A Study of Different Methods & techniques for Stemming in Gujarati Text Mining

Aneri Boradia[1],

M. Phil Student, Gujarat Vidyapith, Ahmedabad.

Neepa Shah[2]

Associate Professor, Dept. of Computer Science, Gujarat Vidyapith, Ahmedabad.

ABSTRACT:

Text Analytics is as familiar as text mining. Text mining is a resource of stemming process and Stemming is the method of commuting incurved words to their word stem. To analyze the data, first we have to understand its word, sentence and content level, and then we can do for the further process. Stemming is a needful step in any language text mining and it is also helpful in Natural Language Processing. Stemming is done for many languages like Tamil, Punjabi, Marathi, Urdu, Hindi, Bengali, English, Gujarat, etc. To understand any sentence or paragraph, morphological analysis is the first step of the process. It is not necessary for a given stem and morphological root of the word to be equal, but it is generally correlated words mapped to the similar stem. We have studied different methods, algorithms and techniques of stemming in Gujarati language as well as other Indian languages. It is comprehensive analysis that we have put in this paper. We have also mentioned stemming errors, Gujarati language morphology, Supervised & Unsupervised learning methods.

KEYWORDS: Text analytics, Natural Language Processing, Information Retrieval System Gujarati Language, Morphology, Stemming.

1. INTRODUCTION:

The Internet plays a very important role in real- life day to day activities. In today's world, the internet has become an essential tool to get valuable information using web mining. We receive only the required information from World Wide Web which we demand to access and save our efforts and time. There is lots of information, videos, images and documents which are added massively on the internet on a regular basis. Text mining is one part of web mining. It has been worked on respective methodology including Information Retrieval methods. During a last few years, there has been an extremely large growth in volume of data produce around the world. The need to allow researchers to find information among these huge collections of data promoted the implementation of some mechanisms to process. These mechanisms are part of the Information retrieval process today. The Information Retrieval process is a full, extensive, and specialized field of research in Information Technology.

Text Mining obtains very rich quality knowledge from text. It works on unstructured information. Many authors and researchers have been doing work on Indian as well as non Indian languages. Gujarati language has been always first to be searched in the domain. Due to its rich morphology, Gujarati language does not have the same structure as any other Indian language. In this type of language documents, many words and contents sharing same morphological can be belonging to the same group. Morphological variants word relates to stemming process. Information Retrieval System helps to stemming for decrease size of the index files. Due to the rapid growth of the internet, stemming algorithm plays vital role in Information Retrieval System for improving of all regional languages. In this paper we cover all the methods of supervised learning & unsupervised learning [5].

2. GUJARATI LANGUAGE MORPHOLOGY:

In the world there are a total of 1690 main languages. Gujarati is the official language of the state of Gujarat; it is spoken by more than 47 million people in the world, giving it the rank of 23rd most spoken language in the world. Gujarati language is written from left side to right side of the page and there is no capitalization in it [5]. To work on Gujarati language first of all we must need to know Gujarati grammar form. How it is work? What is the meaning of this word or sentence? What are verbs, nouns, pronouns, adjectives, adverbs, conjunction etc.. Gujarati language grammar is the study of word order, morphological and syntactic structure and conjugation. So we study on Gujarati language grammar and identify its terms and meanings in detail. We also studied there architecture.

The Gujarati language morphology is bit complex. This language has three genders (masculine, neuter, feminine), two numbers (singular, plural) three nouns (nominative, oblique and locative) and two persons (first, second). Noun is assign by either its meaning or ending. Adjective is assign by its ending based on cases. Verb is inflected based on combination of gender, number, person etc. The Gujarati Language has 34 consonants and 12 vowels. The vowels can be dependent (I , ø , e, È , O), case to give a sound to the consonant or it can be independent (આ, ઊ, ઓં, ઈ, ચો, ઉં) [12]. These dependent vowels are written as postfix or prefix in the word means in upper part or lower part or top part or bottom part to the consonant.

2.1 What is Morphology in Gujarati language?

Morphology is the first determinant step in Natural Language Unit. It is study of the way the words are built up from smaller meaning – bearing units. Morphology is the descriptive analysis of the word. Below we display Gujarati language architecture.

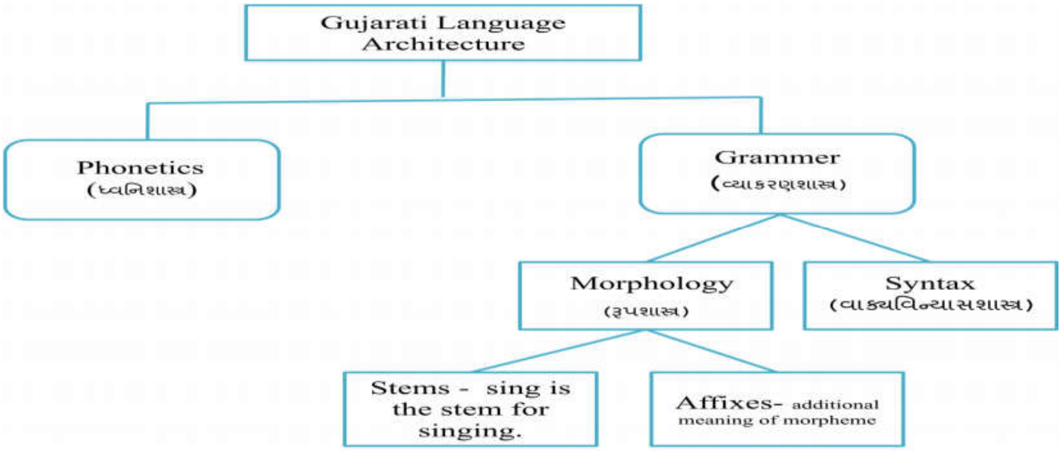

Gujarati Language Architecture
Phonetics (ધ્વનિશાસ્ત્ર)
Grammer (વ્યાકરણશાસ્ત્ર)
Morphology (રૂપશાસ્ત્ર)
Syntax (વાક્યવિન્યાસશાસ્ત્ર)
Stems - sing is the stem for singing.
Affixes- additional meaning of morpheme

Fig. 1 – Gujarati Language architecture

Gujarati grammar has been divided in two parts: Open class and close class. Open class contains noun, Verb, Adjective and Adverb, which are useful in our future work. Close class contains pronoun (સર્વનામ) and Conjunction (સંયોજન), but at this level close class is out of my scope. In future we will try to research on pronoun, conjunctions, and adverb.

Open class contains following terms:

Noun (નામ): A word that is the name of person, name of specific things or idea is called noun.

Adjective (વિશેષણ): A word that tells you much and more about noun is called adjective.

Verb (ક્રિયાપદ): A group or single word is used to indicate that something happens or exists is called verb.

NOUN :	Stem + Affixes	Noun /Adjective /verb
	અપેક્ષા + નું	અપેક્ષાનું
	અપેક્ષા + ને	અપેક્ષાને
	અપેક્ષા + માંથી	અપેક્ષામાંથી
ADJECTIVE :	સાર + ઓ	સારો
	સાર + ઈ	સારી
	સાર + રૂ	સારું
VERB :	રમ + વું	રમવું
	રમ + ઈ +ને	રમીને
	જમ + વું	જમવું

Table 1- Common words with their stem and affixes

3. STEMMING LEARNING METHODS IN GUJARATI LANGUAGE:

Stemming learning methods can be classified in to two groups: supervised learning and unsupervised learning. Researchers use different kinds of algorithm to explore structure in big data.

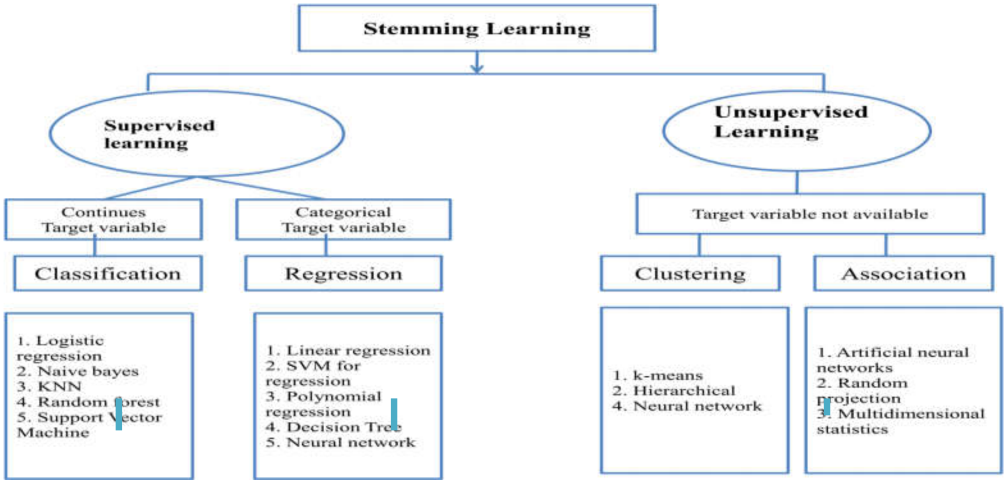


Fig.2 stemming learning method

3.1 SUPERVISED LEARNING METHOD :

Majority of researchers and learners are using the supervised learning method. It is a mixture of artificial intelligence and machine learning with the task of data mining. In this method we have to input variable and output variable with labels. We process with the help of algorithm to learn mapping function from input to the output. In this learning method, data is labeled, i.e. some data is already tagged with right output. Supervised learning algorithm analyses the training data and process for correct output from labeled data. This method consists of two types of approaches:

3.1.1 CLASSIFICATION & REGRESSION :

Classification is used for a model where output variable being determined as a class. Learning from the label data to create a model then predicting a target class for the given input data. Logistic regression method is one of the famous machine learning algorithms. It is used for classification tasks. Logistic Regression measures the relationship between the dependent variable and one or more independent variables by estimating probabilities using its underlying logistic function. Logistic regression separates our input into two regions by a linear boundary, one for each class, therefore it is required that our data is linearly separable, like data points. Two advanced techniques for helping 'support vector machine' (SVM) deal with imbalanced training data and the difficulty of obtaining human-annotated examples – two problems that frequently arise in NLP datasets. SVMs is reducing the need for labeled training examples in both the standard inductive and transductive settings with the help of in text and hypertext categorization.

Classification and Regression used K-Nearest Neighbors algorithm. KNN is a non-parametric method. The input comprises of the *k* closest training examples in the feature space in both cases. The output depends on KNN. Either KNN used for classification or regression. Naive Bayes is a easy technique for constructing classifiers models. That model assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. Decision tree can be used to represent decisions and decision making. Operations research and operations management used Decision trees. NLP is highly knowledgeable, interdisciplinary, involving concepts in computer science, linguistics, logic, and psychology. Many aspects of the field deal with linguistic features of computation and NLP seeks to model language computationally so NLP has a special role in computer science [9].

Polynomial regression is a form of regression analysis in which the relationship between the independent variable and dependent variable is model of degree polynomial.

Polynomial regression is technically a special case of multiple linear regressions; the interpretation of a fitted polynomial regression model requires a somewhat different perspective. A random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction [6], [9].

Example: Assume that we have applied K-Nearest Neighbor classifier technique on our data. KNN is non-parametric recognition method. First step is pre process the document. To pre-processes the sentence we divide texts into number of individual tokens to reduce the unwanted contents from sentence. No training is required. It estimates the posteriori probability from frequency of nearest neighbor of unknown pattern. It store data as a template and whenever new sentence comes it finds the lowest distance between unknown data and training sample's templates. Then after it calculates most of nearest neighbor and assign the class label.

૧. નેમી ઘરે આવી. ૨. નેમી અપેક્ષાસાથે ઘરે આવી.

In above example first preprocess on sentence and generate the result. નેમી, અપેક્ષા, ઘર these are recognize as noun. આવી, આવવું, આવશે this words are known as verb. And અપેક્ષા (આશા) is also recognizing as adjective.

Noun: નેમી, અપેક્ષા, ઘર   verb: આવી   adjective: અપેક્ષા

3.2 UNSUPERVISED LEARNING:

3.2.1 CLUSTERING & ASSOCIATION:

Unsupervised learning is the training of an artificial intelligence (AI) algorithm using information that is neither classified nor labeled and allowing the algorithm to act on that information without guidance [9]. Learning from unlabeled data is to differentiating the given input data. k- Means clustering is a method of vector quantization, which is popular for cluster analysis in data mining. K-means clustering ideal to division *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster[9]. Hierarchical clustering represents "bottom up" approach. This bottom up approach observation starts in its own cluster, and pairs of clusters are merged as one move up the hierarchy. A Hidden Markov model is finite state machine. This model can be considered a normalization of a mixture model, where the hidden variables (or latent variables), which control the mixture component to be selected for each observation, are related through a Markov process [9].

4. ERRORS OF STEMMING:

Languages are not perfectly punctual constructs, and therefore stemming process make mistakes. There are two types of error measurements in stemming algorithm.

Over stemming: It is a type of error in stemming processing. When two individual words are stemmed to the same root it's called over-stemming. It is also called as false positive [2]. Two different words belong to different groups, and remain typical after stemming, and stemmer has responded correctly. If though are converted to same stem, is counted as an over stemming error.

Ex.: "ઉત્તપમ ", "ઉત્તમ" and "ઉત્તર" and this is example of over stemming: though these three words are etymologically related but their meanings are different, so working as synonyms in search engine.

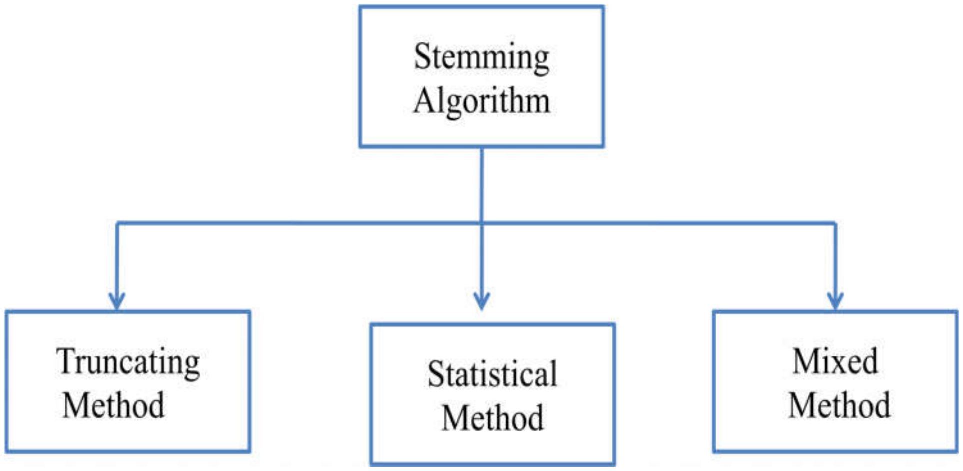
Under stemming: This error happens when two different words should be stemmed to the same root are not. It is also called as false negative [2]. Two different words belong to same group, and remain typical after stemming, and connection is correct, they are converted to other steams, is counted as an under stemming error.

Ex.: “અપેક્ષાએ”, “અપેક્ષાની”, “અપેક્ષામાંથી”, “અપેક્ષાનુ” this Gujarati word keeps morphology, so these near synonyms are not combining.

By counting these errors examples, we can gain good sense into the process of stemmer, and compare with different stemmer.

5. STUDY OF DIFFERENT STEMMING ALGORITHMS:

There are main three type of stemming algorithms in data mining. Researchers and Learners have worked on its techniques and methods for evaluating the strength and similarity of stemming algorithms in different ways and generate results with a high accuracy rate. We describe all methods in below given figure. Also we have made a comparison table for all stemmer. In this table we have described brief information with its advantage and disadvantage. These algorithms are used for all languages.



1. Lovins Stemmer
2. Porter stemmer
3. Paice / Husk stemmer
4. Dawson Stemmer

1. N-Gram Stemmer
2. Yass Stemmer

1. Krovetz stemmer
2. Xerox stemmer
3. Corpus based stemmer
4.Light weight stemmer
5. Hybrid stemmer
6. DHIYA stemmer

Fig 4.Stemming algorithms

5.1 COMPARISON OF DIFFERENT STEMMERS:

Algorithm	Description	Method	Accuracy	Advantage	Disadvantage
Truncating Method					
Lovins Stemmer **[1968]**	This algorithm process very fast and it has impressively traded space for time. It has huge suffix dataset. Remove suffix and double letter and also help in irregular plurals. [2]	Transformation rules / Truncating method	36%	It is faster. It has effectively traded space for time, and with its large suffix set. [2]	It is time consuming. Also not all suffixes available. It is also dependent on technical vocabulary being used by author.
Dawson Stemmer **[1974]**	This algorithm covers a much more comprehensive list of about 1200 suffixes. They are organized as a set of branched character trees for fast access[5].	Truncating method	Not mentioned.	This algorithm Covers more suffixes than Lovins stemmer. It is fast in execution.	It is very complex structure for understand. Lacks a standard reusable implementation.
Porter stemmer **[1980]**	This algorithm process for removing in flexional endings means longest match suffixes from words in English. [11]	Suffix Stripping	33%	Produces the best output as compared to other stemmers. It creates Less error rate. [2]	The stems produced are not always real words. It has at least five steps and sixty rules and hence is time consuming
Paice / Husk stemmer **[1990]**	This algorithm follow rule based approach. They developed pre define rule set. That tries to find rule by the last character of the word and perform cut or replace purpose[2].	Qualitative evaluation Method based on the over- and Under stemming / Truncating method	96%	It is a single form so too much care of both cut or replace in validation with rule applied.	This algorithm may be generating over stemming error. It is very bulky algorithm.
Statistical Method					
N-Gram Stemmer	This algorithm widely used in	probability distribution /	82.5%	It is not language	This algorithm is time

[1974]	statistical approach (NLP). It can use for efficient matching and convert the word curvature in its root. [2]	suffix stripping		dependent. So that it is more useful in may application.	consuming. It is not practical system.
YASS stemmer **[2007]**	This stemmer is corpus based and purely unsupervised method. It is based on suffixing.[5]	Suffix Stripper	89.9%	This stemmer is not language dependent. It is used for all language. and not necessary to knowing language's morphology.	It need meaningful computing power.
Mixed Method					
Krovetz Stemmer **[1993]**	This stemmer produces the words, not stems. It removes inflections in very accurate manner. This algorithm is often used as a first step for document, before using another stemmer. [5]	Hybrid approach	7%	This stemmer is very light compare to other stemmer. It's comparable effectiveness.	This stemmer produce lower false positive rate, sometimes higher false negative rate.
Xerox stemmer **[1994]**	This stemmer works on English lexical database. This can analyze and generate inflectional morphology. [2]	Linguistic Method	Not mention	This stemmer applied for large document. It's help to remove prefixes in our document. [5]	It is work only on English document, so it is language dependent. If word is not part of lexicon this method cant contain the word.
Corpus based stemmer	This stemmer effort to based carry off some of difficulty of porter stemmer.	Proposed method	94.85%	Using this stemmer over stemming and under stemming difficulty are solved.	For this researchers must develop the statistical measure for each increase of collection and processing.
Light weight Stemmer	This algorithm removes all possible prefixes	Rule based method	91.5%	This algorithm is able to hold most of	Using this algorithm we faced over

for Gujarati	from a word to have possible output. Also removes all possible suffixes. [5]			morphological variants. [5]	stemming and under stemming error.
Hybrid Stemmer for Gujarati	This algorithm is lightweight and removes the curvature endings with the use of EMILE corpus. [13]	Goldsmith's Approach = take -all-splits Method. [Rule based]	67.86%	It's performance is good. It's useful in dictionary search or data deflation. [13]	It is easily over come. It removes only curvature endings. [13]
DHIYA Stemmer for Gujarati	It can be done on morphological level. Author have identified 52 most appearing inflections and studied their possible sequence. [12]	Rule based / Morphological analyzer	92.41%	Handcrafted rule set created by author. Accuracy rate is good compare to other stemmer.[12	Over stemming and under stemming is occur. The words that not belong to Gujarati language are not stemmed consistently.[12]

6. CONCLUSION:

We have presented a comparative study of multifarious stemming. As a study of all the methods and techniques, there is an average difference between the algorithms. Many algorithms of stemming are based on rule based approach. Some of them remove suffix, some remove prefix, some work on duplicate data, etc. After studies, we assume that while some algorithm and techniques perform well in one way, the other techniques and algorithm perform well in other way. No algorithms and methods give a 100% accuracy rate, but are good enough to be applied in stemming approach in text mining, NLP and IRS. The main problem is that in Gujarati language some algorithms can't work on outside words from lexicon, some are very time consuming, and some are not implemented properly for successful output. How to identify verb, noun, adverb and adjective? Over stemming & under stemming approach makes error in document. Gujarati language works on lightweight stemmer & hybrid approach stemmer. It is work on rule base method. These techniques were developed to be useful in applications like corpus compression, information retrieval and pre processing modules in NLP problems.

7. FUTURE EXTENSION:

Many researchers have investigated stemmer techniques and methods in this field since long times. However, there are still some open questions, such as how to evaluate a stemmer independently of information retrieval system. Algorithm and techniques need some modification to perform better work, reduce the error rate and provide a high accuracy rate with help of supervised and unsupervised techniques. We will focus on which method is useful to identify Gujarati morphology and give us proper output. We will also try to developing an algorithm to remove suffix, prefix, duplicate word and give proper output. And also identify noun, adjective, verb and adverb.

REFERENCES:

1. Ananthakrishnan Ramanathan and Durgesh D Rao, "A Lightweight Stemmer for Hindi" *Information Storage and Retrieval*, 10(1):253–260

2. Anjali Ganesh Jivani., "A Comparative Study of Stemming Algorithms" Department of Computer Science & Engineering The Maharaja Sayajirao University of Baroda, Comp. Tech. Appl., Vol. 2 (6), 1930-1938.

3. Bijal, Dalwadi., & Sanket Suthar . (2014). "Overview of Stemming Algorithms for Indian and Non -Indian Languages*." International journal of computer science and information technologies.* Vol. 5(2), 1144 - 1146.

4. C. D. Paice, "An Evaluation Method for Stemming Algorithms," Proc. 17th Annu. Int. ACM SIGIR Conf.

5. Chandrakant Patel, Jayehkumar Patel. 2016 "Improving a lightweight stemmer for Gujarati Language", *International journal of Information Sciences and Techniques*, Vol. 6, N0. 1/2, March 2016.

6. J. Damerau, Nitin Indurkhya, Sholom M. Weiss, and Tong Zhang, ' Text Mining - Predictive Methods for Analyzing Unstructured Information' , ISBN – 0-387-95433-3 , 2005 , Spinger.

7. J. Goldsmith, (2001) "Unsupervised learning of the morphology of a natural language", *Computational Linguistics*, Vol. 27, No. 2, pp. 153-198.

8. R. Sheth and B. C. Patel, "Stemming Techniques and Naïve Approach for Gujarati Stemmer," pp. 975–8887, 2012

9. Jiawei Han, Micheline Kamber, Jian Pei , 'Data mining : Concepts and Techniques' 3rd / e , Waltham, MA 02451 , USA. 2012.

10. Julie Beth Lovins. 1968. "Development of a Stemming Algorithm" *Mechanical Translation and Computational Linguistics*, Vol. 11 (1) and Vol. 11 (2),March and June 1968.

11. F. Porter, "An algorithm for suffix stripping," Progr. Electron. Libr. Inf. Syst., vol. 14, no. 3, pp. 130– 137, 1980.

12. Ms. Jikitsha. Sheth and Dr. Bankim Patel, "Dhiya: A stemmer for morphological level analysis of Gujarati language", *2014 International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT),* 978-1-4799-2900-9/14/©2014 IEEE

13. P. Patel, K. Popat, and P. Bhattacharyya, "Hybrid Stemmer for Gujarati," Comput. Linguist.,no. August, pp. 51–55, 2010.

14. Pandit Prabodh, Joshi Daya Shankar, "Grammar – Meaning and form " University Granthnirman Board - Gujarat State, JB Shandil, 1978.

15. Desai Urmi Ghanhyambhai "Morphology – An Introduction" ,Parth Publication– 2007