Project Report Titled

TEXT CATEGORIZATION USING TF-IDF VALUE FOR A TEXTUAL DATASET AND PREDICTING THE CLASS LABEL FOR THE DOCUMENTS

Submitted in partial fulfillment of the
requirements of the degree of
Bachelor of Technology
(Information Technology)

By

Sanket Dhabale (151080004)
Pranay Manthanwar (151080007)
Pankaj Dandewad (151080056)
Chetan Ghodam (151080057)

Under the guidance of

Prof. Mahesh Shirole

Associate Professor

Department of Computer Engineering
and Information Technology,
Veermata Jijabai Technological Institute
(Autonomous Institute Affiliated to Mumbai University)
Mumbai – 400019

(2015 – 2019)

DECLARATION OF STUDENTS

We declare that work embodied in stage-I of this Project titled "**Text Categorization Using Machine Learning Algorithm**" form our own contribution of work under the guidance of **Prof. Mahesh Shirole** at the Department of Computer Engineering and Information Technology, Veermata Jijabai Technological Institute. The report reflects the work done during period of Stage-I.

Sanket Dhabale
(151080004)

Pranay Manthanwar
(151080007)

Pankaj Dandewad
(151080056)

Chetan Ghodam
(151080057)

Approval Sheet

CERTIFICATE

This is to certify that ,

1. **Mr. Sanket Dhabale** (151080004),
2. **Mr. Pranay Manthanwar** (151080007)
3. **Mr. Pankaj Dandewad** (151080056),
4. **Mr. Chetan Ghodam** (151080057),

students of B.Tech. (Information Technology), Veermata Jijabai Technological Institute, Mumbai have successfully completed the stage-I of project titled "**Text Categorization Using Machine Learning Algorithm**" under the guidance of **Prof. Mahesh Shirole**.

Prof. Mahesh Shirole
Associate Professor
Veermata Jijabai Technological Institute

Dr. V. B. Nikam
Head of Computer Engineering and
Information Technology Department

Dr. Dhiren Patel

Director, Veermata Jijabai Technological Institute

CONTENTS

ABSTRACT:

Considering the idea of predicting and classify of documents based on various features like the document text data and other metadata associated with the documents, along with generating labels for the list of documents. So we are going to classify the documents accordingly and generate a specified class for the dataset. For that we need to explore the fields of Machine Learning and provide an appropriate data classification model for the project. Most of the dataset is text based so many algorithms would be the perfect algorithm to predict and classify the documents according to the expected and obtained output.

INTRODUCTION:

Since the age of internet many of the sports, businesses, advertisement companies, political matters, etc documents are released in different languages and it’s quite difficult to predict their documents type or category in which they fall and thus they land up at an unstable position for their expanding their agendas about the information they are providing. Thus our project comes into picture. The project mainly deals with the prediction and classification of the documents based on the text document they had provided along with the dataset. It is more convenient, reliable and also provides the user with wide range of classified and categorized data about the documents.

At first, we have to train our model to classify text data according to their label and provide the test dataset to the model. Thus error in the classification can be calculated. This allows our scope to improve the model.

We are about to use machine learning algorithm that provides appropriate string data classification with maximum accuracy.

MOTIVATION:

We can see the increase in the amount of information available online over various platforms available to us. Also we can see that the number of languages used for text categorization is present in abundance. With abundance of information in various languages, complexity in classification arises. As the text can be in English, Spanish, Italian or any regional languages like Marathi, Hindi. Thus we can’t assign a class just comparing it to another language.

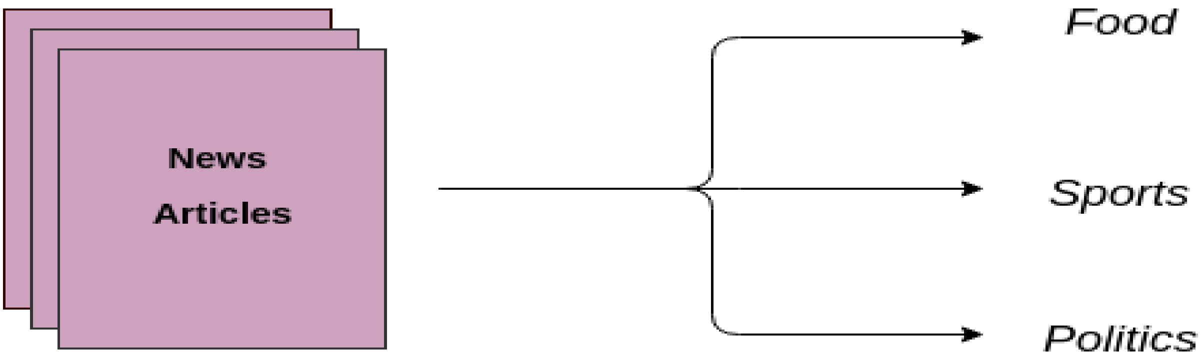
Text classification is different from binary classification. Multiple types of documents are present that is assigning a label to a document is not binary but requires much more computation. Not to cause any bias in assigning labels due to multiple classes is the main motivation behind the project.

PROBLEM STATEMENT:

Starting from the internet age, many more documents have been flooding over the internet. Lots and lots of documents are found within a minute. Thus assigning a binary class to a document is not preferred and it is not so convenient to do so. Also it is very important to classify these online documents for easy and fast access for computational purposes or finding anything on WWW. Our proposed system does the job of classifying and labeling a document its accurate class. Along with English text classification, we are also going to consider the Marathi text classification and label the documents perfectly. Two separate databases will be required for the algorithm to understand and classify the document accordingly.

LITERATURE SUMMARY:

Sentiment analysis is one of the most common use cases for classifiers. This kind of analysis is used detect positive or negative sentiment from a user or customer in their comments, tweets, reviews, etc. In language detection, an incoming piece of text will be analysed against a list of languages (e.g.: Spanish, English, French, etc.) to programmatically detect the language of the given text. Working with a set of apparel products and you want to automatically classify them using their descriptions. Text classification is often used for organize text by topic. This is commonly used for emails, support tickets, reviews, articles, etc.
We can train a topic classifier to tag what incoming texts are about, provided that we provide associations between texts and tags that a machine learning model can learn from.



A number of extra text based features can also be created which sometimes are helpful for improving text classification models. Some examples are:

Textual Feature:

- Text Length The count of characters in the documents text, including punctuation and spaces.
- Character Count The number of alphabetical characters in the documents text.
- Word Count The number of words in the documents.
- Unique Word Count The number of unique words in the documents.
- Sentence Count The number of sentences in the documents.
- Automated Readability Index, the automated readability index score, which is a measure of text readability, can be computed.
- Frequency distribution of Part of Speech Tags:

    - Noun Count
    - Verb Count
    - Adjective Count
    - Adverb Count
    - Pronoun Count

LITERATURE SUMMARY REVIEW:

These topics discuss the work done by the various authors, students and researchers in brief under the domain of classification and pre-processing the textual data.

SR.no.	Title of paper, publisher /event	Autor of publication	Problem they solved	Technology they used	Methodology used	Input provided	Output obtained	Summary of work	Future work proposed/ possible extension of work
1	Text Categorization with Support Vector Machines: Learning with Many Relevant Features	Thorsten Joachims	Classification of Text Document	Polynomial and RBF kernels	SVM	Text Document	Text classification based on predefined categories	It uses Support Vector Machine for the text classification. They are fully Automated and eliminate the need of manual parameter turing. The document may fall in multiple, one or not any of the categories. The representation of each category is treated as separate binary classification. With the information retrival system, the documents are transformed into representable format for the learning algo and classification task. Ordering of the words doesn't matter so it generates the category for the corresponding word in the set including its number of repetition. It has a very high dimensional feature sets which doesn't over-fit the feature set for classification and also generalizes the accuracy.	SVMs do not require any parameter tuning, since they can nd good parameter settings automatically. All this makes SVMs a very promising and easy-to-use method for learning text classifiers
2	Interaction of	Janez Brank	Classification	Tensorflow	SVM, differ	Text Docu	Text classificati	The method SVM and other classifier for	Expand the study to

	Feature Selection Methods and Linear Classification Models	, Marko Grobelnik	of Text Document		ent classifier ie. Naïve Bayes, Perceptron-I, Linear SVM	ment	on based on predefined categories	better accuracy. The Naïve Bayes Classifier, Perceptron-I, Linear SVM is used as the classifier. One of the classifier's weaknesses can be hindered by another. Using this concept, the classification is done. Here the feature selection is done in two ways. One is using full features set and second one is using selected features set and study them for future references for linear classification. Feature selection is based on score of the feature. Top ranked are kept while other are discarded. It has seen the SVM has outperformed perceptron- based and Naïve Bayes- based classifiers for the text categorization.	additional classifiers, linear and non-linear, and diversify the feature scoring algorithms to include those that possibly include information of feature dependencies or similar characteristics, leading to more sophisticated data modeling.
3	An Empirical Comparison of Text Categorization Methods	Ana CardosoCachopo and Arlindo Limede Oliveira	Classification of Text Document	ModApte	Latent Semantic Analysis (LSA), SVM, KNN	Messages and News group	Assigning one or more number of Text Categorization	The classification is based on the messages, newsgroups and categorized into 10 different sections. Both the inputs are used for comparing the SVM and LSA results. Pre-processing of dataset is done by removing words with length smaller than 3 or greater than 20. Removed numbers, made the upper and lower cases same. The tf-idf (term frequency – inverse document frequency) is used for computing the index	We plan to investigate if further improvements can be applied to the SVMs and k-NN LSA models. If possible, this would further enhance the superiority of these methods observed in this experiments

								term weight of a document. The LSA then lowers the dimension of the original set of vectors with the new ones which comprises mostly a generalized word or class for the document. It is seen that k-NN LSA shows more promising classification than any other used method.	.
4	Integrating Feature and Instance Selection for Text Classification	Dimitris Fragoudis,Dimitris Meretakis, Spiros LikothanaSsis	Classification of Text Document	ModApte-training-test, Newsgroup, Reuters	FIS (Feature and Instance Selection)	Text Document	Text classification based on predefined categories	Most of the time feature selection does the job of reducing the dataset. Thus the FIS algorithm pre-processes the dataset by selecting the features and instances then dataset is provided to algorithm. Naïve Bayes, TAN and LB classifier has produced better results with resultant dataset of FIS algorithm. Also provides best result compared to SVM. The results were compared between MI (Mutual Information) and FIS dataset classifier. It was observed that FIS had more promising results than MI. Naïve Bayes, TAN and LB algorithm used the data set which was the resultant dataset of the MI and FIS. Among which FIS had more accurate results.	Can be extend FIS for dealing with multiclass problems and to apply it to structured data in addition to text.
5	Pruning Training Corpus to Speedup Text Classifica	Jihong Guan, Shuigeng Zhou	Classification of Text Document	VC++ 6.0 under Windows 2000, PC with P4 1.4GHz	KNN	Text Document	Text classification based on predefined categories	The method is based on the pruning of dataset by clustering method. The classification is done by using the KNN and Linear classifier.	Improves by the factor of larger than 4, with less than 3%

	tion			CPU and 256MHz memory				They both alone are not efficiently producing results as they can by doing both. Firstly the clustering calculates the difference between the documents vector as corresponding to features selected. By treating each training class as a distinctive cluster, then using a genetic algorithm to select a subset of document features such that the difference among all clusters is maximized. The pruning method has dataset of document D. A document d also has other documents in its classified class. The features and similarities score of the document is tallied with the Class in Dataset D. Thus we are pruning the dataset For further classification based on the score of a document and selecting only the class which has the score near to the document.	degradation of micro-averaging performance. So can be put to use where unnecessary features are bulk or irrelevant.
6	Accuracy improvement of automatic text classification based on feature transformation and Multi-classifier combination	Xuexian Han Guowei Zu, Wataru Ohyama1	Classification of Text Document	ModApte-training-test, Newsgroup, Reuters.	Euclidean distance, SVM-Linear, Linear discriminant function,	Text Document	Text classification based on predefined categories	The procedure of the automatic text classification consists of four general steps for feature vector generation, dimension reduction, learning and classification. The study done in this report tells us that the use of multiple classifiers can be done for better and efficient classification of documents. The classifiers alone are not	Can be used where biased dataset comes into picture with multiple dimensionality.

								efficient enough but the working together it overcomes each- others drawback. Based on the score of the document's feature, dataset can be reduced dimensionally if the feature doesn't have the needed count. Among all the classifiers, the SVM-Linear had the best outcome with reduced dimensionality.	
7	Combining Multiple K-Nearest Neighbour Classifiers for Text Classification by Reducts	Yongguang Bao and Naohiro Ishii	Classification of Text Document	ModApte-training-test, Newsgroup, Reuters.	K-nearest Neighbour, KNN Classifier, RkNN.	Text Document	Text classification based on predefined categories	It uses basic K- nearest neighbour for the classification. Alone K-nearest neighbour is sufficient so multiple feature set has been put to use. It combines multiple KNN classifiers. To select the feature of the subset, the MFS were build on trail and error. To overcome this problem, random selection of MFS was done. This made the problem NP-hard. The multiple reducts can be formulated precisely and in a unified way within the framework of Rough Sets theory. This theory generates multiple reducts which improves the performance of KNN classifier.	Multiple reducts to improve the performance of the k-nearest neighbor classifier which is easiest classifier. So future use might be restricted.
8	Feature Selection using Improved Mutual Information for Text Classifica	Jana Novovicova , Anton on Malik and Pavel	Classification of Text Document	Reuters	Naive Bayes Classifier, Best individual features(BI	Text Document	Text classification based on predefined categories	In text classification, usually a document representation using a bag-of-words approach is employed. This representation scheme leads to very high dimensional feature space. A predefined	Many areas of future work remain. Ongoing work includes comparison on the other

	tion	Pudil			F), Sequential forward selection (SFS)			number of the best features are taken to form the best feature subset. Scoring of individual words can be performed. Best individual features (BIF) methods evaluate all the n words individually according to a given criterion, sort them and select the best k words. Sequential forward selection (SFS) methods firstly select the best single word evaluated by given criterion. Then, add one word at a time until the number of selected words reaches desired k words. r SFS methods do not result in the optimal words subset but they take note of dependencies between words as opposed to the BIF methods. Therefore SFS often give better results than BIF.	text classifiers, for example, support vector machines and k-nearest neighbor.
9	Discretizing Continuous Attributes in AdaBoost for Text Categorization	Pio Nardiello, Fabrizio Sebastiani, and Alessandro Sperduti	Classification of Text Document	Reuters and newsgroup	AdaBoost	Text Document	Text classification based on predefined categories	Based on the idea of adaptive boosting, a version of boosting in which members of the committee can be sequentially generated after learning from the classification mistakes of previously generated members of the same committee, AdaBoost.MH is a realization of the well-known AdaBoost algorithm, which is specifically aimed at multi-label TC4, and which uses decision trees composed of a root and two leaves	AdaBoost.MH is in the restricted lot of the peak text categorization performers nowadays, a lot where the margins for performance improvement are slimmer and slimmer.

								only as weak hypotheses. Algorithms attempt to optimally split the interval on which these attributes range into a sequence of disjoint subintervals. This split engenders a new vector (binary) representation for documents, in which a binary term indicates that the original non-binary weight belongs or does not belong to a given sub-interval	
10	A comparative study on feature selection in text categorization	Yaming Yang, Jan O Pederson	Classification of Text Document	Reuters, OHSUMED	KNN, Linear Least Square Fit (LLSF), Document frequency, Information gain, Chi-test	Text Document	Text classification based on predefined categories	Document Frequency (DF) Threshold is the simplest for vocabulary reduction. Easily scales to very large corpus. Due to widely received assumption of info retrival, DF is not used. The Information Gain (IG) measures the bit of information and obtains the category of the document by the presence or absence of terms. With each term Information gain is calculated and few are discarded which has less value than already predefined threshold. Thus conditional probability is put to use for term *t* and category *c*. The Chi-test measures the lack of independence between the term t and category c and can be compared to Chi square distribution with one degree of freedom. Thus the IG or DF combined with KNN or LLSF gives us efficient	Eases the computation and power over the application used for high level performance. From Neural Network to Text categorization The methods can be used significantly.

								results for the classification.	
11	“Text Categorization with Support Vector Machines.”	Machine Learning, 46, 423–444, 2002 c ,2002 Kluwer Academic Publishers. Manufactured in The Netherlands	Classification of text document	linear kernel, 2nd order polynomial kernel, Gaussian rbf-kernel	SVM	Text document	text classification, lemmatization, stemming	In this we study about (SVM) support vector machines . The SVM are capable of effectively processing feature vectors of some 10 000 dimensions, given that these are sparse. And also support vector machines provide a fast and effective means for learning text classifier’s from examples we study different mappings of frequencies to input space, and combine these mappings with different kernel functions	In future work we want to see if the results can be generalized to other languages i.e. Slavic, romance, and non-Indo-Europeans. If the results were positive, a generic algorithm would be found that worked well on nearly any language.
12	1. Text categorization based on Concept indexing and principal component analysis.	Ke H., Shaoping M 2002	They find that this algorithm can effectively reduce dimensionality without sacrificing categorization accuracy.	salton	Concept indexing, principle component analysis, Vsm,KNN,Baysean classifier.	Text document	Classified data on	They uses the vector space model and feature selection of the text document is represented by a vector and all subsequent calculation based, many ML technology have been successfully applied to text categorization. Concept indexing is simple and effective way to reduce dimension. For effective in data compression and feature extraction we use PCA,they applied pca to ci subspace.	This method for put forwarded in the paper is meaningfull to online text categorisation, application of more machine learning.
13	“Improving SVM Text Classification Performance through Threshold Adjustment”	Clairvoyance Corporation, 5001 Baum Boulevard, Suite 700, Pittsbu	Classification of text document	Corpora, threshold adjusting algorithm	SVM	Text document	automatic process for adjusting the thresholds of generic SVM which incorporates a user utility	In general, support vector machines (SVM), when applied to text classification provide excellent precision, but poor recall. So to improve Recall we customizing SVMs. Customizing Means to adjust the threshold	the proposed thresholding approach is independent of the learnt model, using it in conjunction with other types of models will

		rgh, PA 15213-1854, USA					model, an integral part of an information management system	associated with an SVM. We describe an automatic process for adjusting the thresholds of generic SVM which incorporates a user utility model, an integral part of an information management system	also form an interesting aspect of future work.
14	"Feature Selection Algorithms to Improve Documents Classification Performance"	Pedro A. C. Sousa 1, João Paulo Pimentão1, Bruno René D. Santos 2, and Fernando Moura -Pires3	Classification of text document	Tensorflow	multi-agents systems, feature selection, Information retrieval , text learning	huge network infrastructures and new information, text document	Improve Document	In this we use the feature selection algorithms were evaluated in order to improve documents' classification performance	for improving documents' classification performance.
15	"An evaluation of statistical approaches to text categorization."	Yiming Yang yiming @cs.cmu.edu April 10, 1997	Classification of text document	Corpus, categorization methods	KNN, LLSF ,neural network and WORD, cross method evaluation	Text documents , previously published results and newly obtained results	Improve Text	This paper is a comparative study of text categorization methods. Fourteen methods are investigated, based on previously published results and newly obtained results from additional experiments. Corpus biases in commonly used document collections are examined using the performance of three classifiers. Problems in previously published experiments are analyzed, and the results of flawed experiments are excluded from the cross-method evaluation. As a result, eleven out of the fourteen methods are remained. A k-nearest neighbor (kNN) classifier was chosen for the performance baseline on several collections; on each collection, the performance scores of other methods were	for improving documents' classification performance.

								normalized using the score of kNN. This provides a common basis for a global observation on methods whose results are only available on individual collections. Windrow-Hoff, k-nearest neighbour, neural networks and the Linear Least Squares Fit mapping are the top-performing classifiers, while the Roccio approaches had relatively poor results compared to the other learning methods. KNN is the only learning method that has scaled to the full domain of MEDLINE categories, showing a graceful behaviour when the target space grows from the level of one hundred categories to a level of tens of thousands An Evaluation of Statistical Approaches to Text	
16	Text categorization based on Concept indexing and principal component analysis.	Ke H., Shaoping M 2002	They find that this algorithm can effectively reduce dimensionality without sacrificing categorization accuracy.	salton	Concept indexing, principle component analysis, Vsm,KNN,Baysean classifier.	Text document	Classified data on	They uses the vector space model and feature selection of the text document is represented by a vector and all subsequent calculation based, many ML technology have been successfully applied to text categorization. Concept indexing is simple and effective way to reduce dimension. For effective in data compression and feature extraction we use PCA,they applied pca to ci subspace.	This method for put forwarded in the paper is meaningfull to online text categorisation, application of more machine learning.
17	A Comparison of Word- and Sense-	Kehagias A., Petridis V., Kaburlasos	: (a) in comparing the merit of words and	Wordnet lexical	MAP, ML,verson space, KNN, Recur	Lexical database	Classified data	They work with WordNet lexical database and distinction between the word and senses. It contains the large number of noun ,verb etc of English language	Nevertheless, in a practical classification task the senses would have to be obtained by a

	based Text Categorization Using Several Classification Algorithms.	V., Fragkou P 2003	senses as classification features and (b) in testing several classification algorithms on the Brown Corpus		sive Version of the MAP algorithm, Maximum Likelihood (ML) Classification			.WordNet provide carefully worked out word and sense vocabularies for English language, as well as the membership of each word into a number of senses.the document they have used in their text categorisation experiment use a subset of the brown corpus .for document representation they used 4 document representation two are word based and two are sense based. And classify algorithm uses are Maximum a posteriori (MAP) classification, batch version, recursive version of MAP algorithm, maximum Likelihood classification and FLNMAP with voting.	disambiguation step which, in all probability, would introduce a significant error
17	Automatic detection of text genre.	B. Kessler, G. Nunberg, and H. Schutze. 1997	They propose a theory of genres as bundles of facets, which correlate with various surface cues, and argue that genre detection based on surface cues is as successful as	Computational linguists	Corpus logistic Regression , Neural Network,	Text data	Classified data on basis of linguistics.	They first linguistic research on genre that uses quantitative method then identify the genres : genetic cues , these cues that have figured prominently in previously work on genre.then applied method like corpus , logistic Regression , Neural Network. For each genre facet ,it compare our result using surface cues .	This theory used in application of genre classification to tagging, summarization.

			detection based on deeper structural properties.						
19	Techniques for Improving the Performance of Naive Bayes for Text Classification	Karl-Michael Schneider 2002	In this they demonstrate that simple modification are able to improve the performance of Naïve Bayes for text classification significantly.	Search engine, web kernel	Rule induction, Naïve bays , decision tree,support vectr machine, clustering	Text document	Classified clustered document	Here they used text document data then then classifying these data by the help of naïve bays classifier, in these Bayesian text classification uses a parametric mixture model to model the generation of document.to make the estimation of parameters tractable , we make the Naïve Bayes assumption that the basic units are distributed independently. For the highly classification accuracy than binary independence model on text document because it model word occurrence frequency one can see that for longer document the classification scores dominated by the word probabilities and the probabilities hardly affect the classification. Feature selection is commonly regarded as a nessarry step in text classification. By taking logarithms and dividing by the length of a document, instead of multiplying conditional probabilities they calculate their geometric mean and thus account for the impact of wrong independence assumptions under varying document lengths. Furthermore, by adding the entropy of (the probability distribution induced by) the document, we account for	The main contribution of this paper is our novel feature scoring function, which is able to distinguish features that improve the clustering of the training documents (and thus are useful for classification) from features that degrade the clustering quality (and thus should be removed)

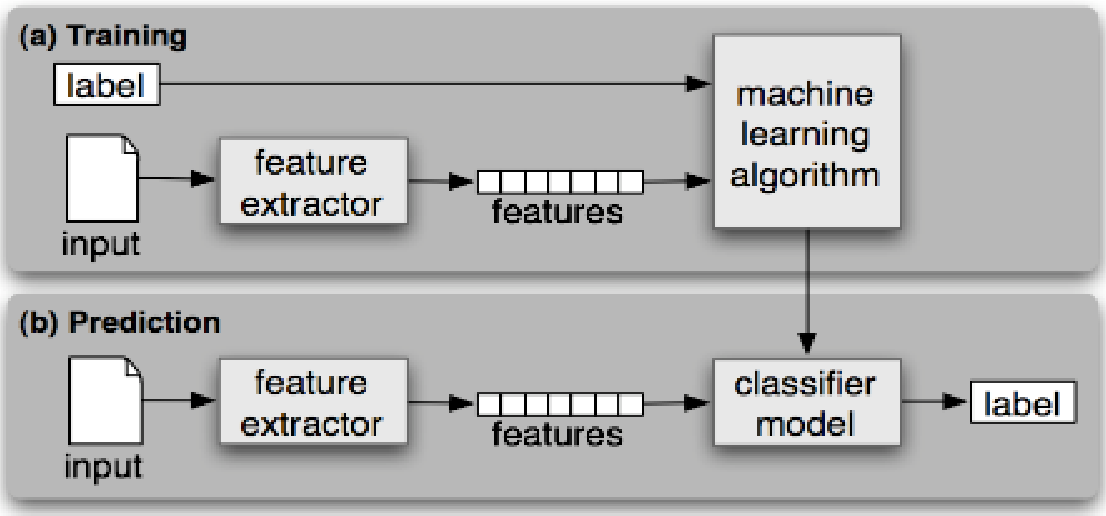
								varying document complexities.	
20	Very Large Bayesian Networks in Text Classification	Klopotek M. and Woch M.	. The paper presents results of empirical evaluation of a Bayesian multinet classifier based on a new method of learning very large tree-like Bayesian networks	Natural language possessing task.	ETC algorithm, naïve bays classifier, e Chow /Liu algorithm	Large data like search engine , language text, petent databases.	Classification of input large data.	In these work they used ETC described in details , it constructs a tree-like Bayesian network but contrary to the Chow/Liu algorithm it does not need to compare all variables with each other so that it saves much calculations of so-called DEP-measure. They estimate also the fitness of ETC to the data bye determining the log likelihood for the artificial test and test data. The goal was to check the quality of the structure of a Bayesian network obtained using ETC algorithm for various DEP functions. Then they compared ETC based multi-net classifier accuracy with Naive Bayes accuracy (NB). On the one hand, though NB is not a particularly good one, it scales quite well for tasks with dozens of thousands of attributes, ETC exhibits a bit higher stability than NB. Standard error values are usually slightly lower than those for NB classifier, though the differences are not striking. It turns out that in spite of the possibility of generation of different trees in case of different sequences of variables the quality of the Bayesian networks obtained is similar they also investigated the complexity of ETC is nlog(N) .then they reduce the ETC complexity, the popular words should be removed from the dictionary. But in some	empirical evaluation of a Bayesian multinet classifier based on a new method of learning very large tree-like Bayesian network

								cases this may deteriorate the accuracy of the classification.	

PROPOSED SYSTEM:

Exiting solution proved to be a good binary classifiers but the text classification is not the binary. The documents contain labels which are not binary in nature. A document might belong to more than one class thus we have to generate algorithm for the purpose of classifying the documents accurately. The solution is divided into two modules. One module handles the English literature part of the document classification. Another module uses Devnagari script for classifying the Marathi documents. English language and one Regional language, in this case MARATHI language, can be used for implementing the text document classification.

Proposed Model:

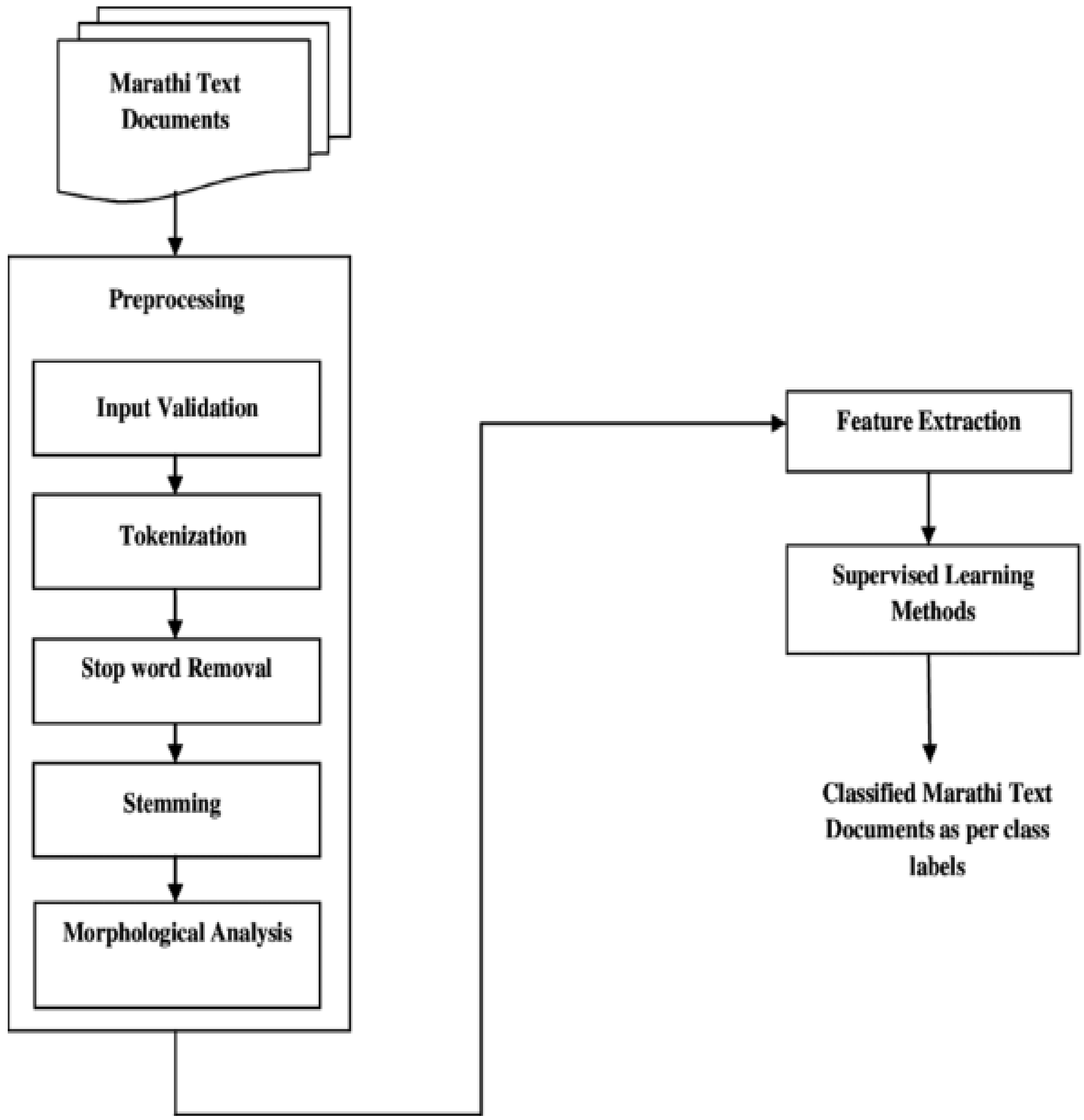


1. **Dataset Preparation:** The first step is the Dataset Preparation step which includes the process of loading a dataset and performing basic pre-processing. The dataset is then splitted into train and validation sets.

2. **Feature Selection:** The next step is the Feature Selection in which the raw dataset is transformed into flat features which can be used in a machine learning model. This step also includes the process of creating new features from the existing data.

3. **Model Training:** The final step is the Model Building step in which a machine learning model is trained on a labelled dataset.

Proposed Model:



Proposed Architecture

<u>STEPS INVOLVED IN PREPROCESSING:</u>

- **Tokenize multi-line comments into single sentences**
    - Make a single sentence of a document
    - Store it in new document
- **Tokenize each sentence into wordS**

  - Generates the token for each and every words
- **Remove stop words** in the tokenized sentence
  - For a selective language, remove all the stop words
- **Morphological Analysis** :
  - Aim to recognize the inner structure of the word
  - Morphological analyzer is expected to produce root words for a given input document
  - Root and stem of word may differ in their forms
- **Lemmatize the words** in the tokenized sentence
  - Put the tokenized value instead of word in a sentence to reduce memory consumption
- **Generate the feature set**
  - All the remaining words will be used for generating the feature set for the classifier model

- ***Random Forest***:

Random forest builds multiple decision trees and merges them together to get a more and stable prediction. Random Forest also doesn’t over-fit the dataset. With variant decision trees, we can obtain different results. The random forest takes the average of the outputs and generalizes the single output with highest number of votes or count.

One of the main feature of random forest is, that it can be used for both classification and regression problems, which form the majority of current machine learning systems. Therefore, in Random Forest, only a random subset of the features is taken into consideration by the algorithm for splitting a node. You can even make trees more random, by additionally using random thresholds for each feature rather than searching for the best possible thresholds (like a normal decision tree does).

Advantages of Random Forest Algorithm:

- The same random forest algorithm or the random forest classifier can use for both classification and the regression task.
- Random forest classifier will handle the missing values.
- When we have more trees in the forest, random forest classifier won’t over-fit the model.
- Can model the random forest classifier for categorical values also.

APPENDIX:

EXPERIMENT:

Input: Text documents

Methodology: Support Vector Machine

Output: Category into which Text documents falls

Summary:

- Fully Automated and eliminate the need of manual parameter turing since they can find good parameter settings automatically
- Document may fall in multiple, one or more categories\
- Each category is treated as separate binary classification
- Max amount of data pre-processing is done for the learning algorithm SVM
- High dimensionality so doesn't over-fit the data and generalizes accuracy
- Eliminates the need of feature selection, thus avoid high computation overload of text categorization.
- Promising and easy to use algorithm for text categorization

REFERENCES:

- Text Categorization with Support Vector Machines: Learning with Many Relevant Features - Thorsten Joachims
- Interaction of Feature Selection Methods and Linear Classification Models - Janez Brank, Marko Grobelnik
- An Empirical Comparison of Text Categorization Methods- Ana CardosoCachopo and Arlindo Limede Oliveira
- Integrating Feature and Instance Selection for Text Classification- Dimitris Fragoudis,Dimitris Meretakis, Spiros LikothanaSsis
- Pruning Training Corpus to Speedup Text Classification- Jihong Guan, Shuigeng Zhou
- Accuracy improvement of automatic text classification based on feature transformation and Multi-classifier combination - Xuexian Han Guowei Zu, Wataru Ohyama1
- Combining Multiple K-Nearest Neighbour Classifiers for Text Classification by Reducts- Yongguang Bao and Naohiro Ishii
- Feature Selection using Improved Mutual Information for Text Classification - Jana Novovicova , Antonon Malik and Pavel Pudil
- Discretizing Continuous Attributes in AdaBoost for Text Categorization-Pio Nardiello, Fabrizio Sebastiani, and Alessandro Sperduti
- A comparative study on feature selection in text categorization-Yaming Yang, Jan O Pederson
- Text Categorization with Support Vector Machines.- Kluwer Academic Publishers
- Text categorization based on Concept indexing and principal component analysis- Ke H., Shaoping M
- Improving SVM Text Classification Performance through Threshold Adjustment-Clairvoyance Corporation,Baum Boulevard, Suite
- Feature Selection Algorithms to Improve Documents Classification-Performance-Pedro A. C. Sousa, Paulo Pimentão, Bruno René D. Santos, Fernando Moura-Pires3.

- An evaluation of statistical approaches to text categorization- Yiming Yang
- Text categorization based on Concept indexing and principal component analysis- Ke H., Shaoping M
- A Comparison of Word- and Sense-based Text Categorization Using Several Classification Algorithms.- Kehagias A., Petridis V., Kaburlasos V., Fragkou P
- Automatic detection of text genre.- B. Kessler, G. Nunberg, and H. Schutze.
- Techniques for Improving the Performance of Naive Bayes for Text Classification-Karl-Michael Schneider
- Very Large Bayesian Networks in Text Classification-Klopotek M. and Woch M.