Influence of GUJarati STEmmeR in Supervised Learning of Web Page Categorization

Chandrakant D. Patel
Research Scholar, Hemchandracharya North Gujarat University, Patan, Gujarat, India.
E-mail: cdpatel4phd@gmail.com

Jayesh M. Patel
Associate Professor, Acharya Motibhai Patel Institute of Computer Studies, Ganpat University, Kherva, Gujarat, India
E-mail: drjayeshpatel.phd@gmail.com

Received: 21 November 2020; Accepted: 19 February 2021; Published: 08 June 2021

Abstract: With the large quantity of information offered on-line, it's equally essential to retrieve correct information for a user query. A large amount of data is available in digital form in multiple languages. The various approaches want to increase the effectiveness of on-line information retrieval but the standard approach tries to retrieve information for a user query is to go looking at the documents within the corpus as a word by word for the given query. This approach is incredibly time intensive and it's going to miss several connected documents that are equally important. So, to avoid these issues, stemming has been extensively utilized in numerous Information Retrieval Systems (IRS) to extend the retrieval accuracy of all languages. These papers go through the problem of stemming with Web Page Categorization on Gujarati language which basically derived the stem words using GUJSTER algorithms [1]. The GUJSTER algorithm is based on morphological rules which is used to derived root or stem word from inflected words of the same class. In particular, we consider the influence of extracted a stem or root word, to check the integrity of the web page classification using supervised machine learning algorithms. This research work is intended to focus on the analysis of Web Page Categorization (WPC) of Gujarati language and concentrate on a research problem to do verify the influence of a stemming algorithm in a WPC application for the Gujarati language with improved accuracy between from 63% to 98% through Machine Learning supervised models with standard ratio 80% as training and 20% as testing.

Index Terms: Stemming, Gujarati Language, Supervised algorithms, Machine Learning, Accuracy.

1. Introduction

In Stemming, morphologically similar words are clustered together under the hypothesis that they are semantically similar [2]–[7]. It is useful for Text Mining, Natural Language Processing (NLP) functions, Text clustering, Text categorization, Text summarization, and application of Text Mining (TM) [1], [8, 9]. It is the initial stage before applying any kind of related categorization algorithm. Additionally, word-stemming is also supportive features for the indexing and search engine, now a day. The main goal of stemming is to remove inflectional or derivational ending from word to its stem form. Stemming, a very beneficial implement under the province of Information Retrieval (IR), supported by nearly every modern indexing and search system at the time of indexing and searching.

Several inflectional and/or derivation stemming algorithms were developed for European and Asian languages [9]. Stemming is a linguistic process in which the various morphological variants of the words are mapped to their base forms [5, 10]. The algorithms or the computer programs that perform the stemming process are called stemmers. Stemmer’s designs range from the simplest techniques, such as the elimination of affixes using a list of repeated affixes, to a more complex design that uses the morphology of words in the inference process to derive a stem.

According to J B Lovins [11], the stemming algorithm can be classified as iteration, longest match, and context-free and context-sensitive. However, the current state of the art approaches of stemmer can be classified as – Language Dependent and Independent including Corpus, Statistical, and Hybrid approaches in [Fig.1]. Each type of stemmer has a typical way to find the stem from variants of the word. The power of a stemmer is the volume of saving in the mass of the vocabulary is acquired. Aggressive or Strong stemmers may decrease the size of the directory for a given corpus extremely.
Many works have been countered on text classification in which the stemming algorithm has been used [12, 13]. The opinions of research are different from each other in the effectiveness of stemming for the aim of classification. While E. Riloff [14] and M. Spitters [15] had over that stemming might not facilitate increase classification accuracy, Buseman had discovered that morphological analysis will increase the performance for a series of classification algorithms applied to German e-mail classification. In recent work, Gustad and Bouma [16] determined that stemming doesn't systematically improve classification accuracy. Additional, however, Cohen et al. [17] found stemming advantageous whereas classifying medical documents. WPC is widely used in the area of vertical search, personalized search and recommendation system, etc. The impact of IR not only searching important terms but also massively depends on the methods of counting words which basically depends on tokenization, stemming and stop word removal process as a pre-processing step before applying text classification.[18] The researchers have following kind of research problems with research focus on influence of stemming algorithm in WPC application.

1. How the morphological rules of Gujarati words are impact.
2. Influence of stemming algorithms on WPC.
3. Which supervised models is most effective while categorization of Gujarati Web page.

The rest of this paper is structured as follows. Second section reviews earlier literature on the stemming and WPC particularly on Gujarati, Hindi and English language. Section three denotes research motivation which indicates modules wise in depth information for it with proposed methodology for WPC is presented. In section four discussed with data analysis and algorithm discussion with result on machine learning supervised models presented and implemented in this research.

2. Related Work

Numerous typical methods that accomplish stemming are in the literature for three languages, including English [Table 1], Hindi [Table 2], and Gujarati [Table 3] with comparison.
Table 1. Comparison study of English stemmer

| Sr. No. | Authors Name & Reference | Data sets | Year | Approach | Evolution methods or Accuracy in % |
|---------|---------------------------|-----------|------|----------|-----------------------------------|
| 1       | Julie Beth Lovins [11]    | Not Given | 1968 | Longest Matching | Not Given |
| 2       | Margaret A. Hafer et. al. [19] | Brown, CPC, ADI | 1974 | Letter Successor | Precision, Recall |
| 3       | M. F. Porter [19, 20]     | Cranfield | 1980 | Suffixes Stripping | Precision, Recall |
| 4       | Donna Harman [22]         | Cranfield, Medlars, CACM | 1991 | Not Given | Ranking with Noise, IDF measure |
| 5       | J. L. Dawson et. al. [23] | Not Given | 1974 | Not Given | Not Given |
| 6       | C. D. Pacey et. al. [24]  | Not Given | 1994 | CISI | UI, UI |
| 7       | Xu and W. Croft et. al. [5] | WEST, WSJ, WS91, ISM | 1998 | Corpus-based | Precision, Recall |
| 8       | Robert Krovetz et. al. [25] | CAACL, NPL, TIME, WEST | 1992 | Inflectional, Derivational | Precision, Recall |
| 9       | James Allan et. al. [26]  | AP89, AP90, FBIS, TREC5, TREC8 | 2003 | Corpus-based | Precision, Recall |
| 10      | Bhamidipati Narayan L et. al. [27] | 20NG, WebKB, WSJ | 2007 | Corpus-based | Precision |
| 11      | Vaarapraaksh Gurusamy et. al. [26] | Moby | 2017 | Performance Analysis | Accuracy |

In [Table 1], a wide variety of approaches developed in six decades stemming from the English language. Most of the stemming algorithms are the corpus-based and suffixes approach which is most suitable for the English language.

Table 2. Comparison study of Hindi stemmer

| Sr. No. | Authors Name & Reference | Data sets | Year | Approach | Evolution methods or Accuracy in % |
|---------|---------------------------|-----------|------|----------|-----------------------------------|
| 1       | Ananthakrishnan Ramanathan et. al. [29] | Hindi news | 2003 | Suffix removal, Lightweight stemmer | Promising result |
| 2       | Leah Larkey et. al. [30]  | Hindi news | 2003 | Lightweight | Not Given |
| 3       | Amarendra Kumar Pandey et. al.[31] | EMILLE | 2008 | Unsupervised | Accuracy 89.9% |
| 4       | Niraj Aswani et. al. [32]  | EMILLE | 2010 | Morphological Analyzer | Precision, Recall |
| 5       | Upenendra Mishra et. al. [33] | Manually | 2012 | Hybrid | Accuracy 91.59% |
| 6       | Anubha Jain et. al. [34]   | FIRE | 2013 | Hybrid | MAP 3.40% |
| 7       | Vishal Gupta [35]          | EMILLE | 2014 | Rule-based (noun) | Accuracy 83.65% |
| 8       | Rakhi Joon et. al. [36]    | CFILT | 2017 | MWES (N-gram) | Accuracy >90% |

In [Table 2], a wide opportunity taken into stemming algorithms with rule-based, hybrid, unsupervised, n-gram, and lightweight for National language Hindi on different kinds of a corpus such as CFILT, FIRE, and EMILLE. The Hindi language got maximum stemming accuracy of 91.59% using a hybrid approach.

Table 3. Comparison study of Gujarati stemmer

| Sr. No. | Authors Name & Reference | Data sets | Year | Approach | Evolution methods or Accuracy in % |
|---------|---------------------------|-----------|------|----------|-----------------------------------|
| 1       | Prasenjit Majumder et. al. [3] | TREC, CLEF | 2007 | Statistical | Precision, Recall |
| 2       | Pratik Patel et. al. [37]  | EMILLE | 2010 | Hybrid | Accuracy 67.86% |
| 3       | Kartik Suba et. al. [38]   | EMILLE | 2011 | Hybrid | Inflectional, Rule-based derivational |
| 4       | Juhi Ameta et. al. [39]    | EMILLE | 2011 | Lightweight | Accuracy 91.5% |
| 5       | Jikish Sheth et. al. [40]  | Not Given | 2011 | Model proposed | Not Given |
| 6       | Jikish Sheth et. al. [41]  | EMILLE | 2012 | Hybrid | Accuracy 92.41% |
| 7       | Bijal Dalwadi et. al. [42] | EMILLE | 2016 | Affix removal | Accuracy 96.63% |
| 8       | Miral Patel et. al. [43]   | MIRI | 2015 | Suffix stripper (noun) | Accuracy 85.83% |

In [Table 3], a hybrid approach is applied for the stemming of the Gujarati language and evaluated on the EMILLE corpus. The rule-based approach is most suitable for the Gujarati language as per the result taken by the researcher that is a maximum of 96.63% for affix removal and 92.41% for the hybrid approach. The first efforts made in 2011 for the
Derivational based stemming; the accuracy was up to 70.7%. In the literature review, the researcher found very few resources for stemming processing on the Gujarati language. So, it is an urgent need to develop a mature stemming algorithm which can help for several Indian languages including Gujarati, Marathi [44], Odia, and Pashto.

3. Methods

3.1. Research Motivation

Web page classification (WPC), also known as Web Page Categorization, is the process of assigning a Web page to one or more predefined category labels [45]. The research concentrates on stemming which involves the following major modules:

A. Pre-processing for the Gujarati language web page

The first module of pre-processing mentioned in the above paragraph as the first step in web page classification process includes only a small percentage of the pre-processing sub-steps: Cleaning or normalizing the data (remove all Non-Gujarati Alphabets, Numerals, and Symbols including HTML tags), Tokenization, and generate the Gujarati words lists.

B. Identifying the stop words list

This module processed only those kinds of words which are less important for categories of the web page. The techniques of stop words identification such as static list or dynamic lists using rules, statistical or hybrid approach.

C. Finding the stem of a word

The Stemming process is language dependent as well as independent. The Information Retrieval System works accurately when there is standard stemming process for all languages. Generally, stemming algorithm is replaced or removes affixes to the word to generate the root word.

D. Evaluation of stemmer on WPC

The fourth module is performed the stemming evaluation using WPC. In the past few years, there are several stemming algorithms, and WPC methods have been developed for the non-Asian languages. Standard benchmark corpora need a predefined training/testing splits or k-fold cross validation statistical methods to allow comparison against different WPC approaches. Otherwise, a typical training/testing split of a labelled corpus is 80: 20.

3.2. Research Methodology

As per the constitution of India, 22 languages are officially included in the census of 2011. The record mentioned that the Gujarati language is scheduled language and categorized as Gujarati, Gujrao / Gujrav, Pattani, Ponchi, Saurashtra / Saurashtri, and others. The Gujarati language got the rank of 6 with 4.58% speakers in the 2011 census and that was increased by 20.40% in the decade of 2001-2011. The family of the Gujarati language mentioned as a member of Indo-Aryan. It is closely allied to Marathi, Hindi, Bengali, and Punjabi [46]. In today’s world, the Gujarati is spoken by more than 68 million people including Gujarat, Rajasthan, Karnataka, Madhya Pradesh, and Maharashtra. Moreover, it is also used beyond India like in USA, Kenya, Pakistan, Fiji, Bangladesh, Oman, South Africa, UK, Uganda, and Zambia. The Jain monk Hemchandracharya, a renowned scholar of Gujarat from the 8th century, composed the grammar book 'Siddhem Shabdanusasan – सिद्धेम शब्दानुशासन' in addition to the Sanskrit language, which is why his praises are sung today not only in Gujarat but also in India. [46, 47].

The approach to the Gujarati WPC proposed in this paper is based on contemporary methods for performing automatic textual classification which initiates operation are remove stop words identification and stemming. This proposed method involves extracting the Gujarati language textual from the web page, encoding the document as features vector with machine learning models. This subsection gives a brief overview of the methodology used in this paper. The methodology is summarized in.
This research methodology logically divided into two sections i.e. up to stemming and classification. The stemming section considered with several subsections such as cleaning, tokenization, stop words while classifier has features vector, selection, and evaluation with a conclusion as depicted in [Fig.2]. In the stemming section, the researcher studied relevant algorithm and approaches which are performed to identify stop words and to do stemming to get an understanding of the field and find a suitable approach and develop an algorithm that can be used for the classifier section. In the classifier section, researcher has found a suitable machine learning algorithm and there are no hard criteria for selecting a specific algorithm or method. This is followed by proposing a general method of Gujarati WPC using text classification methods. To evaluate this method, a framework for the Gujarati WPC is developed that implements the general method and provides support for several algorithms that have been considering for study. The general method performs, to establish how well several experiments are done by performing Gujarati WPC using the framework (with the several classifiers) on standard data sets.

4. Algorithm and Data Analysis

4.1. Algorithm Discussion

The algorithm has been defined with specific purpose for the given set of inputs. In this discussion, we have implemented proposed algorithm with input and desired output such as syllable tokenizer, search dynamic stop words and stemming algorithm for the given research work.

A. Syllable Tokenizer algorithm

Tokenization is a process of computer science to break the mass of texts into smaller units. In other words, tokenization is a process of segmenting a string of characters into words. The unit may consider as sentences, words, keywords, phrases, symbols but that are dependent upon application and languages [49].

Table 4. Sample output of Gujarati Syllable tokenizer

| Word     | Syllable of Word |
|----------|-----------------|
| આં          | ["આ"]         |
| કોવીડ      | ["કો", "વી", "ડ"] |
| મહામારીમાં | ["મ", "હા", "મા", "રી", "માં"] |
| લોકડાઉન   | ["લો", "ક", "ડા", "ઉ", "ન"] |
| જરૂરી      | ["જ", "રૂ", "રી"] |
| પગલ ૛ે    | ["પ", "ગ", "લ ૛ે"] |
| છે          | ["પ"]         |

Tokenization process applies to larger text to divide into smaller text. In text mining, a tokenization is paragraph > sentences > words > syllable > character. The researcher proposed and implemented Gujarati syllable tokenizer for stemming and stop words operations basis on meaningful syllable with sample output in [Table 4].
Influence of GUJarati STEmmeR in Supervised Learning of Web Page Categorization

Step 01 Read a word
Step 02 Mention list of all vowels, consonants and diacritics
Step 03 Declare C-tokens as list and s as string
Step 04 Read each character from word
Step 05 if character in vowel list or consonant list
   Go to Step – 06
   Otherwise
   Go to Step – 10
Step 06 if s last character is Halant (‘્’):
   Go to Step – 07
   Otherwise
   Go to Step – 08
Step 07 Append current character into string
Step 08 Append string to C-tokens list
Step 09 Create string as empty and add current character into string
Step 10 if character in Diacritic list:
   Go to Step – 11
Step 11 Append current character into string
Step 12 Append string to C-tokens list and return C-tokens

B. Dynamic Stop words Identification algorithm

Stop words are frequent words that carry no information [50]. They are topic-neutral words. Removing stop words helps in saving space and simplifies further steps in classification [50, 51]. Choosing appropriate stop words is of language-specific and domain-specific choice, i.e. a word considered as stop word in one domain might be a good discriminator but not in another domain. The usefulness of stop words in web page categorization is also task-dependent [53]. A word in Gujarati language has mainly two kinds as per usage of diacritics such as non-diacritic word or diacritic word. A non-diacritic word doesn’t contain Halant (‘્’) sign for example, પર and ઉપર while a diacritic word does contain Halant (‘્’) sign for example, અને to અને. To identify words based on its length, we need syllable tokenizer for word into syllable tokens. Such as “ઉપર” > “ઉ”, “પ”, “ર”, and “અને” > “અ”, “ન”, “ે”. Here, we consider count of syllable that is the length of that string. So, in the above example we consider the length of word 3 and 2 respectively.

Step 01 Read words and create a list of unique words
Step 02 Read each word from unique words
Step 03 If word is non-diacritic word
   Apply non-diacritic words rules.
   Go to Step – 05
Step 04 If word is diacritic word:
   Apply diacritic words rules.
   Go to Step – 05
Step 05 Append word to stop word list and return stop words

C. GUJarati STEmmeR (GUJSTER) algorithm

A word might consist of a stem or root and affixes. The stem is that the main linguistics elements because it contains that means. The role of stemmer is to spot the stem of a given word, such all the inflected words of same stem get mapped there to. Word stemming is one in every of the fundamental method exhausted applications Natural Language Processing (NLP), Information Retrieval (IR), Machine Learning or Translation (MT), Language Modelling (LM) and alternative information applications. Since inception of NLP domain, there are several approach developed to stemming the words for Indian and non-Indian language discussed in literature review.
The text classification is widely used in Natural Language Processing (NLP). Using text classification, the researcher can train model to supervised learning. To categories the Gujarati web page, researcher used supervised learning algorithms that can predicate with labelled corpus. Using label, an algorithm can understand the patterns and correlation between words. The corpus used in this research paper is hybrid in nature. It is collected from several domains such as EMILLE, TDIL and manually. A machine learning models can’t work directly on texts. So, features engineering is a process to transform texts into useful numeric features that act as input for that and will improve the quality of the model with enhanced performance. When research has text data, there are several methods to gain features that represent the texts as features sets.

In Bag of Words (BOW), the orders of terms in sentences are not considered and for this application, it is not important that orders terms can classify well than unordered terms. The researcher has selected TF-IDF vectors mechanism to generate the features engineering of the document in the corpus. The valid reasons for selection are...It is simple and easy to understand the results for the particular domain. It is very fast process and takes less waiting. The researcher can implement features generation process to resolve the over fitting problem. In addition, Joshi Hardik et. al. [54] have compared various IR (Information Retrieval) models and concluded that TF-IDF (Term Frequency / Inverse Term Frequency) model performed well for the Gujarati newswire corpus.

A numeric features can easily predicate by machine learning models. So, the researcher must create a mapping for each category with numerical value. The machine learning model quality will prove by the test datasets while training datasets. The researcher has selected randomly 80:20 ratios for the training and test, respectively. The researcher also provide the hyper parameter tuning process to cross validation on the training data and fit with final model and then going to evaluate the model with unseen data to obtain some metrics which is less biased as possible.

4.2. Data Analysis

The stemmer datasets which include all kinds of characters with Gujarati and non-Gujarati for the testing are 1028 files and the average length of sentence in the corpus is 13 to 15 and the average word length is 7. These all kinds of datasets are hybrid approach because few of the resources are collected from internet and others are collected or crafted manually. The datasets collected from EMILLE, TDIL and other sources.

Table 5. Statistics of Datasets

| Type                        | Features      | Count         |
|-----------------------------|---------------|---------------|
| Without Filter (Non-Gujarati & Gujarati Characters) | Total Sentences | 862865        |
|                             | Total Words   | 11287001      |
|                             | Total Unique Words | 656806       |
| With Filter (Gujarati Characters) | Total Sentences | 839466        |
|                             | Total Words   | 11080306      |
|                             | Total Unique Words | 485949        |

The statistics of words are as per the length with the help of the Syllable word algorithm mentioned in [Table 5]. The minimum length of Gujarati words is one and the maximum length of Gujarati words is fourteen in our corpus.
5. Results and Discussion

The researcher has tested several machine learning models to work out which one may fit better to the information and properly capture the relationships across the points and their labels. The researcher got only used classic machine learning models rather than deep learning models thanks to the insufficient amount of knowledge. The researcher got, which might probably result in over fit models that doesn’t generalize well on unseen data. The researcher has tested & evaluated the following models such as Random Forest, Support Vector Machine, K Nearest Neighbors, Multinomial Naïve Bayes, Multinomial Logistic Regression, and Gradient Boosting.
Table 8. Classification report of multinomial naive bayes in balanced multi-categories models

| Category | Precision | Recall | F1-Score | Support |
|----------|-----------|--------|----------|---------|
| 0        | 1.00      | 1.00   | 1.00     | 1       |
| 1        | 1.00      | 1.00   | 1.00     | 1       |
| 2        | 1.00      | 0.67   | 0.80     | 3       |
| 3        | 1.00      | 0.50   | 0.67     | 4       |
| 5        | 0.50      | 1.00   | 0.67     | 1       |
| 9        | 0.50      | 1.00   | 0.67     | 2       |

|                    |            |        |          |         |
| Accuracy           |            |        | 0.75     | 12      |
| Macro Average      | 0.83       | 0.86   | 0.80     | 12      |
| Weighted Average   | 0.88       | 0.75   | 0.76     | 12      |

Fig. 3. Confusion matrix of Multinomial Naive Bayes in Balanced Multi-categories Models

Table 9. Classification report of logistic regression in balanced single category model

| Category | Precision | Recall | F1-Score | Support |
|----------|-----------|--------|----------|---------|
| 0        | 0.93      | 0.97   | 0.95     | 29      |
| 1        | 0.96      | 0.93   | 0.94     | 27      |
| 2        | 1.00      | 1.00   | 1.00     | 23      |
| 3        | 1.00      | 1.00   | 1.00     | 25      |

|                    |            |        |          |         |
| Accuracy           |            |        | 0.97     | 104     |
| Macro Average      | 0.97       | 0.97   | 0.97     | 104     |
| Weighted Average   | 0.97       | 0.97   | 0.97     | 104     |
Each one of them has multiple hyper parameters that must be tuned. The researcher has followed the subsequent methodology when defining the most effective set of hyper parameters for every model. Firstly, the researcher has decided which hyper parameters, wish to tune for every model, taking under consideration those which will have more influence within the model behaviour, and considering that a high number of parameters which require lots of computational time.

Then, researcher has defined a grid of possible values and performed a Randomized Search using 3-Fold Cross Validation (with 50 iterations). Finally, once researcher has get the model with the most effective hyper parameters, performed a Grid Search using 3-Fold Cross Validation centered in those values so as to exhaustively search within the hyper parameter space for the most effective performing combination. The researcher has followed this system because with the randomized search able to cover a way wider range of values for every hyper-parameter without incurring in really high execution time. Once researcher has narrowed down the range for everyone, where to concentrate on search and explicitly specify every combination of settings to do.

The reason after electing K = 3 as the number of folds and 50 repetitions in the randomized search originated from the trade-off between smaller performance time and testing a high number of mixtures. When electing the superlative model in the process, the researcher has elected the accuracy as the evaluation metric. To check the performance on test sets that is totally unseen data, the researcher need to train the model with hyper parameters tuning process including cross validation and fitting the model on train sets. In WPC problems, there are several performance metrics which will be accustomed to gain insights on how the model is performed. The researcher has classified the datasets into balanced and unbalanced with same category (news) or multi-category documents (Asiannet, Astrology, Business, Crime, Health, Sports and Spiritual). There is a summary of the various models and their evaluation metrics on multi categories depicted in [Table 8] and [Table 9] with confusion matrix values as [Fig.3] and [Fig.4].

5. Conclusions

Now, the conclusion of this research paper denotes that the influence of stemmer is more effective on web page categorization on Gujarati language. The few classification reports of above selected supervised models are mentioned above with confusion matrix. These same techniques can apply to other languages using mapping of the language because most of the Indian languages are derived from the Indo-Aryan. The stemmer can implement using unsupervised model through machine learning. The researcher didn’t find any limitation of this proposed work but supposed to limited with features then to enhance this proposed techniques basis on positive feedback. The report of classification indicates Accuracy 75% to 97% for the WPC in Gujarati language web page. The confusion matrix indicates the error related to classification because stemming operation provides opportunities to achieve target output.

References

[1] C. D. Patel and J. M. Patel, “GUJSTER: a Rule based stemmer using Dictionary Approach,” IEEE - International Conference on Inventive Communication and Computational Technologies (ICICCT 2017) GUJSTER:, no. IEEE, pp. 496–499, 2017.
[2] D. Harman, “How effective is suffixing?,“ Journal of the American Society for Information Science, vol. 42, no. 1, pp. 7–15, 1991.
[3] P. Majumder, M. Mitra, S. Parui, and G. Kole, “YASS: Yet Another Suffix Stripper,” ACM Transaction of Information
Influence of GUJarati STEmmmeR in Supervised Learning of Web Page Categorization

Authors’ Profiles

Chandrakant D. Patel was born on May 19, 1981. He received the BCA and MCA degree from Hemchandracharya North Gujarat University, Patan in 2002 and 2005, respectively. He has 14+ years’ experience in Academic with UG and PG Courses. Prior to joining Acharya Motibhai Patel Institute of Computer Studies (PG Department), he worked with Shri C. J. Patel College of Computer Studies (UG Department), Visnagar from 2005 to 2011. His areas of interest include Web Page Categorization and Stemmer in Indian languages. He has also written a book for Hemchandracharya North Gujarat University BCA students titled “Computer Organization” in 2011. He has completed OP sponsored by UGC at ASC, Rajkot. He has 2 national and 6 international papers in his credit.

Dr. Jayeshkumar M. Patel, having rich experience of 16 Years in Academics, Industry, Research and International exposure, is holding Doctorate in ERP (Computer Science) from North Gujarat University. Rewarding his research work, he has been awarded “Career Award For Young Teachers” from AICTE, Delhi. He is working as a recognised Ph.D. guide at G.T.U., H. N. G.U, Ganpat University and also with many other reputed universities. He has good number of research under his name and presented more than 47 research papers in International and National Journals and Conferences. He has delivered number of expert talk in SANDHAN Programme and UGC Sponsored Programme. He is also the member of board of studies and selection committee of different universities.

How to cite this paper: Chandrakant D. Patel, Jayesh M. Patel, "Influence of GUJarati STEmmmeR in Supervised Learning of Web Page Categorization", International Journal of Intelligent Systems and Applications(IJISA), Vol.13, No.3, pp.23-34, 2021. DOI: 10.5815/ijitcs.2021.03.34