stemming, process words are reduced to their root. Stemming is one of the mechanisms of preprocessing which is responsible for a reduction of words into their roots. It is responsible for a reduction of words to base form. Stemmers are used to optimize retrieval performance and to reduce the size of indexing files. Stemming generally increases memory at cost of decreased precision. The various stemming algorithm has been developed to optimize the data. There are two points which must follow while using the stemmer

1. Words should keep separate which have the different meaning.
2. For Morphological forms of a word, an assumption made that word have same meaning mapped to the same root.

See figure 3, stemming is described below with suitable example.

Example:–

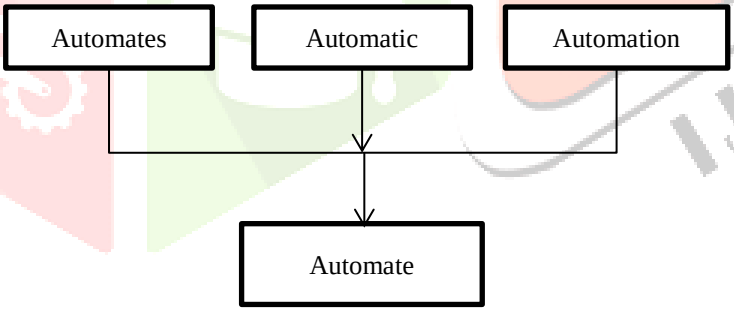

Figure 3 - It represents the stemming process that is the words automate, automatic and automation reduced to their stem word automate.

2.1.3.1 Stemming algorithm

There are various stemming algorithm which are explained below:-

2.1.3.1.1 Brute Force Method

It is one of the simplest methods of stemming, it is based on the dictionary which is consisting of inflection of a word, and the dictionary is used as a look up. The disadvantage of brute force method is that its speed is very slow, in this whole stemming process requires too many resources in storage.

2.1.3.1.2 Lovins Stemmer

In 1968 Lovins proposed this effective stemmer, it is responsible for removing longest suffix from the word. Due to single pass algorithm, it removes only one suffix from a word at maximum. Its advantages are that it is very fast.

Disadvantage

1. It is very time-consuming.
2. Many times it fails to form words from terms or to match the stem of like meaning words.

2.1.3.1.3 Porters

It is proposed in 1980 and it most successful stemming algorithm, it has 60 rules and its rules are very easy to understand, the Lovins algorithm is heavier algorithm than porters algorithm. Basically, it is based on the idea that suffixes in the English languages are mostly made up of the grouping of smaller and simpler suffixes, this algorithm not able to concentrate on lexicon study of the word when original dimensionality changes.

2.1.3.1.4 Paice/Husk Stemmer

It is an iterative algorithm which contains 120 rules.

Disadvantage

This algorithm leads to the over stemming problem because it is a very heavy algorithm.

Advantage

- Simple and easy to understand.
- It is an iterative algorithm so each iteration taking care of both replacement and deletion.

2.1.3.1.5 Dawson Stemmer

It is an extended version of the Lovins algorithm but it covers about 1200 suffixes. But it is very complex in nature and it lacks standard reusable implementation.

2.1.4 TF/ IDF

TF stands for term frequency and it is used as a weighting factor in text data mining and in information retrieval process. It helps in telling how the word important to a document. It is widely used for filtering stop words in various fields of classification and categorization. Term frequency means the number of times word or term occur in the document and inverse frequency used for measuring the importance of the word in the textual document. When its feature introduced it reduces the term weight which occurs very frequently and it is also responsible for increasing the term weight which occurs rarely.

Example

Document (D) = 120 tokens

Appeared (t) = 5 times

TF = No. of times the token appear in the document/ Total No. of Token in the document

= 5/120

=1/24

IDF stands for Inverse Document Frequency and it prizes token that are rare overall in a dataset, and it adds more meaning to the document.

It is computed as follows.

IDF (t) = Total number of document/ number of document in S that contains t

=N/n (t)

See table 1, summarized view of text mining techniques given below.

Table 1: Summarization of Text Mining Techniques

	Text Classifier	**Information Extraction**	**Summarization**	**Question Answering**	**Machine Translation**
Stop Word	√	×	×	×	×
Stemming	√	×	×	×	×
Tokenization	√	√	√	√	√

It depicts the overall view of preprocessing techniques of text mining in summarized form.

3. MISSING VALUE

Data cleaning is one of the most important tasks in preprocessing and it is very difficult to deal with or placing missing value in the dataset is one of the major tasks of preprocessing. The popular method of dealing with missing value is either delete or replace with mean. Data can be missing from dataset because of the following reason

- Malfunctioning of equipment
- Data may be lost
- Inconsistent with other recorded data so it deleted

Missing Data can be handled

- If we are able to identify its reason and pattern.
- How much data is missed?
- It can be handled by imputation method.

Types of Missingness

- **MCAR**
  Missing completely at random is depends upon observed and unobserved measurement.
- **MAR**
  Missing at random only related to observed data.
- **MNAR**
  It stands for missing not at random, it depends upon missing and observed data values both, and it is related to unseen data so it is impossible to determine its mechanism.

Former Strategy of dealing with missing value

- Replacing missing values by the most frequent value of the attribute.
- “For numerical attributes, the missing attribute value may be replaced by the attribute average value."
- Assigning all possible values of the attribute.
- Considering missing attribute values as special values.

Methods of handling missing data

- Deletion methods
- Single imputation method
- Model-based Methods

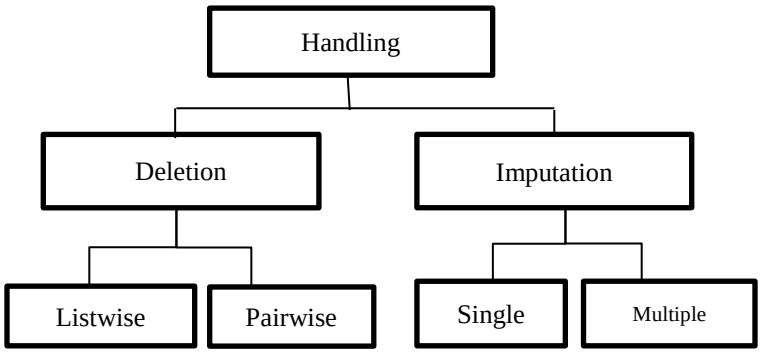

Figure 4- It depicts the handling method of missing value.

3.1 Deletion Methods

It is the simplest method for dealing with missing values; it is suitable for MCAR. Generally, it has two ways of dealing with MCAR that is Listwise deletion method and Pairwise deletion method. See table 2, listwise deletion method discussed below with suitable example.

Example

Table 2: Listwise Deletion Method

Roll No.	Name	Gender	Age	Subject	Marks
1	A	F	22	Maths	77
2	B			Maths	88
3	C	M	23	Maths	55
4	D	F		Maths	89
5	E	M	21	Maths	

It depicts the listwise deletion method shown by dark lines in row 2, 4, 5. In this method if any row has missing attribute the entire row is deleted in the above example row 2, 4, 5 is deleted from the dataset.

Example

See table 3, pairwise deletion method discussed below with suitable example.

Table 3: Pairwise Deletion Method

Roll No.	Name	Gender	Age	Subject	Marks
1	A	F	22	Maths	77
2	B	____	____	Maths	88
3	C	M	23	Maths	55
4	D	F	____	Maths	89
5	E	M	21	Maths	____

It depicts the pairwise deletion method shown by dark lines in attributes gender, age and number. In this method we delete only those cells having missing values entire row or column not deleted this method gives the highest accuracy.

3.2 Mean Substitution Method

In this method, missing values replaced by their mean and if the missingness is categorical then it is replaced by most frequent occurrence. See table 4, mean substitution method discussed below with suitable example.

For Example:-

Table 4: This table having some missing values in some attributes

Roll No.	Name	Gender	Age	Subject	Marks
1	A	F	22	Maths	77
2	B	F	____	Maths	88
3	C	M	23	Maths	55
4	D	F	____	Maths	89
5	E	M	21	Maths	_____

Mean of age = (22+23+21)/3

=22

Mean of marks = (77+88+55+89)/4

=77.25

So the missing age value is replaced by 22 and marks by 77.25 that are both by their mean value shown in the table 5 below:

Table 5: After substitution of the mean value in missing attributes age and marks shown by dark line.

Roll No.	Name	Gender	Age	Subject	Marks
1	A	F	22	Maths	77
2	B	F	22	Maths	88
3	C	M	23	Maths	55
4	D	F	22	Maths	89
5	E	M	21	Maths	77.25

Table 6: Summarized View of All Missing Handling Method

Handling Methods	Approach	Description	Assumptions to achieve unbiased estimates	Advantages	Disadvantages
Listwise deletion or Complete case analysis	Traditional method or deletion method	Drop the observation with the missing value from analyses.	MCAR	Easy to use. High accuracy.	Reduces sample size. Estimates may be biased if data, not MCAR
Pairwise deletion or average case analysis	Traditional method or deletion method	Drop the observation with the missing value only for analyses using that variable.	MCAR	It is simple to use. Keeps as many cases as possible for each analysis. Highest accuracy.	Can't compare analysis because sample different each time.
Mean / mode Substitution	Single imputation	Replace the missing value with the mean/mode of the variable.	MCAR, only when estimating mean	Easy to use. Saves data; Preserves sample size. Moderate accuracy	Potentially biased results. Reduces variability.
Dummy variable method	Single imputation	Include an indicator variable for those missing the variable.	MCAR, only when estimating mean	Uses all available information about missing observation	Results in biased Estimates.
Single regression	Single imputation	Replaces missing values with the predicted score from a regression equation.	MCAR, only when estimating mean	Uses information from observed data.	Overestimates model fit and correlation estimates weaken variance.
Multiple imputations	Model-based	Create multiple data sets that impute different values to the missing.	MAR(but can handle both MAR and MNAR)	Variability more accurate for each missing value.	Room for error when specifying models.
Full information maximum likelihood	Model-based	Uses all information but does not impute values to the missing.	MAR(but can handle both MAR and MNAR)	Uses full information log likelihood to calculate. Unbiased parameter estimates with MCAR/MAR	Standard errors biased downward can be adjusted by using observed information matrix.

It depicts the summarized view of all missing value handling method.

4. DISCRETIZATION

Data discretization is one of the major tasks of preprocessing by the discretization method we are able to convert the continuous value into intervals, it is easier to deal with discrete attribute instead of a continuous attribute. Its main goal to reduce the number of values a continuous attribute into discrete values.

Many researchers like Apte and Hong, Cendrowaka and Clark and Niblett are not able to deal with continuous values so they can do their operation on discretized attributes, working with discretized attributes often needs less memory and it takes less preprocessing time, its disadvantage is it lost some information while discretizing, loss of information increases the rate of misclassification. Discretization methods can be of

supervised or unsupervised, local and global. Equal frequency binning is unsupervised or local discretization algorithm.

CAIM–it stands for class-attribute independence maximization and it is responsible for minimization of loss between class-attribute interdependency and it finds the minimum number of intervals.

Supervised discretization is better than the unsupervised method of discretization this conclusion made because of the study of (Holte, 1993) and (Dougherty, 1995) respectively.

In 1995 entropy-based discretization is proposed by Pfahringer and it is better than supervised discretization this is proposed by Kohavi and Sahami.

(Marzuki & Ahmad, 2007) suggest different discretization method which performs well in different areas. Many discretization methods responsible for a loss of information so MIL discretizer introduced in 2011, it has much scope for improvement in this algorithm, its performance basically compared with the supervised MIL. There are various methods for discretization like

4.1 Unsupervised Discretization

This type of discretization having no class label and it is of two types which are described below:-

4.1.1 Equal-Width Binning

In this method, the range is divided into N intervals of equal size.

For example:-

Age:- 6,7,8,8,12,19,22,22,23,25,34,34,41,45,45.

Interval width calculated by: - W (width) = (maximum-minimum)/N

Suppose we have to divide into 3 intervals then:

= (45-6)/3

=13.

Thus, we have three intervals which are given below in table 7.

Table 7: It depicts Equal Width Binning intervals and its counts.

Intervals	Counts
[6,19]	6
[20,33]	4
[34,47]	5

4.1.2 Equal-Frequency Binnig

In this method the range that is the value of attributes is divided into N intervals, each containing near about same number of elements or sample. See table 8, there is 3 bins having different number of samples and having different number of counts.

For example:-

Age:- 6,7,8,8,12,19,22,22,23,25,34,34,41,45,45.

Bin Size=3

Table 8: It depicts EFB with 3 bins having different number of samples and its counts.

Bins	Samples	Counts
Bin 1	6,7,8,8,12	5
Bin 2	19,22,22,22,23,25	5
Bin 3	34,34,41,45,45	5

4.2 Supervised Discretization

In this method, an attribute is discretized by using class information, and it generates intervals which lead to less information loss. Entropy-based discretization is one of the examples of supervised discretization.

4.2.1 Entropy-based Discretization

It is one of the most commonly used discretization methods, it is a supervised discretization technique and it is based on the class label. It is responsible for finding best splits so that bins are as pure as possible.

Goal: - Split with maximum information gain that is less information loss. Entropy can be defined as the following:-

Entropy (D) = -p1*log2 (P1)-P2*log2 (P2)

Entropy is ranging between 0 and 1, if the entropy is low it means dataset is relatively pure and if the entropy is high it means the data set is mixed.

Suppose a given set of samples S, if S is divided into two intervals S1 and S2 using boundary T, then expected information requirement after partitioning is

E(S, T) = |S1|/|S| Entropy (s1) +|S2|/|Entropy (s2).

See figure 7 and table 9 Entropy based discretization described below.

Example: This table consists of two attributes O-ring failure and temperature.

		O-Ring Failure	
Tempt		Y	N
	<=60	3	0
	>60	4	17

Figure 7- It depicts O-Ring Failure.

Table 9: It depicts entropy-based discretization having attributes O-ring failure and Temperature, which are given below.

O-Ring Failure	Temperature
Y	53
Y	56
Y	57
N	63
N	66
N	67
N	67
N	67
N	68
N	69
N	70
Y	70
Y	70
Y	70
N	72
N	73
N	75
Y	75
N	76
N	76
N	78
N	79
N	80
N	81

Step 1:- Calculate the Entropy
P1 (Yes) =7/24; P2 (No) =17/24;
E (failure) =0.871;
Step2:- Calculate Entropy for the target given bin
P1 (<=60) = 3/24; P2 (>60) = 21/24
E (Failure, Temperature) = P1 (<=60)*E (3, 0) +P2 (>60)*E (4, 17)
=3/24*0+21/24*0.7
=0.6125
Step3:- Calculate the information gain-
Information Gain=E(S)-E(S, A)
Information Gain (Failure, Temperature) = 0.256
Note:-Interval (<=60,>60) gives highest information gain that's why this boundary is chosen.

Table 10: Comparison Table for Discretization Method

Method	Main Points	Information loss	Status	Implementation	Which one is better in dealing with outlier	Class label	Supervised	Unsupervised
Equal-width binning	Interval size of each bin is same.	More	Good	Simple	×	×	×	√
Equal frequency binning	Each bin contains approximately same number of data.	More	Better	Simple	√	×	×	√
Entropy based binning	Finds best split so it gives more information gain	Less	Best	Complex	√	√	√	×

It depicts comparisons between equal- width binning, equal frequency binning and entropy-based binning.

5. CONCLUSION

This paper provides information regarding preprocessing of data or text with examples and comparisons with other techniques especially in context of missing value and discretization. Observation has been made that, preprocessing is a very crucial task of text data mining or data mining. Handling missing value by Mean Replacement is the former strategy but still valuable but sometimes it leads to inaccuracy and inconsistencies. By deletion method more accuracy is achieved but it leads to loss of information and some level of inconsistencies. Deletion method produces a biased result if data are not MCAR. It is also found that entropy-based discretization is better but it is complex in nature. Discretization by equal width binning and by equal frequency binning is simple and suitable for small dataset but it leads to information loss. Unsupervised discretization has no class label and they are not able to deal smartly with outlier so there is some scope of study in this section.

REFERENCES

[1] Agbele, K., K., Adesina, A., O., Azeez, N., A. and Abidoye, A., P. (2012).Context-Aware Stemming Algorithm for Semantically Related Root Words. African Journal of Computing and ICT, 5(4), 33-42.

[2] Agre, G., Peev, P. (2002). On Supervised and Unsupervised Discretization. Institute of Information Technologies, 1113 Sofia, Faculty of Mathematics and Informatics, Sofia University, and 1000 Sofia, 2(2).

[3] Alosh, M.(2009).The Impact of Missing Data in a Generalized Integer-Valued Autoregression Model for Count Data. J Biopharm Statistical, vol.19, no. 6, pp.1039-1054.

[4] Bernaarrds, C., A., Belin, T., R. and Schafer, J., L. (2007).Robustness of a Multivariate Normal Approximation for Imputation of Incomplete Binary Data. Statistics in Medicine, 26, 1368-1382.

[5] Blackwell, M., Honaker, H. and King, G. (2007) A Unified Approach to Measurement Error and Missing: Details and extensions, in Sociological Methods and Research, 46, 342-69.

[6] Buuren, S., V. (2011).Multiple Imputations of multilevel data. J.J Hox and J.K Robert (Eds.).Handbook of advanced multilevel analyses, New York, 173-196.

[7] Buuren, S., V. (2007).Multiple Imputation of discrete and continuous data by fully Conditional Specification. Statistical Methods in Medical Research, 16(3), 219-242.

[8] Cano, A., Nguyen, D., T., Ventura & Cios, K., J.(2016). Ur-Caim: Improved Caim discretization for unbalanced and balanced data. Soft Computing, 20, 173-188.
[9] Carroll, R., J., & Stefanski, L., A. (1990).Approximate Quasi-Likelihood Estimation in Models with surrogate predictors. Journal of the American Statistical Association, 85, 652-663.
[10] Cismondi, F., Fialho, A., S., Vieira, S., M., Reti, S., R., Sousa, J., M., C. & Finkelstein, S., N. (2013). Missing data in medical databases: impute, delete or classify? .Artificial Intelligence in Medicine, 58, 63-72.
[11] Craig, K. & Bandalos, D., L.(2001). The Relative Performance of full information maximum likelihood estimation for missing data in structural equation models. Structural Equation Modelling, 8(3), 430-457.
[12] Dastida, B., G. & Schafer, J., L.(2003).Multiple Edit/Multiple Imputation for Multivariate Continuous Data. Journal of the American Statistical Association, 98, 807-817.
[13] Dougherty, J., Kohavi, R. & Sahavi, M.(1995). Supervised and Unsupervised Discretization of Continuous Attributes. Proceedings of the 12th International Conference on Machine Learning, 194-202.
[14] Dougherty, J., Kohavi, R. & Sahami, M.(1995). Supervised and Unsupervised Discretization of Continuous features. Proceedings of the 12th International Conference, 12, 194-202.
[15] Efron & Bradley.(2013).Large-Scale Inference: Empirical Bayes Methods for Estimation, testing, and Prediction. Cambridge University Press, United Kingdom.
[16] Garcia, S., Luengo, J., Saez, J., A., Lopez, V. & Herrera, F. (2013). A Survey of discretization techniques: Taxonomy and Empirical analysis in supervised. IEEE Transaction Knowledge Data Engineering, 25(4), 734-750.
[17] Gupta, V. & Lehal, G., S. (2013). A Survey of Common Stemming Techniques and Existing Stemmers for Indian languages. Journal of emerging Technology in Web Intelligence,5(2), 36-39.
[18] Gorisek, Ales, Pahor, Mauka (2016).Missing Value Imputation using Contemporary Computer Capabilities: an application to financial statements data in large panels. Economic and Business Review, 19, 97-119.
[19] Jivani, A., G., (2011). A comparative Study of Stemming Algorithm. International Journal of the Computer, Technology, and Application, 2, 1930-1938.
[20] Holte, R., C. (1993).Very Simple Classification Rules Perform Well on Most Commonly Used Datasets. Machine Learning, 11, 63-91.
[21] Honaker, J. & King, G. (2010).What to do about Missing Values in time series cross section Data. American Journal of Political Science, 54 (1), 561-581.
[23] Kurgan, L., A. & Cios, K.J. (2004) CAIM Discretization Algorithm. IEEE Transactions on Knowledge and Data Engineering, 16(2), 145-153.
[24] Liu, H., & Setiono, R. (1997).Feature selection via discretization. IEEE Transaction Knowledge Data Engineering, 9(4), 642-645.
[25] Marzuki, Z., & Ahmad, F. (2007).Data Mining Discretization Methods and Performances. Proceeding of the International Conference on Electrical Engineering and Informatics Institute of Technology Bandung, Indonesia, 535-537.
[26] Mistler, S., A. & Enders, C., K. (2016). A Comparison of Joint Model and Fully Conditional Specification Imputation for Multilevel Missing Data. Journal of Educational and Behavioral Statistics, 42(4), 432-436.
[27] Pfahringer, B. (1995).Supervised and Unsupervised Discretization of Continuous features. Proceedings of the 12th International Conference on Machine Learning, 456-463.
[28] Raghunathan, T., E., Lepkowski, J., M., Hoetoyk, J., V. & Solenberger, P. (2001). A Multivariate Technique for multiply imputing missing values using a sequence of regression models. Survey of Methodology, 27(1), 85-95.
[29] Ramasubramaniam, C. and Ramya, R. (2013). Effective Pre-Processing Activities in Text Mining using Improved Porters Stemming algorithm. International Journal of advanced Research in computer and communication engineering, 2, 4536-4538.

[30] Read, S., H., Lewis, S. C., Halbesma, N. & wild, S., H. (2017).Measuring the Association between Body Mass Index and All- cause Mortality in the Presence of Missing Data: Analysis from the Scottish National Diabetes register. American Journal of Epidemiology, 185(8), 641-649.
[31] Rosenfeld, A., llluz, R., Goltesman, D. & Last, M.(2018).Using Discretization for Extending the Set of Predictive Features. Eurasip Journal on Advances in Signal Processing.
[32] Rosenfeld, A., Grahan, D., G., Hamoudi, R., Butawan, R., Eneh, V., Khan, S., Miah, H., Niranjan, M. & Lovat, L., B.(2015). MIAT: A novel attributes selection approach to better predict upper gastrointestinal cancer. IEEE International Conference on Data Science and Advanced Analytics, 1-7.
[33] Rubin, D., B. (1988).An Overview of Multiple Imputation. Proceedings of the Survey Research Section, American Statistical Association, 79-84.
[34] Ruiz, F., J., Angulo, C. & Agell, N. (2008). IDD: A Supervised internal distance-based method for discretization. IEEE Transaction Knowledge Data Engineering, 20(9), 1230-1238.
[35] Schafer, J., L. (1999).Multiple Imputation: A Primer. Statistical Methods in Medical Research, 8(1), 3-15.
[36] Shin, Y. & Raudenbush, S., W. (2010).A latent Cluster Mean Approach to the Contextual effect model with missing data. Journal of Education and Behavioral Statistics, 35(1), 26-53.
[37] Srividhya, V. & Anitha, R. (2010).Evaluating Preprocessing Techniques in Text Categorization. International Journal of Computer Science and Application, 47(11), 36-39.
[38] Sukanya, M. & Biruntha, S. (2012).Techniques on Text Mining. IEEE, International Conference on Advanced Communication Control and Computing Technologies (ICACCCT), 269-271.
[39] Ventura, D. and Martinez, T., R.(1995). An empirical Comparison of Discretization Methods. Proceedings of the 10th International Symposium on Computer and Information Sciences, 443-450.
[40] Young, W., Weckman, G. & Holland, W. (2011).A Survey of Methodologies for the treatment of missing values within datasets: limitations and benefits. Theoretical Issues in Ergonomics Science, 12(1), 15-43.
[41] Zhong, N., Li, Y. & Wu, S., T. (2012).Effective Pattern Discovery for Text Mining. IEEE Transaction on Knowledge and Data Engineering, 24(1), 30-44.

Authors Profile

Neeta Yadav pursued Bachelor of technology from UNSIET, VBS Purvanchal University in 2016 in Information Technology. She is currently pursuing M.Tech in Computer Science and Engineering from Kamla Nehru Institute of Technology, Sultanpur. Her area of interest is Data mining and Warehousing, Software Testing, Cryptography and Network Security and Computer Network. Currently, she is doing her dissertation in the field of Data mining.

Dr. Neelendra Badal is Professor in the Department of Computer Science and Engineering at Kamla Nehru Institute of Technology (KNIT), at Sultanpur (U.P.). He pursued Ph.D. in 2009 from MNNIT, Allahabad in Computer Science and Engineering. He is Chartered Engineering (CE) from Institution of Engineers (IE), India. He has published more than 65 papers in International/National Journals, Conferences and Seminars. He is a Life member of IE, IETE, ISTE, and CSI India. His research interests are Distributed System, Parallel Processing, GIS, Data Warehouse and Data Mining, Software Engineering, and Networking.