International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 07 | July 2020 www.irjet.net p-ISSN: 2395-0072

Hybrid Machine Learning based Kannada Next Word Prediction

Nandini B R[1], Prof. Hamsaveni M[2], Prof. Charunayana V[3]

[1,2,3]*Assistant Professor, Dept. of CSE Vidyavardhaka College of Engineering*

ABSTRACT: *Natural language Processing is one among the most trending technology in the era of AI products. There has been lots of techniques involved starting from rule-based approach to complex deep learning-based approach. Advancement of technology has made the Natural language processing for South Indian languages several techniques such as naive Bayes, k-Nearest- Neighbour and decision tree algorithms have been used for many NLP tasks like Entity resolution, categorization. In this paper next word prediction system has implemented for regional language Kannada, Karnataka state's spoken language. The objective is to present combination of Naïve Bayes and latent semantic analysis will work equally well compared to complex technique with minimal resource usage for mid-scale data by optimizing it using stochastic gradient.*

Keywords: Indic-NLP, Word Prediction, Naïve Bayes, LSI, Stochastic Gradient Descent

1. INTRODUCTION

In present days, there is an expanding request by disabled individuals to utilize PC programs, for the most part for social Interactions. The vast majority of these Interactions are made by Chat messages, which makes a difficult task for physical impeded individuals to speak with others. Along these lines, one method which treats this difficult that got consideration by scientists is called word Prediction. Word Prediction is a word processing feature highlight that points diminish the quantity of keystrokes [1] essential for composing words. For the most part, these models foresee the following word given a lot of words dependent on a unique circumstance. Hence, the Natural Language Processing (NLP) tasks, which performs errands such word forecast through comprehension and understanding of writings has got well known.

The Naive Bayes is a methodology that assumes the conditional independence of its variables. It requires less resource and processing time than approaches that Consider dependency Be that as it may, the Naive Bayes exactness will in general be most noticeably awful than those as a result of its information misfortune during the independence suspicion. The NLP area despite everything has investigates for an exact strategy that could be executed and anticipate in an appropriate time. In this unique circumstance, the idle semantic examination (LSA) was made. This procedure is utilized to semantically investigate messages through the connection between the words in various content levels, as expressions, sections, among others. Other than the high exactness of the LSA procedure, its deductions despite everything depend just on the content recurrence, which implies that it doesn't consider word orders and thusly the content sentence structure.

Although Deep learning approaches achieves a great precision in text generation tasks, however, it has some issues. The computational cost and time to train the amount of data needed is higher compared to other approaches. On top of that, learned knowledge is in the weights of neurons connections, also not possible to interpret what was learned, which means that it is black box

This paper proposes a new hybrid model to predict words using two well-known models: the Naive Bayes and the LSA. In addition, to this we optimize parameters used to improve the prediction precision through the Stochastic Gradient Descent technique.

Related work

In an investigation of predicting words [2] specialist built up a sentence completion technique dependent on N-gram language models and they inferred a k best Viterbi bar scan decoder for firmly finishing a sentence. We likewise watched utilization of Artificial Intelligence [3] for word forecast. Here syntactic and semantic examination is finished utilizing the diagram base up method for word predict. Another analyst recommends a methodology [4] of word predict by means of a Clustered Optimal Binary Search Tree. They recommend utilizing a system network to assemble ideal parallel pursuit tree which likewise contain additional connection with the goal that bigram and the trigram of the language likewise introduced to accomplish ideal execution of word expectation. Here the scientist does a lexical examination of most plausible showed up word in input text.

In a paper, classification based approach word prediction [5] has presented as an effective technique of word prediction using machine learning algorithm new feature extraction and selection techniques adapted from information gain or mutual information (IG/MI) and Chi square (X2). Some researchers use Ngram language model for word completion of Urdu words [6] and also Hindi words [7] for detecting disambiguation in Hindi word. There are some related works also on Bangla language using N-gram language model such as grammar checker of Bangla language [8], checking the correctness of Bangla

word [9] and verification of Bangla sentence structure [10].

There are various word prediction tools that we experience in real times, such as AutoComplete application by Microsoft, AutoFill application by Google Chrome, Typing Aid [11], LetMeType [12] etc

In our paper we have proposed a combination of Naïve Bayes and LSI for training and stochastic gradient descent for optimization.

Methodology

Our approach starts with data collection for word prediction and data transformation. Followed by pre-processing and training. We will go through each section one by one in details

Data Collection and Analysis: Data collection has done by scrapping the kannada tourism website.

Word cloud representation of dataset can be seen in fig 1

As per Mallamma V. Reddy[13] , Kannada is a morphologically rich language wherein morphemes consolidate with the root words as postfixes.

Kannada grammarians partition the expressions of the language into three classifications specifically: ***Declinable words*****:** Morphology of declinable words, as observed in numerous Dravidian dialects is genuinely basic contrasted with action words. Kannada words are of three sexes and furthermore declinable and conjugable words have two numbers-particular and plural.

Conjugable words**:** The action word is considerably more intricate than the things. There are three people specifically first, second and third individual. Tense of action words is past, present or future. Angle might be basic, persistent or great.

Uninflected words: Uninflected words might be delegated intensifiers, postpositions, conjunctions and interpositions. In Kannada, nearby words are regularly joined and articulated as single word which is called Morphophonemics or ?? Composition or ಸ?ಸ is where at least two words consolidate together to frame a general new word which protects the significance of the mix of the words.

Hence preprocessing technique will be slightly different

Fig 1: Word cloud Representation of Dataset

Data Preparation: As we are using supervised learning method, it will be having input pattern and its output label, i.e. for this word prediction problem word to be predicted will be output label. The model is initially fit on a training dataset

After Dataset is prepared entire process of proposed system has shown below.

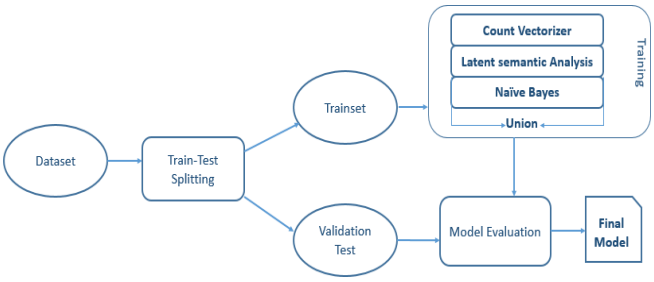

Fig 2: Proposed System

Preprocessing: It include remove unwanted symbols and tokenization. Too much of standardization of words has not done to avoid grammatical error. After pre-processing word tokenization has done to build vector space.

Training: To perform inferences, it is necessary to create and train the LSA and Naive Bayes sub- models. These sub models are trained in order to represent two types of information. The former, Naive Bayes, is used to store co-occurrences word pattern; the later, LSA, is used to obtain the semantic behavior of pattern

Naïve Bayes Algorithm is a probabilistic machine learning method which is mainly used for classification task. It assumes independence of variable

It works based on Bayesian inference given below.

$$P(\text{ಪ್ರವಾಸಿಗರಿಗೆ} \mid \text{ಕರ್ನಾಟಕ ಪ್ರವಾಸೋದ್ಯಮವು}) = \frac{P(\text{ಕರ್ನಾಟಕ ಪ್ರವಾಸೋದ್ಯಮವು} \mid \text{ಪ್ರವಾಸಿಗರಿಗೆ})\ P(\text{ಪ್ರವಾಸಿಗರಿಗೆ})}{P(\text{ಕರ್ನಾಟಕ ಪ್ರವಾಸೋದ್ಯಮವು})}$$

Where $P(\text{ಪ್ರವಾಸಿಗರಿಗೆ})$ is given by Number of occurrences of ? ?ಗ?? in trainset to number of words in trainset

LSA learns latent topics by doing matrix decomposition on the vector space using Singular value decomposition. It is typically using as a dimension reduction technique.

Optimization

Stochastic gradient descent is a very popular and common algorithm optimization Machine Learning algorithms. It performs one updation of randomly at a time and find local minima faster compared to other methods

Results

Experiment has conducted for scraped data around 10000 phrases with 5 SGD iteration with 50 epochs. Accuracy and log loss have been used as performance metric. Accuracy obtained for final iteration is 0.62 (62%) with log loss 0.01400.

Below is the change in loss for each step by setting change in iteration 5. It can be observed that los converges for every 9th epoch.

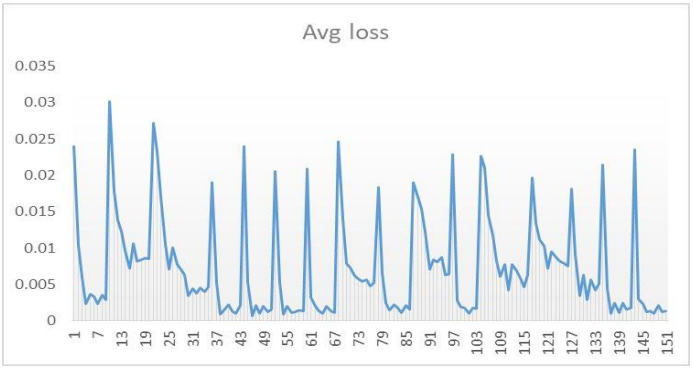

Fig 3 Loss per Steps

There is an enormous Growth of language model for NLP tasks from n-gram to now a day's hot topic Attention and transformers. Comparing the word prediction model for the data set considered with predefined models like :

- Bidirectional Encoder Representations from Transformers is a technique for NLP pre- training developed by Google (BERT)
- Universal Language Model Fine-tuning (ULMFIT) Accuracy of proposed model is almost equal to other existing model using less amount of resource. In Below figure system usage refer to System load.

System load refers to RAM and CPU usage and also our word prediction model is easily interpretable

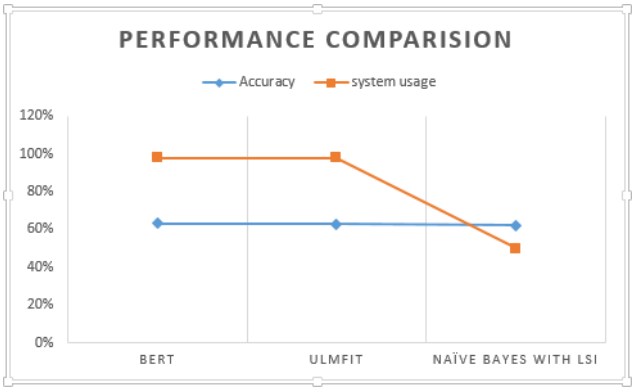

Fig 4 Performance Comparison Graph

Conclusion

Next word prediction has been implemented using combination of Naïve Bayes and LSI for Kannada language. Model has trained with various patterns generated from combination of bigram, trigram and 4-gram to yield better accuracy. Also, stochastic gradient descent has been used for model optimization. Experimentation has done with different learning rate. Final model will yield accuracy of 60-70% accuracy. As part of enhancement character ngram in the vector

Space model will still improve the model performance.

References

[1] C. Aliprandi, N. Carmignani, N. Deha, P. Mancarella, and M. Rubino, "Advances in nlp applied to word prediction," 2008.

[2] Steffen Bickel, Peter Haider and Tobais Scheffer, (2005), "Predicting Sentences using N- Gram Language Models," In Proceedings of Conference on Empirical Methods in Natural language Processing.

[3] Nestor Garay-Vitoria and Julio Gonzalez- Abascal, (2005), "Application of Artificial Intelligence Methods in a Word-Prediction Aid," Laboratory of Human-Computer Interaction for Special Needs.

[4] Eyas El-Qawasmeh, (2004), "Word Prediction via a Clustered Optimal Binary Search Tree," International Arab Journal of Information Technology, Vol. 1, No. 1.

[5] Hisham Al-Mubaid, (2007), "A Learning-Classification Based Approach for Word Prediction", The International Arab Journal of Information Technology, Vol. 4, No. 3.

[6] Qaiser Abbas, (2014), "A Stochastic Prediction Interface for Urdu", Intelligent Systems and Applications, Vol.7, No.1, pp 94-100.

[7] Umrinder Pal Singh, Vishal Goyal and Anisha Rani, (2014), "Disambiguating Hindi Words Using NGram Smoothing Models", International Journal of Engineering Sciences, Vol.10, Issue June, pp 2629.

[8] Jahangir Alam, Naushad Uzzaman and Mumit khan, (2006), "N-gram based Statistical Grammar Checker for Bangla and English", In Proceedings of International Conference on Computer and Information Technology.

[9] Nur Hossain Khan, Gonesh Chandra Saha, Bappa Sarker and Md. Habibur Rahman, (2014), "Checking the Correctness of Bangla Words using N-Gram", International Journal of Computer Application, Vol. 89, No. 11. ss

[10] Nur Hossain Khan, Md. Farukuzzaman Khan, Md. Mojahidul Islam, Md. Habibur Rahman and Bappa Sarker, (2014), "Verification of Bangla Sentence Structure using N-Gram," Global Journal of Computer Science and Technology, vol.14, issue-1 .

[11] Typing aid Auto completion utility downloadable at

https://autohotkey.com/board/topic/49517-ahk-11typingaid-v2220-word-autocompletion-utility/

[12] Handy free windows program https://letmetype.en.softonic.com/

[13] Mallamma V. Reddy and M. Hanumanthappa, Indic Language Machine Translation Tool for NLP.