


# Predicting Marathi News Class Using Semantic Entity-Driven Clustering Approach

Jatinderkumar R. Saini, Symbiosis Institute of Computer Studies and Research, Symbiosis International University (Deemed), India

 <https://orcid.org/0000-0001-5205-5263>

Prafulla Bharat Bafna, Symbiosis International University (Deemed), India

## ABSTRACT

Document management is a need for an era, and managing documents in the regional languages is a significant and untouched area. Marathi corpus consisting of news is processed to form group entity document matrix Marathi (GEDMM), vector space model for Marathi (VSMM), and hysynset vector space model for Marathi (HSVSMM). GEDMM uses entity group extracted using condition random field (CRF). The frequent terms are used to construct VSMM using TF-IDF. HSVSMM uses synsets using hypernyms-hyponyms and synonyms. GEDMM and HSVSMM use dimension reduction by selecting significant feature groups. Hierarchical agglomerative clustering (HAC) is used and a dendrogram is produced to visualize the clusters. The performance analysis is carried out using several parameters like entropy, purity, misclassification error, and accuracy. The clusters produced using GEDMM shows the minimum entropy and the highest purity. A random forest classifier is applied, and the results are evaluated using misclassification error and accuracy.

## KEYWORDS

Classification, Clustering, Condition Random Field (CRF), Entity Group, Marathi, Named Entity Recognition, Vector Space Model

## INTRODUCTION

In the contemporary era, the data is present in different languages, processing English text is common and enriched with strong literature but processing data of regional languages like Marathi is a critical task. Abundant data is available in Marathi and classifying Marathi text using dimension reduction by selecting appropriate tokens to involve context in the process of classification is the need of an era.

The textual data is a popular means of information exchange. The available data is divided as structured (tabular form), unstructured (reviews, comments, emails, etc) and semi-structured. (HTML) (Maksimenko et al., 2020) (Gao et al., 2020). Several techniques are available to mine the structured data (Gharehchopogh, & Khalifelu 2011). To process, a huge amount of data text mining techniques are available to generate useful patterns. Clustering, classifications are popularly used techniques (Larsen & Aone, 1999) to identify patterns in the data. There are several steps, which need to be carried out to process the text. Text documents are formed using sentences. The sentences need to be fragmented into tokens by the removal of blanks or other punctuations. Stop words (Meyer et al., 2008) (Aggarwal & Zhai, 2012). do not contribute to the decision-making process and increases

DOI: 10.4018/JCIT.20211001.0a12

This article published as an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>) which permits unrestricted use, distribution, and production in any medium, provided the author of the original work and original publication source are properly credited.

noisy features. These are removed to reduce the dimensions. Lemmatization is an effective way to bring words into their meaning root form (Sharma, 2019). (Kubosawa et al., 2019). For e.g. In the sentence 'राम मंदरिसाठी २८ वर्ष उपास करणारी आधुनिक शबरी', first tokenization is implemented which results into 8 tokens. मंदरिसाठी will be converted to "मंदरि" as a lemma. '२८' will be ignored in the preprocessing steps. 'मंदरि' is tagged as a noun by Part of speech (PoS) tagger.

POS tagging is also a significant activity that identifies nouns, adjectives and other parts of speech and helps to perform morphological analysis of the text. (Bustikova et al., 2020).

Unique lemmas can be weighted using TF-IDF measures that is a deterministic statistical procedure but is preferred due to its simple and effective calculations. It is based on a frequency count of the words and is normalized by considering the word occurrences in the entire corpus.

Corpus is set of processed text data and various operations are applied on it e.g. Preprocessing, linguistic analysis and so on. Processing corpus of the English language is an easy task due to the wide availability of resources, but processing a corpus in regional languages (Hanumanthappa, & Swamy, 2014). like Marathi (Bafna and Saini, 2020 (2)), etc is a challenging and untouched domain of text mining (Vijaymeena and Kavitha, 2016.).

Named Entity Recognition (NER) (Nadeau & Sekine, 2007) is assumed as a subset of information mining. It first detects all proper nouns called as Named Entity from the corpus and assigns each entity with a tag as location, person, etc.

Condition random field (CRF) is a classifier model that finds out the dependency between class and the entity. It gives normalized result by avoiding bias. CRF is used for multi-class problems (Kim et al., 2020)

Clustering is an unsupervised technique which groups unlabeled objects into several groups. Classification is used as a data mining technique for labeled dataset. Clustering can be carried out using several ways. Hierarchical agglomerative clustering is one of the ways and follows a bottom-top approach. It is popularly uses technique for generating groups of texts due to its high accuracy. Every object initiates with its cluster and pairs of clusters are combined while moving up the levels in the hierarchy.

Entropy shows the wrongly classified data objects. Minimum entropy represents relevant clusters. VSMM (Vector space model for Marathi) has documents occurring in rows and dimensions occupying columns. HVSMM represents hysynsets (hyponyms-hyponyms and synonyms in one group) representing features of the matrix. Group Entity document matrix Marathi (GEDMM) represents grouped entities as columns and Marathi news as rows. In all the matrices feature weight is recorded.

The random forest classifier is based upon decision trees. Different numbers of decision trees are produced on training data. The error gets reduced and stabled after a specific number of trees and termed as algorithm convergence. A confusion matrix is produced for either training or testing dataset to check the accuracy of prediction.

There is need of having more accurate techniques to classify Marathi text. Dimension reduction, choosing significant tokens and involving context in the process of classification is the need of an era.

This research unique because

1. Named entity recognition is carried out on the Marathi corpus.
2. Entities belonging to one class are grouped to form GEDMM.
3. GEDMM, VSMM and HVSMM are constructed to carry out clustering and classification.

The paper is arranged as per the mentioned manner. The literature review in the form of current work is depicted as the second section. The detailing of the research methodology is the third section. Results and discussions are stated as the fourth section. The paper completes after the conclusion, denoted by the fifth section.

## LITERATURE REVIEW

The text mining approach is used to generate different questions' lists. It is used to understand the skill set level of a student. The named entity recognizer (NER) is used to generate the MCQ type questions based on distractor. Using n blooms taxonomy questions having descriptive answers (subjective) are prepared using NLP techniques. The system could dynamically generate the questions and reduces memory the occupation concept. (Deena et al., 2020)

The most important research topics are searched using different text mining approaches. Topic modeling using LDA and co-word analysis is used on more than 1000 published research papers. To visualize the data co-occurrence maps are used. Different topics and words are extracted from the research papers. Four clusters of topics are generated and represent the thematic rule. (Soltani et al., 2020). Generally, in the NLP applications researchers determine the significant NLP features using pragmatic knowledge it may result in large efforts in preprocessing or feature engineering. NLP feature specification language are presented which automate reuse of features amongst semantically relevant application. It results into identifying optimal NLP features.

Semantic, statistical and linguistic features are considered for multiple steps of analysis like corpus, documents, paragraphs, sentences, phrases, words. Considering user feedback, a recommendation system is built to enable the features based on the applications. This work could be integrated into machine learning techniques to generate statistically and semantically significant features and to built semantically relevant models. (Srinivasan, 2020).

To extract entities for Indian languages is a difficult task as these languages do not use the concept of capitalization, also these are free order, morphologically and inflectionally rich. Indian names are ambiguous too (Patil et al., 2016). The instances are extracted using techniques namely Inductive Logic Programming (ILP). It constructs rules for mining instances of different named entities. It minimizes the efforts taken by developer/ linguist and constructs an interactive interface. A linguist is expected to input his intuition and background knowledge regarding named entities. The tagged data is used for Hindi and Marathi sentences. Rules and background knowledge are composed manually due to less availability of resources. The approach generates significant patterns. The experiments are scaled to large training data set (Patel et al., 2009).

Handcrafted rules, Hidden markov model and gazetteers are used to identify named entities in the Marathi language. 12 different kinds of NEs are identified. Lemmatization is used to remove inflections. Pseudo word smoothing and replacement is used due to the presence of the sparse data. The viterbi algorithm handles word disambiguation. Grammar rules and gazetteers are used to improve the performance of the system (Patil et al., 2017).

Feature selection and reduction increase the accuracy of the classifier Improved Chi-square method of feature selection known as ImpCHI is applied on the Arabic text to classify. The features are reduced till 900. The technique performs better than usual chi-square and mutual information methods. SVM classifier is used on 5070 sets of Arabic text documents generating total six classes. SVM and Chi-square in integration perform better. Performance evaluation is carried out using f-measures, recall and precision (Kou et al., 2020).

Parameters like author, subject, title can be used to classify the text documents. Arabic text classification methods are studied and presented. These methods focus on neural nets and deep learning. LSTM, CNN, RNN and FFNN are reviewed with their types. Through methodical learning, 12 research papers in the classification field are explored to determine the accuracy of each technique. Several evaluation criteria are applied for the classification. Outcomes provide analysis based on deep learning models to enhance text classification accuracy for Arabic documents (Wahdan et al., 2020)

Named entity recognition is popularly used in English language. But it is slowly evolving in other languages also like Bengali, Urdu, Hindi, Oriya and so on. The dataset scarcity is an issue for the other languages causing relatively less progress in these domains. (Nasar et al., 2021). The actual representation of mapping between the documents and their text is critical to improve classification

performance of document classification task. A preprocessing techniques such as Term weighting represents documents in a vector space and assigns appropriate weights to terms. Computing the correct term weight affects the accuracy of the text classification. MONO weighting scheme is proposed that uses the non-occurrence information of tokens in a comparatively effective manner. Intra-class text document scaling is implemented and supplies more accurate representations of unique capabilities of tokens. These tokens' counting is different in documents but same in the specific class. SRTF-MONO and TF-MONO are proposed using MONO. 3 datasets with two classifiers were used for experiments. A comparative analysis is carried out using 7 weighting schemes. Evaluation is carried out using several parameters (Dogan & Uysal, 2020). To organize the ever-growing textual data classification is applied which produces the groups based on the content of the documents. Text pre-processing, text representation, training model and implementation of model are the four steps which are implemented on BBC news data set. Three classifiers namely random forest logistic regression and K-nearest neighbor are implemented. The comparative analysis is performed using confusion matrix, support, accuracy precision, F1-score (Shah et al., 2020).

Feature selection is an important step in text classification. Impacts of globalization methods are studied based on feature selection. Binary-class balanced (BCB), Multi-class balanced (MCB), Multi-class unbalanced (MCU), binary-class unbalanced (BCU) are used in integration with weighted-sum (AVG), summation (SUM) and maximum (MAX). DFSS, odds ratio along with chi-square (CHI2). Are used on different datasets. These are Polarity Reuters-21578, and 20Newsgroup. Classifiers are applied as Decision tree and SVM. According to the experimental AVG performs best in the case of all algorithms (Parlak et al., 2020).

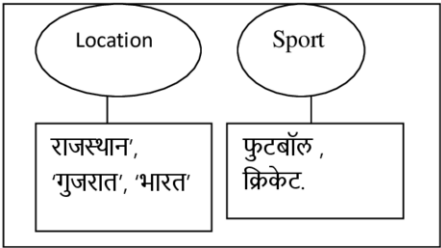
The challenges of text mining initiate right from the conversion of unstructured data to structured one. Choosing the important words, dimension reduction or feature selection, understanding meaning of the words are common challenges which affect the accuracy of the decision making. Multilingual text is another open challenge. The proposed work focuses on all these challenges (Avasthi et al., 2020)

## RESEARCH METHODOLOGY

There is abundance of textual data and the need is to classify the data for the fast information retrieval. Plenty of the techniques are available to process English text but to process text in regional languages like Marathi are scarce being official language of Goa and Maharashtra, thus Marathi news act as input dataset. Considering context of the text while classifying them, results into accurate classes thus context is involved while converting unstructured data into the structured one. Dimension reduction is achieved by forming synsets which facilitates to select significant tokens. Figure 1 shows the need of the proposed technique.

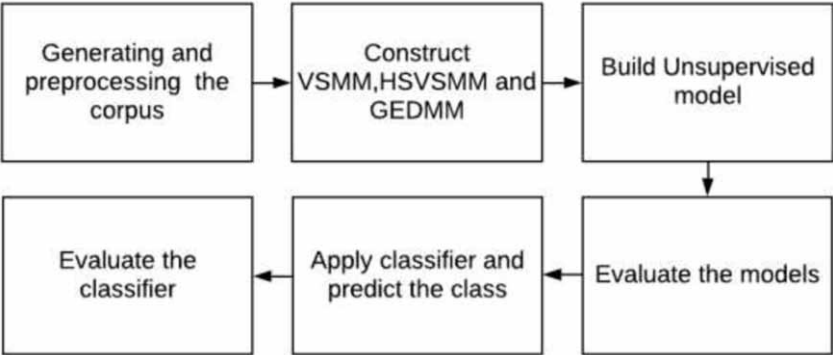
The experimental set up established to group Marathi documents using semantics and context of Marathi words. The corpus of Marathi news is constructed and NER is applied and the entities present in the same domain are grouped to form GEDMM. Preprocessing is followed by calculating

Figure 1. The need of proposed technique



TF-IDF measures of the tokens. HSVSMM which uses hysynset based on hypernyms-hyponyms and synonyms relationship and VSMM (Bafna and Saini, 2020) are generated. Figure 2 shows the entities which are grouped to construct an entity group. The relevance measure of each entity is used to calculate entity group weight. For e.g. ‘राजस्थान’, ‘गुजरात’, ‘भारत’ are the names of locations which are grouped to form one feature. Feature weight is addition of the relevance measure of the entities. Feature weight of sport is sum of relevance measures of फुटबॉल and क्रिकेट.

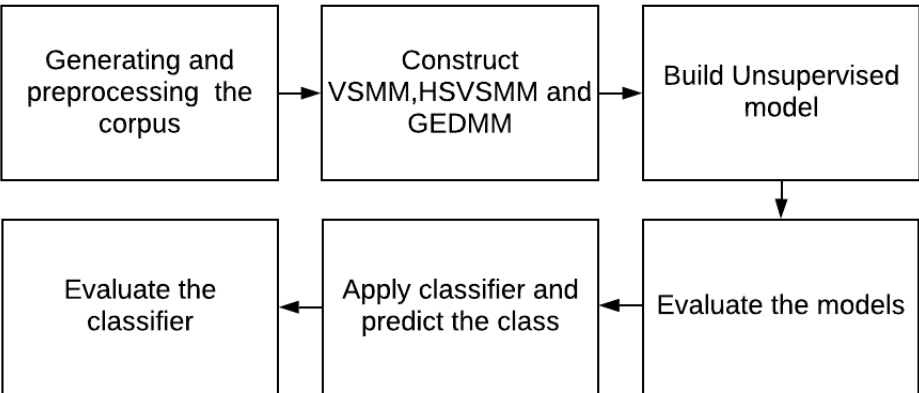
Figure 2. Entity group



Thus choosing significant entity groups acts as input to produce GEDMM. Cosine matrices are formulated to apply HAC. Dendrogram is presented along with a comparative analysis of a proposed technique. Different parameters are used for comparative analysis.

Figure 3 states the steps in research methodology. Research methodology initiates with selection of appropriate packages. Library udpipe available in R is used to process the Marathi text. (<http://cran.r-project.org/web/packages/udpipe/vignettes/udpipe-annotation.html>). ‘UTF8’ Encoding style is used to read Marathi text. The Marathi language framework/model is required. The model needs to be download (Marathi\_Model <- udpipe\_download\_model(language = “Marathi”)), load (Marathi\_

Figure 3. Steps in research methodology



model <- udpipe\_load\_model()) and call (model\_marathi<- udpipe\_download\_model(language = "Marathi")). library(pROC), library(caret) library(mlbench) etc. are used. Different plots are obtained after building model using command fit=train(type=.,data=train, method='knn',tuneLength=20,trControl=trControl,preProc= C("Center", "Scale")) etc. Following steps depict the research methodology.

1. Generating and preprocessing the corpus: Data in the form of Marathi News is downloaded from various websites (<https://www.sumanasa.com/loksatta/news>)([www.lokmat.com](http://www.lokmat.com))([www.loksatta.com](http://www.loksatta.com))

An experimental set up is established on 1215 news. The preprocessing of the corpus initiates with tokenization which results in 1,35,234 tokens. Constructing entity document matrix: Library(rJava), library(NLP), library(openNLP) are used for extraction (produces information of tokens) and classification (classifies the tokens) of entities. Extractors, their IDs and classifiers are available (<https://tdilddc.in>) for the Marathi language. Conditional Random Field (CRF) is used by the NER tagger and it titles the sequences of words. NER is applied to the corpus. The extracted entities are assigned with the relevance measures. The entities belonging to the same domain are grouped to form the entity group. The weight of each entity group is calculated. It is the sum of the relevance measures of all entities present in the group. The threshold is set to be 75%, thus the top 75% entities are chosen to formulate Group Entity document matrix Marathi (GEDMM). Total 35,189 entities are extracted. Grouping relevant entities has resulted into 29,145 groups. There are total 11, 234 significant entity groups (above threshold). These represent the attributes of GEDMM.

2. Construct VSMM and HSVSMM: Lemmatization is carried out along with special characters, numeric etc. removal. POS tagger is applied to select nouns, adverbs, adjectives. TF-IDF is applied and features are selected using a threshold. The total number of words/tokens with significant POS are 1,09,234. Total lemmas are 34,142. The significant terms chosen based on 75% threshold of TF-IDF weight are 13,221.VSMM is constructed using these 13, 221 terms. To construct HSVSMM hysynsets are formed and significant hysynsets act as attributes of HSVSMM. Total number of significant hysynsets are 11,132
3. Apply HAC and performance evaluation

Hierarchical agglomerative clustering using single linkage is applied on the three matrices that is GEDMM, VSMM and HSVSMM. The cosine measure is applied on each of the matrix to get the clusters and dendrograms are plot to show the document domain (class). Entropy and purity is calculated for each dendrogram and comparative result is presented. The most accurate clusters are formed using GEDMM.

4. Apply classifier and performance evaluation

GEDMM produces the best results and thus is used as input to random forest classifier. It is the most suitable algorithm for multi-class text data. Confusion matrix is constructed to evaluate the performance of the classification and prediction. The functions like confusionMatrix(p,train\$type), p1=predict(model,train) are used to carry out the validation steps.

## RESULTS AND DISCUSSION

The corpus is pre-processed and annotated using Marathi-ufal-ud-2.4-190531.udpipe library (udpipe\_annotate(udmodel\_Marathi)). The annotation details are specified in Table 1 for varied size corpus. It

Table 1. count of features for varied data size

Data set size	Total tokens	Removal of stop words	Total lemmas	Significant terms	Synsets
125	10750	10009	5012	4050	2250
380	40,135	30913	10,312	7189	4350
570	50,021	40989	20,112	11023	5160
1215	1,35,234	1, 09,234	34,142	13, 221	11,132

states the number of tokens obtained at each step from corpus preprocessing to the synset formation for varied data size. The records are displayed for varied data size from the range initiating from 125 to 1215. The number of synsets are increased as the dataset is incremented. For e.g. For 125 news, total existing tokens are 10,750, after removing the stop words it results into 10009 tokens. Total unique lemmas are 5,012. The significant terms that are above chosen threshold are 4050. Formation of synsets results into dimension reduction to form 2250 features.

Table 2 shows the entities retrieved by NER. These are classified into different domains like place, person name and so on. For eg. ‘महाराष्ट्र’ (Maharashtra) is identified as place. The value in the form of probability represents the relevance measure for each entity which is 0.894 for ‘महाराष्ट्र’ (Maharashtra). The function used for identifying the location/ place is `location_ann<-Maxent_Entity_Annotator(kind = “location”)`. Similarly to identify the person name `location_ann<-Maxent_Entity_Annotator(kind = “person”)` is used. CRF Extractor produces information about tokens and classifier assigns a class to the token. For Eg. , ‘कम्प्युटर’ belongs to “Computer” class.

Table 2. Detection of entity domain

#> # A tibble: 5 x 4	
#> reqcategory_id	probability label
#> 1 महाराष्ट्रराज्यातपरीक्षांनघेण्याच्यानर्णियावरराज्यसरकारठामआहे.	0.894 place,
#> 2 गुगलनेयानदीनधोरणाचीघोषणाकेलीआहे	0.649 internet
#> 3 असंस्पष्टकिरणपुनरियांनादीदलिआहे....	0.451 person
#>4 कम्प्युटरसमिुलेशनच्यासहाय्यानेयाग्रहाचेअस्तित्वजाणवले	0.240 Computer...
#> 5 एकाग्रहाचाशोधसंशोधकांनालागला	0.252 science

The entities retrieved in such a way are shown in Table 3 with their relevance measures. Total number of entities extracted are 35,189. The entities belonging to sports domain are ‘संघ’ (team) and क्रिकेट (cricket) having their relevance measures as 0.43 and 0.69 respectively.

Table 4 shows the features of GEDMM in the form of either entity group or a single entity along with the feature weight. Summing up the relevance measure of संघ and क्रिकेट (0.43+ 0.53) results in

Table 3. Entity and feature relevance measures

Entity	महाराष्ट्र,	राजस्थान	संघ	क्रिकेट	गुगल
Relevance measure	0.11	0.33	0.43	0.53	0.76

Table 4. Features of GEDMM and relevance measure

Entity group	महाराष्ट्र, राजस्थान	संघ, क्रिकेट	फळ,अन्न	..	इंजनि
Relevance measure	0.44	0.96	0.76	..	0.

0.96. So this entity group is considered while forming GEDMM. Due to the summing of, the entity group is selected which otherwise would have been ignored due to low threshold though significant. Total entities are reduced to 29,145 due to their grouping.

The significant features are used to construct GEDMM. These significant features can be in the form of entity group. The entity groups having feature weight/ relevance measure above the threshold act as attributes of GEDMM. For eg. सनिमा, अभनिता are entities from the entertainment domain which are combined to form entity group. The relevance measure of this entity group differs from news to news depending upon its significance with respect to the specific news. That is for N1 it is 0.11 whereas for N2 and N3 it is 0.56 and 0.78 respectively. In Table 5, news are placed in rows and column represents features in the form of entity groups which are 11,234 in total.

Table 5. Formation of GEDMM using significant features

News	सनिमा, अभनिता	संघ, क्रिकेट	..	आकाशगंगा
N1	0.11	0.32	..	0.09
N2	0.56	0.11	..	0
N3	0.78	0.12	..	1
..	..	..	..	..
N1215	0.89	0.21		0.11

To construct VSMM nad HVSMM, the TF-IDF measure for each token is calculated. There are total tokens of which TF-IDF weight is measured. Token and its TF-IDF weight is reflected in Table 6. For e.g. token ‘राधिका’ has a weight 0.04.

Table 6. Term and TF-IDF count for VSMM and HVSMM

Term	राधिका	ओळख	गुणवान	आगामी	व्यक्तिरिखा
TF-IDF weight	0.04	0.12	0.45	0.34	0.76

Table 7 shows feature weights calculated to decide whether the particular synset is to be included in the construction of HVSMM. There are total 12,144 synsets. The feature weight of गावकरी and लोक, is 0.32 and 0.63 respectively so it's total feature weight is 0.95.

Table 8 shows significant features used to construct HVSMM. The features above the threshold value are used to construct HVSMM. There are total 11,132 selected synset groups/features. All the features depicted as column heads of Table 8 has feature weight greater than 0.75.



Table 7. synsets and their weights for HSVSMM

token	आकाशगंगा	दंड	..	{गावकरी, लोक}
TF-IDF weight	0.02	0.05	..	0.95

Table 8. Significant synset and TF-IDF count for HSVSMM

Significant tokens	गावकरी,लोक	भारतीयहडिस्थानी	क्रिकेट, खेळ	वचिर
TF-IDF weight	0.98	0.83	0.77	0.71

In Table 9 news occupy rows and features are represented using columns. Features can be single term or a synset (group of terms) There are total features which are used to construct HSVSMM using 1215 news. ‘भारत’, ‘हडिस्थान’ formulates one synset which means ‘India’ whereas ‘भीती’ is a single term representing feature. The significance of or feature weight for भारत,हडिस्थान with respect to N1 and N2 is 0.81, 0.24 respectively. Total 11,132 features act as column heads.

Table 9. Feature vectors of HSVSMM

News	भारत हडिस्थान	पुणे महाराष्ट्र	..	भीती
N1	0.81	0.35	..	0.11
N2	0.24	0.13	..	0.56
N1215	0.51	0.19		0.11

Table 10 shows the mutual similarity between news. The most similar news have cosine measure near to 1. Cosine similarity measure value zero indicates there is no common word between that specific news. Thus Similarity between N1 and N1 is always as indicated by table’s first entry.

Table 10. Cosine similarity matrix for news using GEDMM

News	N1	N2	..	N1215
N1	1	0.34	..	0.56
N2	0.34	1	..	0.01
..	..	..	..	..
N1215	0.01		..	1

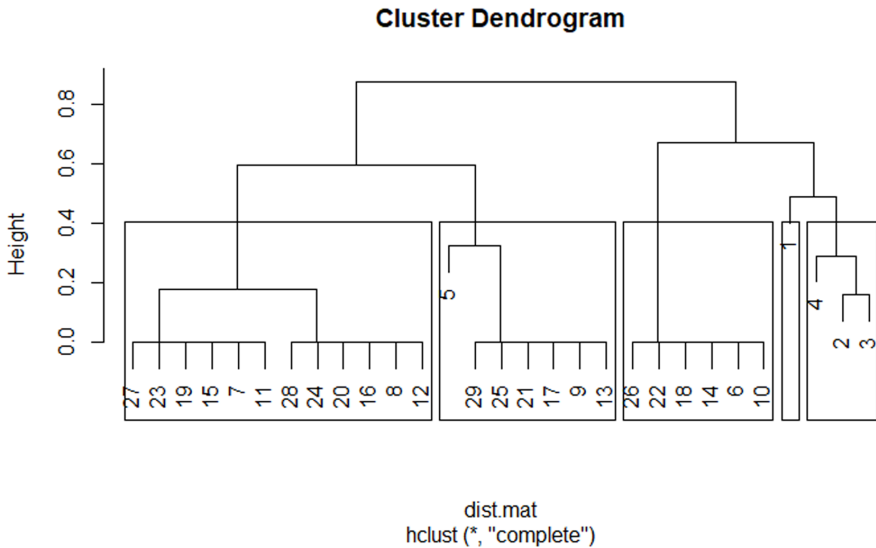
Cluster quality is validated using entropy and purity. Both of the parameters are measured for different datasets with different size. For GEDMM, Entropy is consistently improved and minimum from small to big corpus indicates less error in the same way, consistently maximum purity indicates accurately clustered news. GEDMM produces the improved clusters which can be titled as sports,

entertainment, computer, location and Person. Table 11 depicts improved entropy using GEDMM which is 0.3 better than HSVSMM (0.4) and VSMM (0.5), also improved purity (0.8) than HSVSMM (0.7) and VSMM (0.6). Figure 4 depicts the dendrogram for 29 news, representing the cosine similarity matrix implemented on GEDMM. Dendrogram is obtained after applying HAC using single linkage on cosine matrix. Five clusters represent five significant entities as ‘computer’, ‘entertainment’, ‘location’, ‘person’ and ‘sport’.

Table 11. Entropy comparison

Dataset size	VSMM		HSVSM		GEDMM	
	Entropy	Purity	Entropy	Purity	Entropy	Purity
125	0.3	0.7	0.2	0.8	0.1	0.9
380	0.3	0.7	0.2	0.8	0.1	0.9
570	0.4	0.6	0.4	0.7	0.2	0.8
1215	0.5	0.6	0.4	0.7	0.3	0.8

Figure 4. Dendrogram



Prediction of news is done using a random forest classifier. GEDMM is used as input to the training model. 70% of news is included in the training dataset and 30% are included in the testing dataset. The algorithm converges at 500 trees. Figure 5 denotes that misclassification error reduces, as trees are incremented.

Table 12 shows the confusion matrix for 972 news which is 80% of 1215 (total news). The maximum values existing at the diagonal position of the matrix confirms the correct predictions. The accuracy of the classifier is 0.83 and the misclassification error is less than 0.4. Parameters of the process/methodology are proved by getting better results. Sports related news are predicted most correctly.

Figure 5. trees produced by random forest

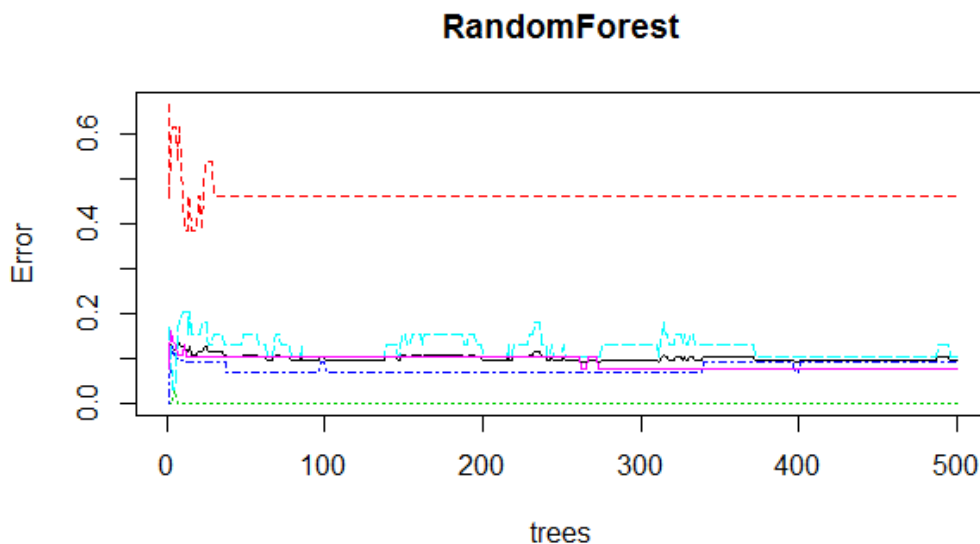


Table 12. Confusion matrix

News domains	Computer	Entertainment	Location	Person	sport	class.error	Precision	Recall
Computer	54	0	0	6	0	0.1034482	0.81	0.89
Entertainment	0	180	27	0	0	0.3000000	0.76	0.88
Location	0	0	240	8	0	0.2000000	0.78	0.78
Person	4	0	0	238	0	0.016597	0.91	0.91
Sport	0	0	0	5	221	0.021111	0.88	0.87

## CONCLUSION

Marathi news are processed to detect it's class which could be one of 'computer', 'entertainment', 'location', 'person' or 'sport'. The approach comprises of 4 phases/steps. It is implemented initiating from the extraction of entities and forming entity groups and hysynsets as a first phase followed by generating GEDMM, VSMM, HSVSMM as a second phase. HAC is applied in the third phase. The approach completes with building random forest and prediction. Clustering and classification are evaluated using several parameters like entropy, purity and respectively. Due to the grouping approach used in GEDMM and HSVSMM, dimension reduction is achieved which results into accuracy. Clustering obtained using GEDMM performs best than VSMM, HSVSMM. Thus random forest is built using GEDMM and achieves an accuracy as 0.83. The work could be extended using other techniques as well as other critical languages. 'Sanskrit' collocations can be extracted to apply entity grouping techniques.

## REFERENCES

- Bafna, P. B., & Saini, J. R. (2020). An Application of Zipf's Law for Prose and Verse Corpora Neutrality for Hindi and Marathi Languages. *International Journal of Advanced Computer Science and Applications*, 11(3). Advance online publication. doi:10.14569/IJACSA.2020.0110331
- Bafna, P. B., & Saini, J. R. (2020). Marathi Text Analysis using Unsupervised Learning and Word Cloud. *International Journal of Engineering and Advanced Technology*, 9(3).
- Aggarwal, C. C., & Zhai, C. (Eds.). (2012). *Mining text data*. Springer Science & Business Media. doi:10.1007/978-1-4614-3223-4
- Hanumanthappa, M., & Swamy, M. N. (2014). A detailed study on Indian languages text mining. *International Journal of Computer Science and Mobile Computing*, 3(11), 54–60.
- Meyer, D., Hornik, K., & Feinerer, I. (2008). Text mining infrastructure in R. *Journal of Statistical Software*, 25(5), 1–54.
- Vijaymeena, M. K., & Kavitha, K. (2016). A survey on similarity measures in text mining. *Machine Learning and Applications: An International Journal*, 3(2), 19–28.
- Larsen, B., & Aone, C. (1999, August). Fast and effective text mining using linear-time document clustering. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 16–22). doi:10.1145/312129.312186
- Gharehchopogh, F. S., & Khalifelou, Z. A. (2011, October). Analysis and evaluation of unstructured data: text mining versus natural language processing. In *2011 5th International Conference on Application of Information and Communication Technologies (AICT)* (pp. 1–4). IEEE. doi:10.1109/ICAICT.2011.6111017
- Bafna, P., Pramod, D., & Vaidya, A. (2016, March). Document clustering: TF-IDF approach. In *2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)* (pp. 61–66). IEEE doi:10.1109/ICEEOT.2016.7754750
- Deena, G., & Raja, K. (2020). Developing the Assessment Questions Automatically to Determine the Cognitive Level of the E-Learner Using NLP Techniques. *International Journal of Service Science, Management, Engineering, and Technology*, 11(2), 95–110. doi:10.4018/IJSSMET.2020040106
- Soltani Delgosha, M., Haji Heydari, N., & Saadatmanesh, H. (2020). Semantic structures of business analytics research: Applying text mining methods. *Information Research*, 25(2).
- Misra, J. (2020). *autoNLP: NLP Feature Recommendations for Text Analytics Applications*. arXiv preprint arXiv:2002.03056.
- Patil, , & Patil, , & Pawar. (2016). Issues and Challenges in Marathi Named Entity Recognition. *International Journal on Natural Language Computing*, 5, 15–30. doi:10.5121/ijnlc.2016.5102
- Patil, N., Patil, A., & Pawar, B. V. (2017, December). Hybrid Approach for Marathi Named Entity Recognition. In *Proceedings of the 14th International Conference on Natural Language Processing (ICON-2017)* (pp. 103–111). Academic Press.
- Nadeau, D., & Sekine, S. (2007). *A survey of named entity recognition and classification*. Academic Press.
- Kim, G., Lee, C., Jo, J., & Lim, H. (2020). Automatic extraction of named entities of cyber threats using a deep Bi-LSTM-CRF network. *International Journal of Machine Learning and Cybernetics*, 11(10), 2341–2355. doi:10.1007/s13042-020-01122-6
- Patel, A., Ramakrishnan, G., & Bhattacharya, P. (2009, July). Incorporating linguistic expertise using ilp for named entity recognition in data hungry indian languages. In *International Conference on Inductive Logic Programming* (pp. 178–185). Springer.
- Bafna, P. B., & Saini, J. R. (2020). Marathi Document: Similarity Measurement using Semantics-based Dimension Reduction Technique. *International Journal of Advanced Computer Science and Applications*, 11(4). Advance online publication. doi:10.14569/IJACSA.2020.0110419

- Kou, G., Yang, P., Peng, Y., Xiao, F., Chen, Y., & Alsaadi, F. E. (2020). Evaluation of feature selection methods for text classification with small datasets using multiple criteria decision-making methods. *Applied Soft Computing*, 86, 105836. doi:10.1016/j.asoc.2019.105836
- Wahdan, K. A., Hantoobi, S., Salloum, S. A., & Shaalan, K. (2020). A systematic review of text classification research based on deep learning models in Arabic language. *Iranian Journal of Electrical and Computer Engineering*, 10(6), 6629–6643.
- Dogan, T., & Uysal, A. K. (2020). A novel term weighting scheme for text classification: TF-MONO. *Journal of Informetrics*, 14(4), 101076. doi:10.1016/j.joi.2020.101076
- Shah, K., Patel, H., Sanghvi, D., & Shah, M. (2020). A comparative analysis of logistic regression, random Forest and KNN models for the text classification. *Augmented Human Research*, 5(1), 1–16. doi:10.1007/s41133-020-00032-0
- Parlak, B., & Uysal, A. K. (2020). The effects of globalisation techniques on feature selection for text classification. *Journal of Information Science*.
- Gao, X., Tan, R., & Li, G. (2020). Research on Text Mining of Material Science Based on Natural Language Processing. *MS&E*, 768(7), 072094. doi:10.1088/1757-899X/768/7/072094
- Sharma, S., Sodhi, A., Nagpal, N., Dureja, A., & Dureja, A. (2019). Practical approaches: Machine learning techniques for image processing and text processing. *International Journal of Information Systems & Management Science*, 2(2).
- Maksimenko, O. I. (2019). Automatic distributive-statistic analysis as system text processing. *RUDN Journal of Language Studies. Semiotics and Semantics*, 10(1), 92–100. doi:10.22363/2313-2299-2019-10-1-92-100
- Kubosawa, S., Tsuchida, M., & Ishikawa, K. (2019). *U.S. Patent No. 10,339,223*. Washington, DC: U.S. Patent and Trademark Office.
- Bustikova, L., Siroky, D. S., Alashri, S., & Alzahrani, S. (2020). Predicting Partisan Responsiveness: A Probabilistic Text Mining Time-Series Approach. *Political Analysis*, 28(1), 47–64. doi:10.1017/pan.2019.18
- Avasthi, S., Chauhan, R., & Acharjya, D. P. (2020). Techniques, Applications, and Issues in Mining Large-Scale Text Databases. In *Advances in Information Communication Technology and Computing* (pp. 385–396). Springer.
- Nasar, Z., Jaffry, S. W., & Malik, M. K. (2021). Named Entity Recognition and Relation Extraction: State-of-the-Art. *ACM Computing Surveys*, 54(1), 1–39. doi:10.1145/3445965

*Prafulla Bafna is an Assistant Professor in the Symbiosis Institute of Computer Studies and Research, Symbiosis International University (SIU), Pune, Maharashtra, India. She has a MPhil (Computer Sc), Yashwantrao Chavan Maharashtra Open University (YCMOU) and M. Sc. (Computer Science), Savitribai Phule Pune University. Her research interests include Data mining, Text Mining and Human computer Interaction(HCI).*