© 2017 IJSRST | Volume 3 | Issue 7 | Print ISSN: 2395-6011 | Online ISSN: 2395-602X
Themed Section: Science and Technology

A Study of Current State of Work done for Classification in Indian Languages

[1]Kaushika Pal, [2]Dr. Biraj V. Patel

[1]Assistant Professor, Sarvajanik College of Engineering and Technology, Surat, Gujarat, India

[2]G.H.Patel, P.G. Department of Computer Science & Technology, Sardar Patel University, V.V. Nagar, Gujarat, India

ABSTRACT

Classification has become an important aspect of study for storing, organizing and retrieving relevant document. So much work has been done in English language. Researchers have now started focusing on Indian language document classification as lot of content is available on web in Indian languages. The purpose of this paper is to study current work done in various Indian languages, and analyze the current situation and future scope to research in classification and related work on Indian languages.

Keywords: Classification, Machine Learning, Pre-processing, Data Mining, Natural Language Processing

I. INTRODUCTION

Classification is one of the most widely used techniques in Machine Learning. The fundamental aim of classification it to predict a class for some input. In current era lot much Hindi language content is being generated in digital form. Content such as blogs, product review, movie review, news articles, various categories of opinions, stories, etc. are increasing in volume gradually, which has given impetus to the techniques like data mining, Natural Language Processing and Machine Learning for automatic classification of documents. Document classification is one of the text mining tasks to manage the information efficiently, by classifying the documents into classes using classification.

Classifying documents involves Document Collection, Preprocessing, Feature Extraction, Feature Selection and Classification.

Pre-processing involves Tokenization, Stop-word removal, stemming.

Tokenization: tokenization is the process of breaking a stream of text into words, phrases, symbols, or other meaningful elements called tokens.

Stop-word removal: Stop words are words which are filtered out before or after processing of natural language data.

Stemming: The goal of both stemming and lemmatization is to reduce inflectional forms and sometimes derivationally related forms of a word to a common base form.

Feature Extraction aims at converting the document sentences into a set of words (i.e, Creating Document Vector) and, at the same time, enriching their semantic meaning. Subtasks involves are part of-speech tagging, meaningful word selection.

Feature Selection also known as variable subset selection or attribute selection is the process of detecting relevant features and removing irrelevant, redundant or noisy date.[21] Feature selection is accomplished by keeping the words with highest score according to predetermined measure of the importance of the word.

Classification: There are two classes of Machine Learning techniques: supervised and unsupervised. In supervised methods, a model is created based on a training set. Categories are predefined and documents within the training dataset are manually tagged with one or more category labels. A classifier is then trained on the dataset which means it can predict a new document’s category from then on. Depending on the classification algorithm or strategy used, the classifier might also provide a confidence measure to indicate how confident it is that the classification label is correct.

IJSRST17372 | Received : 15 Sep 2017 | Accepted : 30 Sep 2017 | September-October-2017 [(3) 7: 403-407]

In this paper classification done in Indian languages, techniques used and performance of various techniques are discussed.

II. Literature Review

Harikrishna D M, K. Sreenivasa Rao (2016). [1] worked on classifying Hindi Short Stories. The dataset consists of 780 Hindi emotional sentences collected from children stories. Stories are classified as Happy, Sad, Anger and Neutral. Feature Selection is done using POS feature, Emotion – Specific feature, POS + ESF. Classifiers used are Naive Bayes (NB), k-nearest neighbour (KNN) and support vector machine (SVM). Their experiment shows SVM models outperformed other models in terms of classification accuracy and feature selection POS + ESF is working better than other feature selection methods.

D M Harikrishna, K. Sreenivasa Rao (2015) [2] Document classification is done using data set is 300 Hindi short stories and classification has been done in Fable, Folk-tale, Legend. Pre-processing done is lemmatization (stemming). Feature selection is done using weighing schemes of keywords, POS density, and analysis of POS tags according to story genres. Classifiers used are Naive Bayes (NB), k-nearest neighbour (KNN) and support vector machine (SVM). Their experiment shows SVM models outperformed other models.

Harikrishna D M, K. Sreenivasa Rao (2015) [3] worked to classify 450 hindi and telugu short stories into Fable, Folk-tale, Legend genre. Pre-processing involves lemmatization (stemming) Stop-words removal. Feature selection is done using Baseline, POS Density, TF, TFIDF, TF+POS Density, TFIDF + POS Density. Performed experiment shows TF+POS combination is working better than other features selection. Classifiers used are Naive Bayes (NB), k-nearest neighbour (KNN) and support vector machine (SVM). It is found that SVM has the highest accuracy.

Harikrishna D M, Gurunath Reddy, K. Sreenivasa Rao (2015) [4] worked to classify 300 short hindi stories into Fable, Folk-tale, Legend genre and then each story genre into emotions like Happy, Anger, Sad and Fear. Feature selection is Baseline, POS Density, TF, TFIDF, TF+POS Density, TFIDF+POS Density for genre classification and POS and Sentence level features are considered for predicting emotions. TF+POS combination is working better than other features selection for genre classification. Classifiers used are Naive Bayes (NB), k-nearest neighbour (KNN) and support vector machine (SVM).

Megha Garg, Bhaskar Sinha, Somnath Chandra (2016) [5] presents SVM based approach for learning, classifying and automatically predicting relationships between Hindi synsets of IndoWordNet. The accuracy found using SVM is 71.87% which can be improved through introduction of language based knowledge.

Garima Nanda, Mohit Dua, Krishma Singla (2016) [6] are using term frequency and Naïve Baye’s Classifier to answer a question asked in hindi language. Worked for tokenizing the question. The system relocates the correct or closest results to the specific question.

Vandana Jha, Manjunath N (2015) [7] propose an hindi opinion mining system the data set used is hindi document containing movie review(200 documents 100 positive and 100 negative) feature used are Unigram, Best words, best words + bigram chi square features. Two approaches proposed: Supervised: Naïve Bayes classifier for machine learning and Unsupervised POS tagging in which considered adjectives as opinion words, where simulations confirms the effectiveness of the proposed approach. POS tagging is working better.

Sumitra Pundlik, Prachi Kasbekar (2016) [8] proposed a model for classification of hindi speech documents into multiple classes with the help of ontology creation. Which include sentiword classification of BOW into positive, negative or neutral using two approaches. 1) Finding Polarity of the word using HindiSentiWordNet (HSWN). 2) Combination of HSWN and Language Model (LM). Claimed that after implementation a hindi document can be classified successfully in multiple classes with the help of ontology if the text is related to ontology. Multiclass Analysis along with Sentiment Analysis is performed. Some of the document class considered like “उत्सव”, “खेती”, “संकट”, “समाज” , शिक्षा” etc.

Nidhi, Vishal Gupta (2012) [9] proposed two new algorithm for punjabi text classification; Ontology based classification and Hybrid approach (combination of naïve bayes and ontology based). Data set used is 180 punjabi text documents.45 files are used as Training Data. Training set contains total 3313 words which are used to train the punjabi text classifier based on Naïve Bayes and Centroid Based. Accuracy of Centroid Based Classification and Naïve Bayes Classification is 71%, Ontology Based Classification 64% , Hybrid approach's accuracy is 85%.

Nidhi, Vishal Gupta (2012) [10] proposed classification algorithm for punjabi text document and created sports based ontology for each class that consists of terms related to that class. Using ontology classification does not need training set or labeled documents.

Pratibha Singh, Ajay Verma, Narendra S. Chaudhari (2011) [11], Handwritten hindi numerals are classified. Binarization, noise removal, size normalization, skeltonization and width normalization are done and then SVM and MLP based classifier are used. SVM outperforms MLP based classifier.

Akanksha Gaur, Sunita Yadav (2015) [12], worked on Handwritten hindi character recognition. First binarization of the image and separations of characters are performed, horizontal bar removed and then SVM uses hyperplane for classification.

Richa Sharma, Shweta Nigam (2014) [13] surveyed on opinion mining in hindi language and found from last few years, enormous increase has been seen in hindi language on the web, researchers has performed opinion mining in hindi. Also discussed techniques and several challenges of hindi based opinion mining. The nature of Indian languages varies a great deal in terms of script, representation level and linguistic characteristics. To understand the behavior of Indian Languages, large amount of work needs to be done in that field.

Pooja Pandey, Sharvari Govilkar (2015) [14] worked on sentiment analysis in hindi using HSWN and improved exiting HSWN by adding missing sentimental words related to Hindi movie domain. Sentiments can be mined from various sources like texts, tweets, news articles, comments, blogs, social media or any other source.

Upendra Mishra, Chandra Prakash (2012) [15] proposed a stemmer for hindi language. Stemming can be used to improve the effectiveness of information retrieval.

Dr. Hanumanthappa, Narayana Swamy.M(2015) [16] have done a detailed study on text mining in Indian languages. Also discussed the need of text mining in Indian languages and made an attempt to propose techniques for Indian Languages Text Mining.

Upendra Singh, Saqib Hasan(2015) [18] surveyed on document classification and classifiers and specified an insight into text classification process, its phases and various classifiers. It's aimed for comparing and contrasting various available classifiers on the basis of few criteria like time complexity and performance. Performance of different algorithms varies according to data collection. SVM with term weighted VSM representation has shown some potential results in the task of text classification up to some extent but claims that universal acceptance of this method remains implausible. Survey was carried on K Nearest Neighbor, Support Vector Machine, Naïve Bayes, Neural Networks, Rocchio's algorithm.

Jasleen Kaur, JatinderKumar Saini (2015) [19] studied of text classification Natural Language Processing Algorithms for Indian Languages. And analyzed text classification on Indo-Aryan, Indo Dravidian, Sino - Tibeto Indian Languages. Also found very few work is done on text classification in Indian Languages and study shows that supervised approach is working well for Indian languages. But Indian languages still need to be explored in terms of text classification.

Yakshi Sharma, Veenu Mangat (2015) [20] proposed an algorithm which uses subjective lexicon method. SentiWordnet were created which contains adjectives and adverbs for a particular reason and assigned polarity to all adjectives and adverbs in the wordnet, then sum up the positive negative polarity for a tweet. Choose the dominating one.

Senthil Kumar B, Bhabitha Varma E (2016) [21] discussed several approaches of text categorization, feature selection methods and applications of text categorization. For Feature selection discussed Embedded Method, Wrapper Method, and Filter Method.

For Feature Selection Singular Value decomposition, Principal Component Analysis, Independent component analysis, Canonical Correlation Analysis, Locally Linear Embedding and Linear Discriminant Analysis, Overview of Machine learning techniques such as Bayesian Classifier, Neural Network Classifier, Support Vector machine, Decision Tree, K – nearest Neighbor are also discussed.

Pooja Bolaj, Shavari Goilkar(2016) [22] worked on Text Categorization techniques for Indian Regional Languages. They discussed various categorization techniques like Decision tree for Bangla, K-nearest Neighbor for Bangla, Telugu and Marathi, Centroid Algorithm for Punjabi, Support Vector machine for Bangla and Urdu, Naïve Bayes for Bangla, Punjabi, Urdu, Telugu and Marathi, Neural Networks for Tamil languages. From the study it was observed that three supervised learning methods Support Vector Machine, Naïve Bayes and K – Nearest Neighbor are most suitable and better results for document classification for Indian regional languages.

III. Conclusion

Carrying classification task on Indian languages is challengeable due to morphological variance. Although many researchers are working on classification on various Indian languages, but much yet has to explored and worked. From the study it is observed that certain work has been proposed but not implemented, certain work is done but on very small scale. Researchers have huge scope of classifying documents in Indian languages. The major challenge in classifying document of Indian languages is that the languages are morphological variant in nature and therefore requires different.

IV. REFERENCES

[1]. Harikrishnna D M, K. Sreenivasa Rao (2016). Emotion-Specific Features for Classifying Emotions in Story Text. IEEE 2016 22nd National Conference on Communication (NCC).

[2]. Harikrishnna D M, K. Sreenivasa Rao (2015). Classification of Children Stories in Hindi Using Keywords and POS Density. In Proceedings of IEEE International Conference on Computer, Communication and Control.

[3]. Harikrishnna D M, K. Sreenivasa Rao (2015). Children Story Classification based on Structure of the Story. IEEE International Conference on Advances in Computing, Communications and Informatics. (pp. 1485-1490).

[4]. Harikrishnna D M, K. Sreenivasa Rao (2015). Multi-stage Children Story Speech Synthesis for Hindi. IEEE 2015 8th International Conference on Contemporary Computing (IC3).

[5]. Megha Garg, Bhaskar Sinha (2015). Identification of Relations from IndoWordNet for Indian Languages using Support Vector Machine. IEEE International Conference on Computing and Network Communications. (pp. 547-552).

[6]. Garima Nanda, Mohit Dua(2016). A Hindi Question Answering System using Machine Learning Approach. IEEE International Conference on Computational Techniques in Information and Communication Technologies.

[7]. Vandana Jha, Manjunath N (2015). HOMS: Hindi Opinion Mining System. 2015 IEEE 2nd International Conference on Recent Trends in Information Systems. (pp. 366-371).

[8]. Sumitra Pundlik, Prachi Kasbekar(2016). Multiclass Classification and Class based Sentiment Analysis for Hindi Language. IEEE International Conference on Advances in Computing, Communications and Informatics (sept 21-24, 2016) (pp. 512-518).

[9]. Nidhi,Vishal Gupta(2012). Domain Based Classification of Punjabi Text Documents using Ontology and Hybrid Based Approach. Proceedings of the 3rd Workshop on South and Southeast Asian Natural Language Processing. (pp. 109-122).

[10]. Nidhi,Vishal Gupta(2012). Algorithm for Punjabi Text Classification. International Journal of Computer Applications. Volume 37 - No. 11. (pp 30-35)

[11]. Pratibha Singh, Ajay Verma, Narendra S. Chaudhari. Performance Analysis of Flexible zone based features to classify Hindi numerals. 2011 3rd International Conference on Electronics Computer Technology. (pp. 292 – 296).

[12]. Akanksha Gaur, Sunita Yadav. Handwritten Hindi Character Recognition using K- Means Clustering

and SVM. 4th International Symposium on Emerging Trends and Technology in Libraries and Information Services.

[13]. Richa Sharma, Shweta Nigam (2014). Opinion Mining In Hindi Language: A Survey. International Journal in Foundation of Computer Science & Technology. Vol. 4. No. 2. (pp. 41-47)

[14]. Pooja Pandey, Sharvari Govilkar (2015). A Framework for Sentiment Analysis in Hindi using HSWN. International Journal of Computer Applications. Volume 119 No. 19. (pp. 23 - 26).

[15]. Upendra Mishra, Chandra Prakash (2012). MAULIK: An Effective Stemmer for Hindi Language. International Journal on Computer Science and Engineering. Vol. 4 No. 5. (pp. 711-717).

[16]. Dr. Hanumanthappa, Narayana Swamy.M (2015). Indian Language Text Mining. Journal of Software Engineering and Simulation. Vol. 2 Issue 10. (pp. 01-04).

[17]. Piyush Arora (2013). Sentiment analysis for Hindi Language. (MS by Research Thesis). International Institute of Information Technology, Hyderabad.

[18]. Uprendra Singh, Saqib Hasan(2015). Survey Paper on Document Classification and Classifiers. In International Journal of Computer Science and Technology - Volum 3 Issue 2.(pp 83-87).

[19]. Jasleen Kaur, Jatinderkumar Saini (2015). A study of Text Classification Natural Language Processing Algorithms for Indian Languages. VNSGU journal of Science and Technology. Volume 4. No.1 (pp. 162 - 167)’

[20]. Yakshi Sharma,Veenu Mangat(2015). A Practical Approach to Sentiment Analysis of Hindi Tweets. International Conference on Next Generation Computing Technologies.

[21]. Senthil Kumar B, Bhabitha Varma E (2016). A survey on Text Categorization. International Journal of Advance Research in Computer and Communication Engineering. Vol. 5, Issue8. (pp. 286-289)

[22]. Pooja Bolaj, Sharvan Govilkar(2016). A Survey on Text Categorization Techniques for Indian Regional Languages. International Journal of Computer Science and Information Technologies. Vol. 7(2). (pp. 480-483)

[23]. Pooja Pandey and Sharvari Govilkar(2015). A survey of Sentiment Classification Techniques used for Indian Regional Languages. International Journal on Computational Science & Applications. Vol. 5, No. 2 (pp. 13-26)

[24]. Hanumanthappa, Narayan Swamy M (2016). Indian Languages Text documents Categorization and Keyword Extraction. International Journal of computer Technology and Applications. Vol.9(3) (pp. 37-45)

Authors :

Kaushika Pal is MCA from VNSGU, Surat, Gujarat, India. Currently she is Assistant Professor at Sarvajanik College of Engineering and Technology, Surat. She has several research papers in National and International journals. Her research areas are Data Mining, Big data, Software Testing, Software engineering.

Dr. Biraj V. Patel is Ph.D(Computer Science) degree from the Sardar Patel University, V.V nagar, Gujarat, India in the year 2014. He has joined as a lecturer in 2008, at department of computer science & Technology, Sardar Patel University. His are of interest is SEO, Data Warehousing and data mining.