International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-2, July 2019

Impact of Stemming on Telugu Text Classification

Narla Swapna, Peneti Subhashini, B Padmaja Rani

Abstract: ***In Text categorization, Information retrieval and document clustering stemming is absolutely necessary especially for morphological rich languages like Indian. The process of stemming is, reducing the inflected or resultant terms to their stem word, root or origin form. However, stemming is a tricky task - particularly for extremely inflected natural languages having a lot of words for the same normalized word form. In Text classification, stemming tries to cut off details like either suffix or prefix of a word and produce basic word. In this paper, we apply various stemming methods on Telugu text classification and ensure the performance of the classifier is effect by stemming. Telugu is suffix oriented language, so we have performed number of experiments on erratically selected Telugu text documents and finally we conceive that the performance of the classifier is improved.***

Keywords: ***Information retrieval, Stemming, Text classification, Telugu***

I. INTRODUCTION

India is a multilingual country. The authorized and local languages of India play a vital role in communication with the citizens living in the country. In the Indian constitution, each and every state officially choose their local languages for communicate at the constitution level needs. The availability local languages constantly growing and amount of textual data in electronic form has rapidly increased. Searching of relevant data is fast when data is organized. So, Text categorization is a very important step. Various algorithms in the field of information Retrieval and Text categorization stressed upon word based representation models due to the fact that the document formation is naturally made up of a collection of words in the light of grammar rules. In Text categorization, most of the experiments are conducted with Word based stemming model. The Word based stemming model is often called as 'sent of words' approach. This stemming model is a language dependent. In general, a stem is a word and it may not be identical to the morphological variant of the word. For highly inflectional languages, stemming model pick up the performance for Text categorization [6]. M NarayanaSwamy et.al [2] has used decision tree, Naïve Byes and K-nearest neighbour classifiers in research. The Kavi Narayana Murthy [3] has discussed the many improvements in automatic text categorization of Indian languages.

Revised Manuscript Received on 30 July 2019.
***** Correspondence Author
Dr. Narla Swapna*, Department of CSE, CMR College of Engineering & Technology, Hyderabad, India
Dr. Peneti Subhashini, Department of CSE, MLRIT, Hyderabad, India
Dr. B Padmaja Rani, Department of CSE, JNTUH, Hyderabad, India

He noted that, Indian languages need large amount of corpora, high-quality of morphological analyzers and good stemmers are very essential to handle the affluence of morphology, particularly for the Dravidian languages.

II. STEMMING

Stemming is one of the most effectual procedures, which is essential in several applications, like document classification, information retrieval system, Natural language processing, machine learning and machine translation. By applying stemming technique during the categorization of text documents, will results in improves the performance of the classifier. Stemmers are classified in to language dependent stemmers and independent stemmers. A language dependent stemmer needs linguistic knowledge where, the independent stemmer does not require. N-gram is a language independent model. It is an alternative to word based models to identify the root word. Various language dependent and independent stemmers are proposed and discussed by Narla swapna, Dr Padmaja Rani [7] for Telugu root word identification. A rule based stemmer or statistical based stemmer can be applied on Telugu. Statistical based stemmers have very less significant for the languages which have lack of corpus. Corpus based stemming techniques are provided by kavi Narayanamurthi et.al [5]. Three different stemming techniques are discussed in his proposal. Yet another statistical trimmer was is also proposed by Dr K.V.N Sunitha, N. Kalyani[4] for Telugu language. They use an unsupervised method for some statistical analysis to trim a Telugu word. A rule based approach for A Telugu morphological analyzer was proposed by Uma maheswararao. G [8]. The proposed method is based on linguisticdatabase and analyze the every word irrespective of it's inflectional and derivational word. A TelStem was proposed by A.P. Siva Kumar, P. Premchand, A. Govardhan[1]. It is an unsupervised Telugu stemmer using Take-all-Splits heuristic and it is not a language specific.

III. TEXT CATEGORIZATION

Text categorization (TC) is the task of assigning a appropriate category to a text document. It has many of application. Traditionally, all incoming documents are examined and classified physically by particular subject expert based on the document contents. In the present Text Categorization prototype system, K-nearest neighbor classification algorithm has been applied. The KNN classifier is based on the hypothesis that the classification of an occurrence is most comparable to the classification of other occurrences that are very nearby in the vector space. When compared to the other text categorization methods, KNN classifier is simple, easy and it is computationally efficient.

It tends to perform extremely well with a lot of data points. The learning process cost is zero. It is more effective when the training data is very large.

The main computation is the arranging of training documents in order to find the *k* nearest neighbors for the test document. In K-Nearest Neighbor classifier, cosine similarity is the simplest approach for categorization. Three different assessment metrics precision (P), recall (R) and F-measure are used to evaluate the performance of KNN. These evaluation metrics proves to be efficient in binary classification in category assignment. Precision is a measure of exactness of the classifier and Recall is a completeness of classifier.

$$Precision\ (P) = \frac{TP}{(TP + FP)} \qquad (a)$$

$$Recall\ (R) = \frac{TP}{(TP + FN)} \qquad (b)$$

$$F\ Measure\ = \frac{2 * P * R}{(P + R)} \qquad (c)$$

IV. EXPERIMENTATIONS AND OUTCOMES

Different investigations were performed on Telugu Corpus by without stemming and with various stemming models [7]. Seven different categories of 1150 documents are taken to implement this work. Performance of the classifier is calculated and compared. The classifier performance of KNN without stemming is shown in Table I.

TABLE I: CLASSIFIER PERFORMANCE WITHOUT STEMMING

S.NO	DATA SET	PRECISION (%)	RECALL (%)	F Measure (%)
1	News /(వార్తలు) /Varthalu	40.9	92	56.6
2	Politics/ (రాజకీయాలు) /Rajakiyalu	68	8	14.3
3	Rivers /(నదులు)/Nadulu	66	14	23.1
4	Literature/ (సాహిత్యము) /Sahithyam	69	0	0
5	Songs/ (పాటలు)/(Patalu)	87	7	12.9
6	Sports/ (క్రీడలు)/kreedalu	100	8	14.8
7	Stories/ (కథలు)/Kathalu	51	8	13.8
Average value		**68.8**	**19.5**	**19.35**

In Table II, classifier performance of KNN with language dependent stemming is exposed.

TABLE II: CLASSIFIER PERFORMANCE WITH LANGUAGE DEPENDENT STEMMER

S.NO	DATA SET	PRECISION (%)	RECALL (%)	F Measure (%)
1	News /(వార్తలు) /Varthalu	40.3	92	56.04
2	Politics/ (రాజకీయాలు) /Rajakiyalu	85.7	48	61.53
3	Rivers /(నదులు)/Nadulu	86	52	64.81
4	Literature/ (సాహిత్యము) /Sahithyam	86	52	64.81
5	Songs/ (పాటలు)/(Patalu)	87	56	68.13
6	Sports/ (క్రీడలు)/kreedalu	66	64	64.98
7	Stories/ (కథలు)/Kathalu	55	64	59.15
Average value		**72.28**	**61.14**	**62.71**

The results of classifier performance with language independent stemming are found in Table III.

TABLE III: CLASSIFIER PERFORMANCE WITH LANGUAGE INDEPENDENT STEMMER

S.NO	DATA SET	PRECISION (%)	RECALL (%)	F Measure (%)
1	News /(వార్తలు) /Varthalu	46	92	61.33
2	Politics/ (రాజకీయాలు) /Rajakiyalu	77	68	72.22
3	Rivers /(నదులు)/Nadulu	94	64	76.15
4	Literature/ (సాహిత్యము) / Sahithyam	65.2	60	62.49
5	Songs/ (పాటలు)/(Patalu)	94	68	78.91
6	Sports/ (క్రీడలు)/kreedalu	77	68	72.22
7	Stories/ (కథలు)/Kathalu	82.6	76	79.16
Average value		**76.54**	**70.85**	**71.70**

In Table IV, results of KNN classifier performance with Hybrid stemming are shown.

TABLE IV: CLASSIFIER PERFORMANCE WITH HYBRID STEMMING

S.NO	DATA SET	PRECISION (%)	RECALL (%)	F Measure (%)
1	News /(వార్తలు) /Varthalu	62.6	92	74.50
2	Politics/ (రాజకీయాలు) /Rajakiyalu	79.1	76	77.51
3	Rivers /(నదులు)/Nadulu	87.5	84	85.7
4	Literature/ (సాహిత్యము) / Sahithyam	95.2	80	86.94
5	Songs/ (పాటలు)/(Patalu)	95.4	84	89.33
6	Sports/ (క్రీడలు)/kreedalu	90	80	84.70
7	Stories/ (కథలు)/Kathalu	83.3	80	81.61
Average value		**84.7**	**82.2**	**82.89**

An average classifier performance of the Telugu text documents by with various stemming models and without stemming is shown in Table V. The average precision, recall and F measures of with stemming models and without stemming is depicted in figure I. From the Table V, the highest precision, recall and F measure are found in case of Hybrid stemming. They are 84.7 %, 82.2 % and 82.89 % respectively.

It is also observed that, Hybrid stemming is found to be a good classifier performer than all the other applied stemming methods. In Hybrid stemming, the recall levels have increased by 62.7% when compared to classifier performance without stemming. It is also observed that Hybrid stemming mechanism performs very well when compared to other stemming methods like language dependent stemming and language independent stemming.

V. CONCLUSION

In this paper, various experimentations are carried out on Telugu text corpus, which is collected from different internet sources to evaluate the performance of the classifier. In order to assess the classifiers, the KNN method is used. The outcomes clearly showed the classifier performance is improved, that is 63.04% more with the use of the stemming.

We conclude that stemming is very essential and crucial step in Telugu Text categorization. The maximum classifier performance is observed as 82.89% with Hybrid stemming.

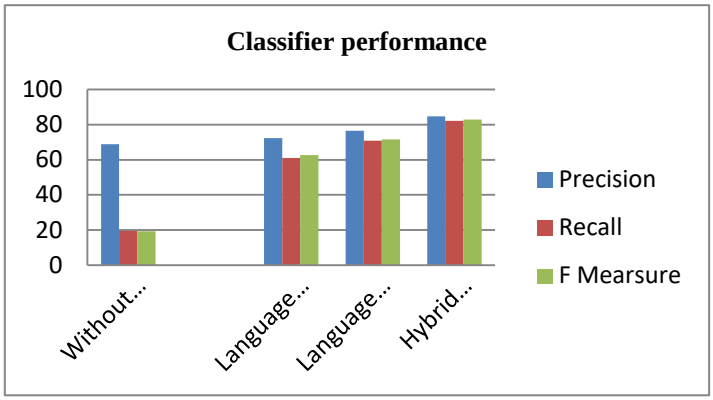

Figure I: Average Precision, Recall and F Measures without stemming and with stemming models

REFERENCES

1. Damashek, M., Gauging Similarity with n-grams: Language-Independent Categorization of Text, Science, Volume 267, February 1995.
2. Indian Language Text Representation and Categorization Using Supervised Learning Algorithm M NarayanaSwamy ,M. Hanumanthappa in ICICA- 2014.
3. KaviNarayana Murthy, "Advances in Automatic Text Categorization" Department of Computer and Information Science, University of Hyderabad.
4. KVN Sunitha, N.Kalyani, "YAST-Yet another statistical trimmer", published in the International Journal of Computer applications in Engg, Technology and Sciences Oct'2008, ISSN:0974-3596, pp:146-157.
5. M.Santosh Kumar, Kavi Narayan Murthy "*Corpus –Based Statistical Approches for Stemming Telugu*" in vishvabarath@tdil [70]
6. Natarajan A., Powell A.L., and French J.C. *Using N-grams to Process Hindi Queries with Transliteration Variations* Technical Report: CS-97-17,1997
7. Swapna Narala, B. Padmaja Rani,K. Ramakrishna, "Experiments in Telugu Language using Language Dependent and Independent Models", InternatIonal Journal of Computer Science and technology(IJCST) , Vol. 7, Issue 4, oct - Dec 2016, ISSN : 0976-8491 (online) | ISSN : 2229-4333 (print).
8. UMA MAHESWARA RAO, G. 1999. A Morphological Analyzer for Telugu (electronic form). Hyderabad: University of Hyderabad.

Table V: Average classifier performance with and without stemming

	Precision (%)	Recall (%)	F Measure (%)
I. Without stemming	68.8	19.5	19.35
II. With stemming			
i. Language Dependent Stemming	72.28	61.14	62.71
ii. Language Independent Stemming	76.54	70.85	71.7
iii. Hybrid stemming (Blended model)	84.7	82.2	82.89

AUTHORS PROFILE

Dr. Narla Swapna, working as an Associate professor in the Dept. of Computer Science & Engineering, CMR College of Engineering & Technology, Hyderabad. Dr. Narla Swapna obtained Ph.D from Jawaharlal Nehru Technological University, Hyderabad. She has 12 years of teaching experience and 14 publications in various National, International Conference and Journals. She held various positions like JKC core team member for nine years for 2006-2015, single point of contact (SPOC), NBA coordinator, TPO and in charge HOD. Her research interests include Machine Learning, Information Retrieval system and Natural Language Processing. She is also a Member of various Technical Associations including ISTE, IEEE etc.

Dr. Subhashini. P working as an Associate professor in the Dept. of Computer Science & Engineering, MLRIT, Hyderabad. Dr. P. Subhashini obtained B.Tech , M.Tech and Ph.D (full time) from Jawaharlal Nehru Technological University, Hyderabad. She held various position like R& D coordinator, JHUB funding member and worked as a website & publicity member for ICCII, an international Conference on Computational Intelligence and Informatics for three consecutive years for 2016-2019, Organized by Dept. of Computer Science & Engineering, JNTUCEH, Hyderabad. Her Research interests include Machine Learning, Information Security, Information Retrieval system, Natural Language Processing and Cloud computing. She has 15 publications in national and international journals.

Dr. B. Padmaja Rani working as Professor in the Department of Computer Science and Engineering and TEQIP-III coordinator, JNTUH College of Engineering, JNTU University Hyderabad. She obtained B.Tech from Osmania University, M.Tech and Ph.D. from JNT University, Hyderabad, India. She is having 24 years of experience in Industry and Academia. Her area of Research includes Information Retrieval, Data Mining, Machine Translation, Computer Networks, and Software Engineering etc. She is guiding 10 Research Scholars in the area of Information Retrieval, Natural language processing and Computer Networks. She has more than 65 publications in various International Journals and Conferences. She is a member of various advisory committees and Technical Bodies. She is also a Member of Various Technical Associations including ISTE, CSI, IEEE etc.