Mining and Polarity Prediction of Product Reviews

Jennifer Sarah Julius and K.M.M.Rajashekharaiah, School of Computer Science & Engineering, KLE Technological University, *jennifersarahjulius@gmail.com,kmmr@bvb.edu*

Abstract—**e-commerce applications are growing day by day so analyzing the reviews posted by the customers regarding products and making informed decision is very important. Sentiment analysis comes into picture. Sentiment analysis is a task in which you identify the polarity of given text using text processing and classification.The goal is to develop a classifier that performs sentiment analysis, by labeling the customers reviews to positive, negative or neutral . From which it is easy to classify text into classes of interest. In this paper, various classifiers to sentiment analysis such as multinomial naive bayes, decision tree, svm and K-nearest neighbour are used. Use of particular algorithms depends on the type of input given.Analyzing and understanding when to use which algorithm is an important aspect and can help in improving accuracy of results.**

Index Terms— **sentiment analysis,classification algorithms,reviews,text mining**

I.INTRODUCTION

With the increasing popularity of e-commerce applications, every day a huge amount of reviews posted in review block are made available online. This review block helps the other customers to make choices regarding the products to understand which products are worth taking and which are not.It also allows the developer of the product to view customers reviews. Therefore, online reviews can be very valuable, as collectively such reviews reflect the "wisdom of crowds" and can be a good indicator of the product's future sales performance.The classification of product reviews into positive,negative or neutral does not actually convey the emotions of customers.If the review is positive,it's a gain to product otherwise it's a loss to the product.For a popular product, the number of reviews can be in hundreds or even thousands.The opinions which are given in the form of reviews are increasing in such a way that it exceeds the analyzing capacity of an individual to make informed decision regarding product[1].In order to solve this problem of decision making regarding opinions,sentimental analysis plays an important role.

A. Sentiment analysis

Sentiment analysis is a Natural Language Processing and Information Extraction task that aims to obtain writer's feelings expressed in positive or negative comments, questions and requests, by analyzing a large numbers of documents[2]. In other words,sentiment analysis is detection of attitudes or its a task in which you identify the polarity of given text using text processing and classification.Sentiment analysis has other names such as opinion extraction,opinion mining,sentiment mining,subjective analysis.

Steps involved in sentiment analysis:

- Text extraction**:**This step involves extracting words from text that influence the outcome of the result.
- Text Refinement:This step involves refining text that involves stemming,stop word removal,tokenization.
- Text Classification**:**This step includes classification of text into its class(positive/negative)[3].

Since sentiment analysis mainly depends on orientation of words,determining semantic orientation of words Hatzivassiloglou and McKeown hypothesize that adjectives separated by "and" have the same polarity, while those separated by "but" have opposite polarity[4].It is also required to consider different kinds of reviews that are posted.

Examples:

- Rules of opinions: Apart from sentiment words and phrases, there are also many other expressions or language compositions that can be used to express or imply sentiments and opinions[1].
- Sentiment shifters: These are expressions that are used to change the sentiment orientations, e.g., from positive to negative or vice versa. Negation words are the most important class of sentiment shifters. For example, the sentence "I don't like this camera" is negative. There are also several other types of sentiment shifters. Such shifters also need to be handled with care because not all occurrences of such words mean sentiment changes. For example, "not" in "not only … but also" does not change sentiment orientation[1].

II.BACKGROUND OF INFORMATION RETRIEVAL IN PRODUCT REVIEW CONTEXT

Information retrieval can be used in many areas of computer science such as media search, search engines, etc.. Information can be retrieved through huge collection of database. Hence, Information Retrieval is basically concerned with the searching and retrieving of knowledge based

information from database. It is the process by which a collection of data is represented , stored and searched for the purpose of knowledge discovery as a response to a user query .When the information is to be retrieved i.e. for example whenever the user makes a query for particular search ,for the search to be more efficient the query should be represented or converted to the same format as that of documents stored in database.

And the query initially can be represented in the form of different types of data i.e. structured, unstructured and semi structured. Structured data refers to any data that resides in a fixed field within a record or file.This includes data that resides in relational database or spread sheets. Structured data first depends on creating a data model ,a model of the types of business data that will be recorded and how they will be stored, processed and accessed .This includes defining what fields of data will be stored and how data that will be stored .Structured data has the advantage of being easily entered, stored, queried and analyzed. Next one is semi structured data, it is the information that does not reside in a relational database but has some organizational properties that make it easier to analyze.It falls in between structured and unstructured data .It has a information associated with it such as meta data tagging that allows element contained to be addressed. Example of semi-structured data are XML ,other mark up languages,email etc.The last one is unstructured data, which refers to information that either does not have a predefined data model or is not organized in predefined manner.Unstructured data my include documents, texts, images, files, videos or web page or word processor document.

Finally we are more concerned about unstructured data rather than other types of data.The user gives a query for searching useful information and that query is an unstructured data.Lets understand what steps are carried out when user gives a query which is unstructured data to get an useful information.Initially, user queries the search engine. Stopwords are removed from the query.The search engine checks for words with similar meaning in WordNet. A separate dictionary is built for words not present in WordNet. The keywords that match with zones,retrieve the query to user.In zone based indexing, exact words are matched. In case exact result is not found the concept of n-ary tree is integrated with zone based indexing[7].Taking this as a basic concept ,we can apply this to our current topic product review meaningful information extraction. Extracting meaningful information can use many techniques like retrieving, searching, filtering, classifying and finally summarizing the useful information.

III.SYSTEM ARCHITECTURE

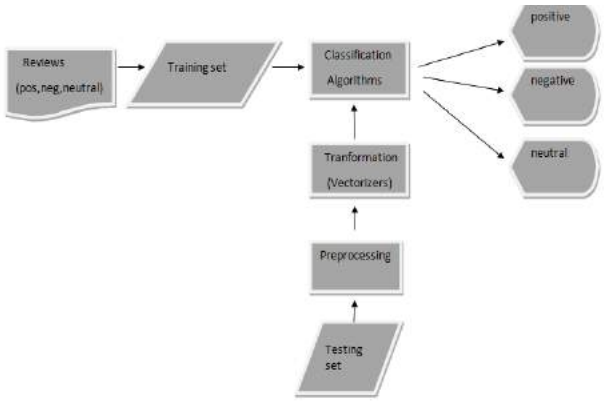

Fig. 1. System architecture for sentiment analysis

The above structure for sentiment analysis has following steps

A. Text preprocessing

When data is given as input it is necessary to preprocess the data.Text preprocessing is the process of preparing and cleaning the data of dataset for classification.It helps to reduce the noise in the text,improve the performance of the classifier and speed up the classification process.Preprocessing data has following 3 steps

- Tokenization**:**It is a kind of pre-processing where running text is segmented into words or sentences. Before any real text processing is to be done, text needs to be segmented into linguistic units such as words, punctuation, numbers, alpha-numerics, etc.This process is called tokenization.
- Stop word removal:In computing, stop words are words which are filtered out before or after processing of natural language data (text).Stop words usually refer to the most common words in a language. Some of the most frequently used stop words for English include "a", "of", "the", "I", "it", and “and” these are generally regarded as 'functional words' which do not carry meaning. When assessing the contents of natural language, the meaning can be conveyed more clearly by ignoring the functional words .Hence it is necessary to remove those words which appear too frequently that provide no information for the task[1].
- Stemming: It is the process for reducing derived words to their stem, or root form i.e.

it mainly removes various suffixes as result in the reduction of number of words, to have exactly matching stems, to minimize storage requirement and maximize the efficiency of IR Model[5].

Once the input(review text in our context) is segmented into words and then stop words are removed and stemming is carried out,the output of this phase is given as input to the transformation phase.

B. Transformation

It is the extraction of features in a format supported by machine learning algorithms from datasets.Following are the machine learning techniques that has to be applied before the text is sent for classification.

- TF-IDF**:**In information retrieval or text mining,the term frequency–inverse document frequency (also called tf-idf), is a well known method to evaluate how important is a word in a document. tf-idf is a very interesting way to convert the textual representation of information into sparse features.The weight of each word in the corpus is calculated with the help of TF-IDF.TF-IDF calculates values for each word in a document defined as below – wd = fw, d* log(|D| fw,D), w represent words, D is collection of documents, d is individual document belongs to D,|D| is size of corpus, fw,d-is number of times w appears in d, fw, -D is number of documents in which w occurs in D[1].
- Count vectorizer**:**It turns a collection of text documents into numerical feature vectors
- Hash vectorizer**:**In machine learning, feature hashing, also known as the hashing trick , is a fast and space-efficient way of vectorizing features, i.e. turning arbitrary features into indices in a vector or matrix. It works by applying a hash function to the features and using their hash values as indices directly.

Hence the role of these vectorizers is to convert the segmented text into their format and help knowing how many times the meaningful word has been repeated when test set is compared with training set. When the test set is given as input, it is compared with the training set and relevant words required for classification are identified.

C. Classification

The ouput of transformation phase is given as input to the classification phase.In this phase whatever the relevant words are obtained in previous phase are passed to this phase and polarity of those words are predicted by the classification algorithms.Classification is a level in sentiment analysis that can described as a process in which we predict qualitative response, or in this case we classify the document into its polarity. Predicting a qualitative response of a document can be referred to as classifying to the class[3]. There are many possible classification techniques that can be used to predict the polarity of document that is product reviews in our context. In this paper,following classification techniques are used.

- Decision Tree:Decision tree induction is the learning of decision tree classifiers constructing tree structure where each internal node (no leaf node) denotes attribute test. Each branch represents test outcome and each external node (leaf node) denotes class prediction. At every node, the algorithm selects best partition data attribute to individual classes. Decision trees are one of the most widely used machine learning algorithms. They are popular because they can be adapted to almost any type of data. They are a supervised machine learning algorithm that divides its training data into smaller and smaller parts in order to identify patterns that can be used for classification. Decision trees are built using a heuristic called recursive partitioning. In the training phase the algorithm learns what decisions have to be made in order to split the labelled training data into its respective classes.Whenever an unknown label is given, inorder to classify it, the data is passed through the tree. At each decision node a specific feature(i.e. opinion word in our context) from the input data is compared with a constant that was identified in the training phase.The decision will be based on whether the feature is greater than or less than the constant, creating a two way split in the tree.The data will eventually pass through these decision nodes until it reaches a leaf node which represents its assigned class.

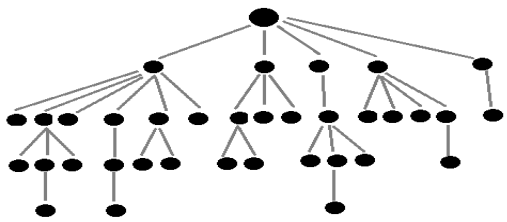

Fig. 2. Decision tree

- Support Vector Machine:Support vector machine can be referred to as supervised machine learning algorithm.It can be used for classification and regression problems. Important property of SVM is that their ability to learn can be independent of dimensionality of feature space.The goal of SVM is to design a hyperplane that classifies all training vectors into two classes.The SVM's algorithm first starts learning from data that has been classified already, which is represented in numerical labels with each number representing a category. SVM then groups the data with the same label in each convex hull. From there, it determines where the hyperplane is by calculating the closest points between the convex hulls.Once SVM determines the points that are closest to each other, it calculates the hyperplane, which is a plane that separates the labels.SVM learns to assign a label to the text and, it classifies text as positive and negative[2]

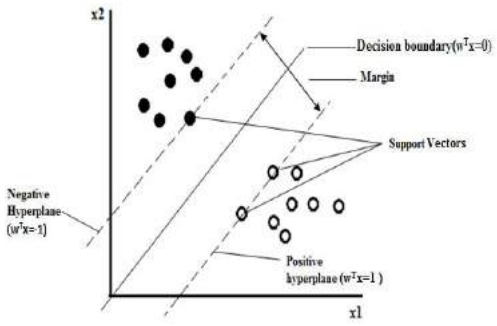

Fig. 3.SVM hyperplane

SVM Benefits-

High Dimension Input Space - while text classification we have to deal with many features (may be more then 1000).Since SVM uses over fitting protection, which does not depend on number of features so they have ability to handle large number of features[6].

Document Vector Space - despite the high dimensionality of the representation, each of the document vectors contain only a few non-zero element.More Text Categorization problems are linearly separable[6].

- Nearest Neighbour:KNN is another supervised learning algorithm, and it is a lazy learner. It is called lazy algorithm because it doesn't learn a discriminative function from the training data but memorizes the training dataset instead i.e it is the simplest classifier mainly depends on the category labels. KNN uses the parameter "k", in classifying the object. The main concept behind the KNN is it trains the system for existing data and find a predefined number of training sample nearest in distance to the new point and estimate the labels from these. For example when we try evaluate the sentiment of an unknown review,the algorithm utilizes the metric to identify the k most similar reviews in the training data set.It then tabulates majority sentiment and assigns that sentiment to the review given as testing data.

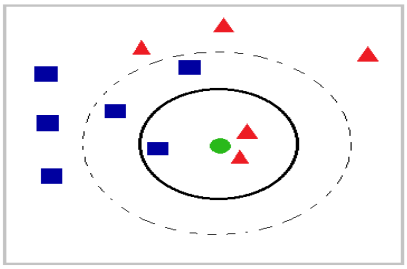

Fig. 4.Example of KNN

In the above figure,the testing data i.e. green circle should be classified to either first class of red triangles or second class of blue classes.If k=3 that is considering solid circle,green circle would be assigned to first class because there are two triangles and one blue square inside the inner circle.

- Multinomial Naive bayes algorithm

  Algorithm:

  Dictionary generation:Count occurrence of all the words in whole data set and make dictionary of some most frequent words

  Feature set generation:It compares all the words in testing set with that training set and calculates the following

$$\hat{P}(c) = \frac{N_c}{N}$$

Conditional Probabilities:

$$\hat{P}(w \mid c) = \frac{count(w,c)+1}{count(c)+|V|}$$

IV.RESULTS AND DISCUSSIONS

This section presents the experimental results on the performance of proposed algorithms. Initially,100 reviews are trained as positive,negative and neutral in a document.Then anaconda prompt is opened.Following commands are executed.

```
>python
>exit()
>cd Desktop
>cd C:/Users/Dell/Desktop/reviewAnalysis
>python reviewUI.py
```

Following GUI is displayed.

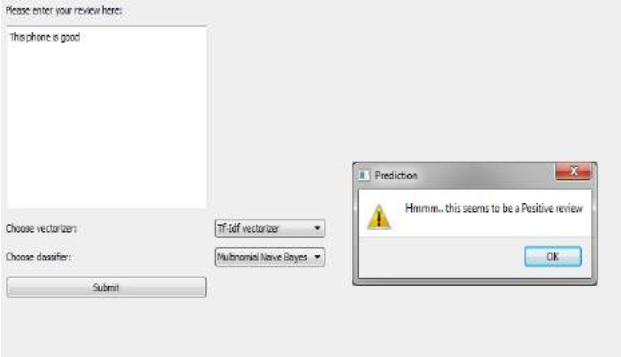

Fig.5.GUI of sentiment analysis

It can be seen from the fig that GUI contains review box in which the review is entered.Once the review is entered,vectorizers(tf-idf,count,hashing)and classifiers(multinomial,decision-tree,SVM,nearest neighbour) of your choice in combo box are choosen. and submit button is pressed and prediction is done based on review entered.

In the above GUI,the review is entered i.e. the phone is good,now tf-idf vectorizer and multinomial naïve bayes classifier is selected and submit button is pressed.The figure 3 shows that the review is correctly predicted as positive.

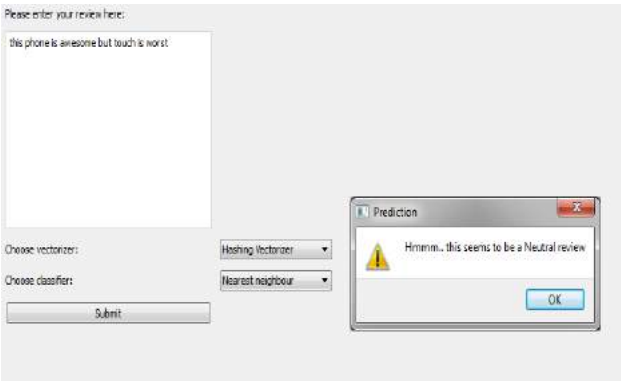

Fig.6.shows the review correctly predicted as neutral when hashing vectorizer is used with nearest neighbour classifier

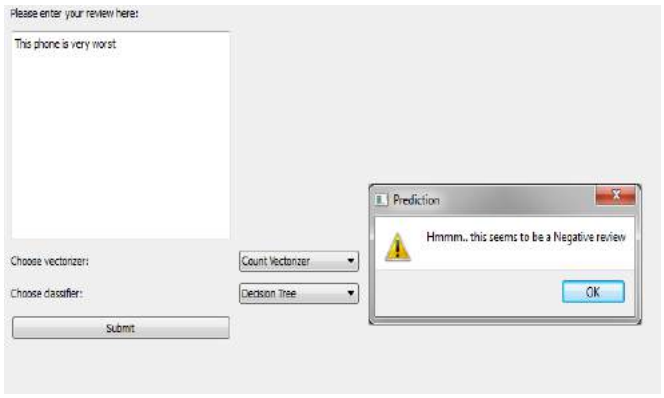

Fig.7. shows the review correctly predicted as negative when count vectorizer is used with decision tree

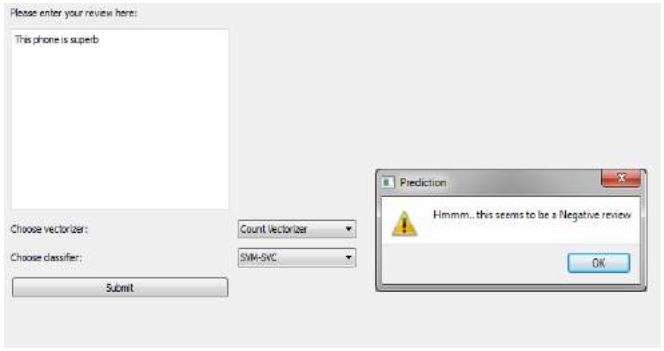

Fig.8. shows the review wrongly predicted as negative irrespective of which vectorizer is used with svm-svc because svm works with huge set of data

Similarly the same procedure is repeated for all combinations of vectorizers and classifiers to predict the review entered.

A.Confusion matrix

A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known.

n=165	**Predicted: NO**	**Predicted: YES**
Actual: NO	50	10
Actual: YES	5	100

Fig.9.shows confusion matrix

B. Quality measures

- Precision: Precision (also called positive predictive value) is the fraction of retrieved instances that are relevant

Precision=tp/(tp+fp)

Where tp: no of true positives, fp: no of false positives

- Recall: Recall (also known as sensitivity) is the fraction of relevant instances that are retrieved.

Recall = tp/(tp+fn)

Where fn: no of false negatives

- F-score: It is the measure of test's accuracy

$$\text{F-score} = \frac{2*\text{precision}*\text{recall}}{\text{Precision} + \text{Recall}}$$

- Support: It is the no of occurrences of each class

Now again anaconda command prompt is opened and following commands are executed

>python main.py

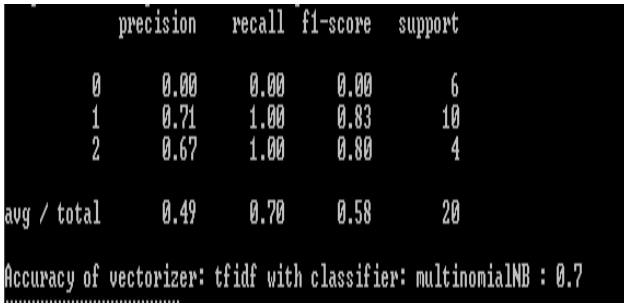

Fig.10. shows the classification accuracy of tf-idf with multinomial naïve bayes

It can be seen from the figure,the index number 0 represents neutral reviews,1 represents positive reviews and 2 represents negative review**.**F1-Score is calculated with the values of precision and recall for positive,negative and neutral reviews

F1score=2*0.71*1.00/(0.71+1.00)=0.83(positive review)

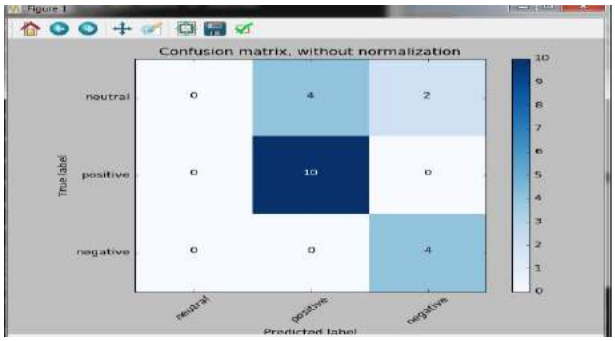

Fig.11. shows confusion matrix of tf-idf with multinomial naïve bayes

Similarly, the confusion matrix and accuracy is displayed for all combinations of vectorizers and classifiers

TABLE I. ACCURACY OF CLASSIFICATION ALGORITHMS

Algorithms used	Accuracy
Tf-idf with multinomial naives bayes	0.7
Tf-idf with decision tree	0.45
Tf-idf with SVM	0.2
Tf-idf with KNN	0.7
Count-vectorizer with decision tree	0.5
Count-vectorizer with multinomial naives bayes	0.7
Count-vectorizer with SVM	0.2
Count-vectorizer with KNN	0.6
Hashing-vectorizer with SVM	0.2
Hashing vectorizer with decision tree	0.55
Hashing vectorizer with multinomial naives bayes	0.65
Hashing vectorizer with KNN	0.65

It can be seen from the table 1 that multinomial naives bayes with any vectorizer choosen gives a better performance when compared to other algorithms.

TABLE II. PRECISION,RECALL AND F1-SCORE OF CLASSIFICATION ALGORITHMS

Algorithms used	Precision	Recall	F1-measure
Tf-idf with multinomial naives bayes	0.49	0.70	0.58
Tf-idf with decision tree	0.49	0.45	0.43
Tf-idf with SVM	0.04	0.20	0.07
Tf-idf with KNN	0.68	0.70	0.68
Count-vectorizer with decision tree	0.41	0.50	0.43
Count-vectorizer with multinomial naives bayes	0.49	0.70	0.58

Count-vectorizer with SVM	0.04	0.20	0.07
Count-vectorizer with KNN	0.60	0.60	0.50
Hashing-vectorizer with SVM	0.04	0.20	0.07
Hashing vectorizer with decision tree	0.39	0.55	0.45
Hashing vectorizer with multinomial naives bayes	0.46	0.65	0.54
Hashing vectorizer with KNN	0.65	0.65	0.64

It can be seen from the table 2 that multinomial naive bayes with any vectorizer choosen gives a better Precision,recall and f1 score when compared to other algorithms.

V.CONCLUSION

In this paper, a set of techniques are discussed and compared for mining and predicting the polarity of product reviews based on data mining and natural language processing methods. The objective is to make prediction of a customer reviews of a product sold online. The experimental results indicate that the techniques used are very promising in performing their tasks. Predicting the reviews is not only useful to customers, but also to product manufacturers to improve their product quality and features thus helping then in marketing.

REFERENCES

[1] Aamera Z. H. Khan, Dr. Mohammad Atique, Dr. V. M. Thakare "Sentiment Analysis Using Support Vector Machine",International Journal of Advanced Research in Computer Science and Software Engineering, Volume 5, Issue 4, April 2015

[2] Jayashri Khairnar, Mayura Kinikar "Machine Learning Algorithms for Opinion Mining and Sentiment Classification" International Journal of Scientific and Research Publications, Volume 3, Issue 6, June 2013

[3] Amit Gupte, Sourabh Joshi, Pratik Gadgul, Akshay Kadam, "Comparative Study of Classification Algorithms used in Sentiment Analysis" Amit Gupte et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (5) , 2014, 6261-6264

[4] V. Hatzivassiloglou and K. R. McKeown, "Predicting the semantic orientation of adjectives," in Proc. 8th Conf. Eur. Chap. Assoc. Comput.Linguist., Morristown, NJ: Assoc. Comput. Linguist, 1997, pp.174–181.

[5] Vikram Singh and Balwinder Saini "AN EFFECTIVE PRE-PROCESSING ALGORITHM FOR INFORMATION RETRIEVAL SYSTEMS" International Journal of Database Management System(IJDMS) Vol.6, No.6, December 2014

[6] Ms. Gaurangi Patil, Ms. Varsha Galande, Mr. Vedant Kekan, Ms. Kalpana Dange "Sentiment Analysis Using Support Vector Machine" International Journal of Innovative Research in Computer and Communication Engineering (An ISO 3297: 2007 Certified Organization) Vol. 2, Issue 1, January 2014

[7]. Komal Shivaji Mule,Arti Waghmare "Improved Indexing Technique for Information Retrieval based on Ontological Concepts"International Journal of Computer Applications (0975 – 8887)National Conference on Advances in Computing (NCAC 2015).