International Journal of Advances in Electrical Engineering

E-ISSN: 2708-4582
P-ISSN: 2708-4574
IJAEE 2023; 4(1): 18-22

www.electricaltechjournal.com
Received: 20-10-2022
Accepted: 22-11-2022

Nagesh Raykar
Shri Jagdishprasad Jhabarmal Tibrewala University, Jhunjhunūn, Rajasthan, India

Prashant Kumbharkar
JSPMs Rajarshi Shahu College of Engineering, Pune, Maharashtra, India

Dand Hiren Jayatilal
Mulund College of Commerce, Mulund West, Mumbai, Maharashtra, India

Correspondence
Nagesh Raykar
Shri Jagdishprasad Jhabarmal Tibrewala University, Jhunjhunūn, Rajasthan, India

Hybrid LSTM technique for phonetic

Nagesh Raykar, Prashant Kumbharkar and Dand Hiren Jayatilal

DOI: https://doi.org/10.22271/27084574.2023.v4.i1a.33

Abstract
Demographic Details is empirically obtained data depicting different elements of a population. Various phonetic-based retrieving strategies are utilized, yet they are useless when trying to appeal to Indian name lists. The purpose of this research is to find Indian names with identical accents but distinct spellings based on data entries acquired from demographic databases. This study document contains the phonetic-based pronunciation technique, Long Short Term Memory, and K-Mean method Hybrid LSTM approach to identify Indian names using databases. The suggested Hybrid LSTM method is compared to a recurrent unit GRU-based technique in order to assess the precision of name prediction for Indian names. The hybrid LSTM model is produced by combining the K-Mean technique and the LSTM technique. When obtaining entries of Indian Names from sources, the constructed LSTM approach provides improved precision. If, as is the case in this instance, the demographic data collected from Indian Names from demographic sources has inaccurate data due to misspelled names, the provided method can be utilized to decrease data redundancy and acquire accurate data.

Keywords: LSTM, K-mean, GRU, demographic data

1. Introduction
Demographic details consists of quantitative information representing many aspects of a population. It is often collected for investigation reasons, product inspection, and in governmental and associated individual services. It might be challenging to obtain names from demographic information, especially for Indian names. When extracting Indian names from precise demographic data, phonetically-based pronunciation methods provide superior results. The method, which is phonetically dependent on pronunciation, was created in English predominantly. Existing phonetic-based techniques are unsuitable for Indian names, and even more so for English-phonetically similar names. The situation becomes much more problematic when distinct Indian names in local dialects are examined. De-duplication was necessary following the selection of related names. Existing phonetic-based name retrieval and name De-duplication methods are various [1, 2]. Soundex, Daitch-Mokotoff Soundex, Cologne Phonetic, Double Metaphone, Caverphone, Metaphone, and Match Rating Approach are indeed the phonetic-based techniques now in use. Those methods are employed to extract data through phonetic inputs [3]. These techniques retrieve information from phonetic inputs. In this circumstance, deduplication and precise data retrieval will be vital.

2. Literature Survey
To develop successful approaches for a machine learning -based method, we evaluated the research papers in this part. The primary prerequisite for a problem formulation and resolution was identified with the aid of a review of the relevant literature. The author will consider the research investigation as both conceptual and experimental. This paper illustrates the authors' coherent de-duplication technique for data de-duplication [4]. In the authors explain effective deduplication approaches for data deduplication. It is said that Cloud storage data can be employed to mitigate the difficulties associated with the deduplication method. The authors also discuss the difficulties and security concerns that arise throughout the deduplication technique. In, the authors present a novel technique for deduplication Indian names. It is argued that detecting identical records is difficult in the absence of domain expertise. Additionally, domain-specific approaches only apply to typical domains. In [5], the authors argue for a phonetically-based data extraction technique.

In a real-world setting, incorrect data entry can make it difficult to get accurate data. Consequently, both a phonetic matching method and a string matching method are required. The authors of [6] assert that India is a multilingual nation with significant demographic variances. They argue that an India-centric concept for the deduplication of structured data in Indian languages is required. Using an adjustable and ascending deep-learning technique, the research provides a novel strategy that is tailored to India-specific demographic data, particularly region-wise names, and addresses. It really is asserted that storing data on the cloud can reduce the challenges associated with the de-duplication approach. The researchers also cover the problems and security considerations that arise even during de-duplication process. The researchers of created a novel technique for deduplication Indian Demographic Names [7]. In a real-world scenario, erroneous data entering can make it challenging to obtain the data appropriately. In additional to a string-matching technique, a phonetic matching method is necessary for the optimum matching strategy. The authors of [8] assert that India is a multilingual society with numerous demographic differences. They suggest that it is vital to develop an India-centric working methodology for deduplication structured data in Indian languages. In addition to a string-matching method, a phonetic matching approach is used to determine the ideal method.

The study uncovered a novel solution that applies a scalable and adaptive deep-learning strategy to handle India-specific names, addresses, and demographic information. In Gated Recurrent Unit (GRU) paper author explained the how GRU will be helpful for the emotion based classification. This article describes the LSTM (Long Short Term Memory). This is used for varnishing gradient problem & How the LSTM solves the complex issues. This article provides an exhaustive examination of existing LSTM cell variations and network designs for prediction of time series. Suggested is an LSTM categorizing with optimized cell state descriptions and LSTM with interaction cell states [9]. Long Short Term Memory also resolves complicated, artificial long-time-lag challenges that past recurrent network techniques have not been able to accomplish [10]. The article addresses the standard k-means clustering approach and evaluates the deficiencies of standard k-means technique, for instance the k-means clustering method requirement to determine the distance between both data object and all cluster centers within every iterative process, resulting in a low clustering effectiveness. This study presents an enhanced k-means algorithm to handle this problem, which requires a basic data structure for storing some information in each repetition for use in the following iteration. The enhanced method avoids repeatedly calculating the distance between each data object and the cluster centers, minimizing processing time. K-means' computing complexity is reduced as a result of the revised approach, as demonstrated by experimental data [11].

3. Existing Approach

3.1 Machine Learning Scheme

The Machine Leaning -based approach is frequently used to obtain the exact fit from of the demographic data and also to carry out phonetic-based De-duplication. In machine learning, the training and testing information are crucial for figuring out what the data mean and what they might mean in the future. Figure 1 shows how the Machine Learning technique works in general.

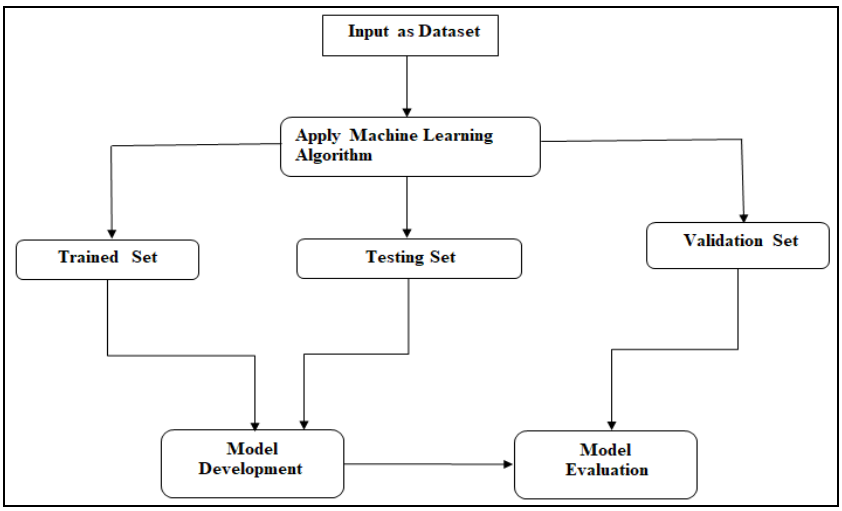

Fig 1: Machine Learning Strategy

3.2 Existing Parallel ML Method for Gated Recurrent Unit

Using the Gate Rate Unit (GRU) method, the problem with the varnishing gradient is fixed. This method is like the LSTM method. The GRU's main differences are: Unlike LSTM, it only has three gates and doesn't keep track of the state of the cells inside. The GRU includes Update Gate, Reset Gate, as well as the Existing Memory Gate [12].

The first Name and last Name were taken from the demographic information using the GRU method. When determining the exact match of a specified first Name and last Name, phonetic pronunciation are taken into account.

The list of the steps that the GRU Method takes as below:

3.2.1 GRU Algorithm

The way the GRU works is very equivalent to how the LSTM works. The reset gate as well as the update gate are what make up the GRU. LSTM, on the other hand, is made

up of the input, output, as well as forgets gate. The reset gate shows how and where to connect the new input to the memory that came before it. This same GRU Algorithm is utilized to solve the RNN's "vanishing gradient" issue.

1. The data has been put into groups based on how similar the sounds are.
2. The data is split into sets called Training, Validation, and Testing.
3. The input is sent to the Input Gate.
4. The information is then sent to the Update Gate.
5. Update Gate figures out Zt for timestamp t to figure out how much information from earlier is to be carried forward. The equation for this is given in Equation- (1)

$$z_t = \sigma(W^{(z)}x_t + U^{(z)}h_{t-1}) \quad (1)$$

6. At the Reset Gate, calculations are done to figure out how much information from before should be forgotten. The Reset Gate is found by using Equation (2).

$$r_t = \sigma(W^{(r)}x_t + U^{(r)}h_{t-1}) \quad (2)$$

7. This step is where you figure out the final output. You get a new memory content that keeps the data you've previously stored by using the Reset Gate. It is worked out like this, as shown in Equation (3):

$$h_t' = \tanh(Wx_t + r_t \odot Uh_{t-1}) \quad (3)$$

8. The last step is to figure out ht. It is a piece of information about the current unit that is stored in a vector. It sends the message to the network. The update gate decides which parts of the current memory (h't) and the memory from the previous stage (ht-1) will be read. The equation that goes with it is given in Equation- (4).

$$h_t = z_t \odot h_{t-1} + (1 - z_t) \odot h_t' \quad (4)$$

3.3 Data Extraction & Data Deduplication process using GRU

3.3.1 Generic Based Deduction Rule: It looks at how similar the words sound, which cuts down the number of names by a lot. It's a way to get rid of possibilities and narrow down the search area. As an example, "Ashish," "Asish," and "Asheesh" can all be shortened to "Asis." Based on the Phonetic reduction law, a number of rules have been set up to turn a name string into a common name string.

3.3.2 Database creation

It is similar to a database in that it was made by putting together random firstName and lastName pairs. Many different name strings can be turned into a unique generic name. The process of turning data into a series of ones and zeros. Then, the K-Mean method is used to make a cluster.

3.3.3 Matching Technique

The Edit Distance method is used to find the ED between two strings with as few steps as possible. To get the normalized distance, the string is split by up to two normal strings. The usual string affiliated with the bins whose standard ED form of the query data is less than the allowed limit is thought to be the likely cause of the duplication.

3.3.4 Working Flow of Existing Methodology

Methods for using the GRU algorithm are shown in Figure 3.

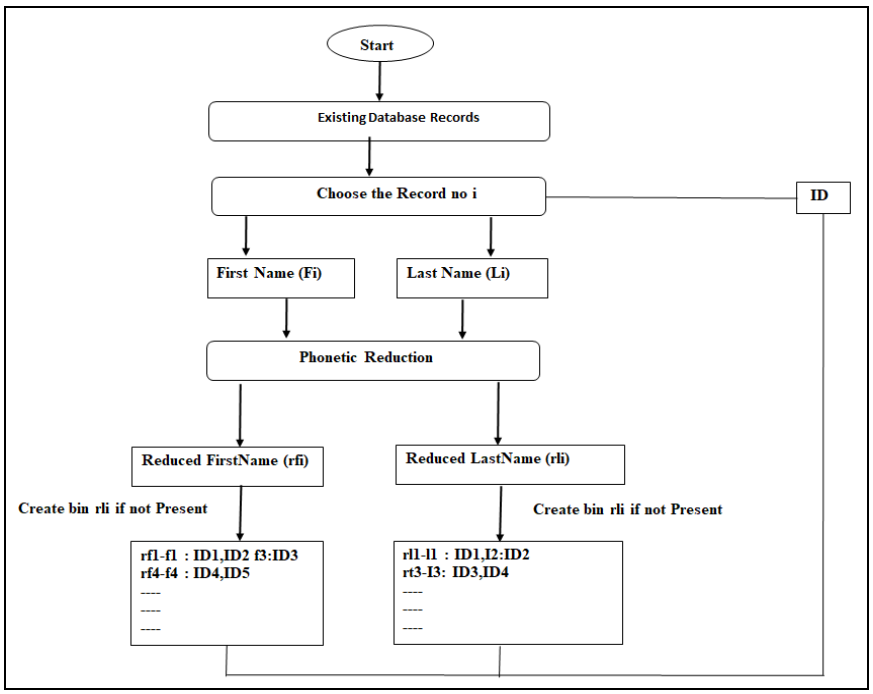

Fig 2: Working Flow of GRU algorithm

4. Proposed Methodology

The Hybrid LSTM Based approach developed for the Suggested approach. Getting accurate information from the Demographic database is very important for making decisions. When methods use character matching, it's hard to pull out names because names are influenced by the languages of different areas. In India, there are a lot of different local languages and dialects. A phonetic-based construction method that uses LSTM and K-Mean to find the exact first Name and last Name pair has indeed been described.

4.1 This is a step-by-step guide to the suggested technique

- The making of first Name and last Name pairs of Indian names for use in demographic data. It has names randomly selected.
- A data set has been split into three sets: Training (40%), Verification (20%), and Running tests (30%).
- The Training data have been grouped using K-Mean. Input data clustering is utilized to train the LSTM Model.
- The LSTM receives a three-dimensional array as input. Each cluster of training data will contain a memory-resident copy of this LSTM model.
- The real input values are evaluated with the aid of a categorized model. We will interpret the output. During the conversion method, these encoded keys are translated into the three-dimensional array of strings.
- For instance – A string "NATA" will initially be derived from the word "Neeta." The text is subsequently encoded and displayed.
- Determine the Cluster closest to the binary format: After getting the encrypted pair, the Classify method will identify the nearest matching group.
- Load into to the LSTM model the nearby cluster: The Cm LSTM Model is preparing it for transmission. This model utilized inputs from the stored LSTM Model. Once the comparison is complete, the subsequent phase will be invoked and the five closest encoded texts will be retrieved. The 5 Most Recent Encoded Texts Are Listed Here: 12. This step extracts the top 5 matches from the encoded results displayed to the user.
- Retrieve Name from Worksheet Actual names are stored alongside every possible pair in the worksheet.
- Create a name array containing the top K matches. Figure 4 depicts the suggested Assembled computation consecutive execution.

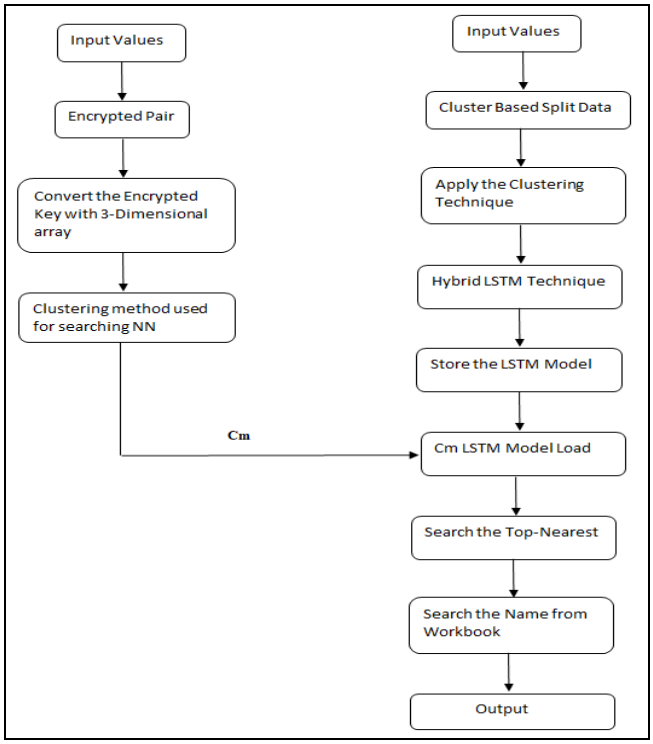

Fig 3: Execution in stages for the Hybrid Algorithm.

5. Experimental Results & Analysis

The results of the proposed constructed strategy and the GRU method have been evaluated. 300,000 submissions have been compared to the outcomes. Accuracy, precision, recall, execution time, and space complexity have been determined utilizing GRU and the suggested hybrid model.

5.1 Algorithm Evaluation

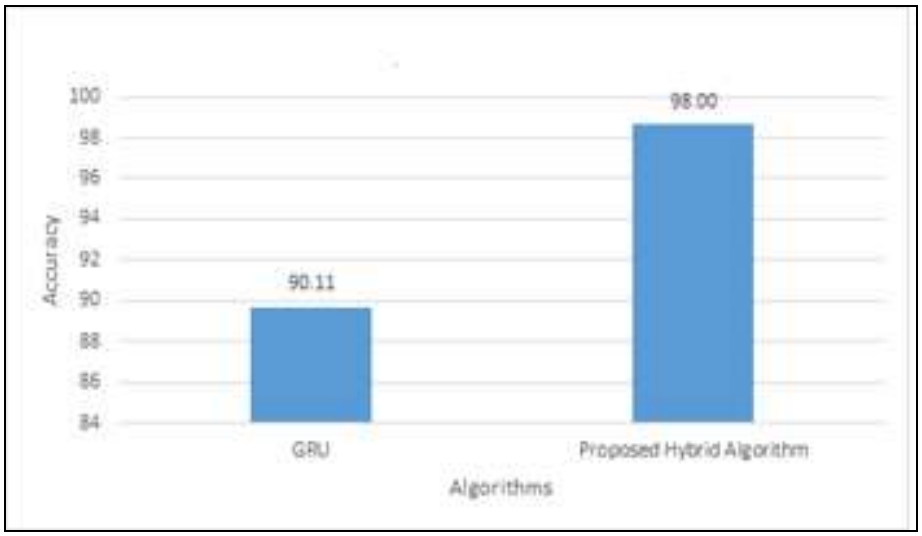

Fig 4: Gate Rated Unit (GRU) and Proposed Hybrid Technique

Compared to GRU, the proposed combined approach offers greater precision. Furthermore, the correctness has been evaluated for a range of papers. There have been examinations of 100K to 300K records.

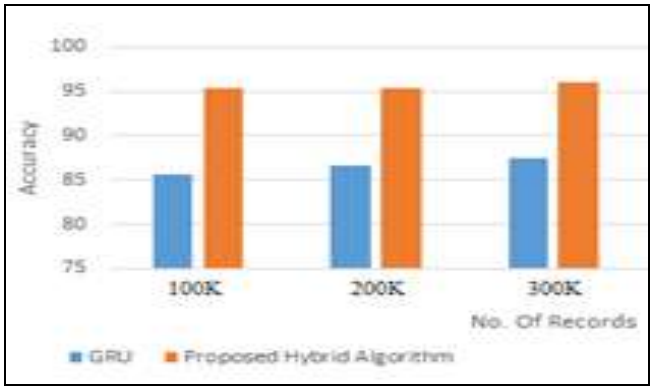

Fig 5: Range of Entries comparison100K-300K

Precision and Recall Experimental Result

Details	Gate Rated Unit (10%)	Proposed Assembled LSTM (10 %)
Precision	8.2%	9.4%
Recall	8.0%	9.3%

6. Conclusion and Future Work

In this research we evaluated the accuracy, recall, and precision performance measures for the Gate Rated Unit with the proposed algorithms. In comparison to the proposed Hybrid Long Short Term M model, Gate Rated Unit has showed an accuracy of 90.11 percentage, which is quite low. The suggested assembled model has shown an accuracy of 98.0 percentage. The accuracy of both models has indeed been evaluated by comparing them to different numbers of entries. It is determined that as the number of records increases, so does the accuracy of both algorithms. In terms of precision, the proposed model outperforms the Gate Rated Unit method. For both fewer and more input, the proposed approach produces more accurate results than Gate Rated Unit. A phonetic-based Hybrid Model has been built for the purpose of efficiently retrieving names using Indian Demographic data. Neither Long Short Term Memory nor Gate Rated Unit display vanishing gradient. The Gate Rated Unit and Assembled models were used to perform a phonetic comparison. When analyzing individual performances, accuracy, recall, precision, execution speed, and memory all have been taken into account. Ultimately, future work will need to check the other countries' regional or local languages' name accuracy using this research or techniques.

7. References

1. Miri A, Rashid F. Secure Textual Data Deduplication Scheme Based on Data Encoding and Compression, 2019 IEEE 10th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON); c2019. p. 0207-0211. doi:10.1109/IEMCON.2019.8936222.
2. Constantinescu C, Pieper J, Li T. Block Size Optimization in Deduplication Systems, Data Compression Conference; c2009 p. 442442. doi:10.1109/DCC.2009.51.
3. Bin Ayub Khan A, Ghazanfar MS, Khan SI. Application of phonetic encoding for analyzing similarity of patient's data: Bangladesh perspective," 2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC); c2017. p. 664-667. doi:10.1109/R10-HTC.2017.8289046.
4. Bhattacharjee K *et al*. A Novel Approach of Deduplication on Indian Demographic Variation for Large Structured Data. In: Nagar, A.K., Jat, D.S., Marín-Raventós, G., Mishra, D.K. (eds) Intelligent Sustainable Syst. Lecture Notes in Networks and Systems, Springer, Singapore. 2022, 334.
5. Kaushik Vandna, Bendale Amit, Nigam, Aditya, Gupta, Phalguni. Certain Reduction Rules Useful for De-Duplication Algorithm of Indian Demographic Data. International Conf.on Advanced Comp. and Comm. Tech; c2014. ACCT.79-84.10.1109/ACCT.2014.85.
6. Hasim Sak, Andrew Senior, Francoise Beaufays. Long Short-Term Memory Based Recurrent Neural Network Architectures for Large Vocabulary Speech Recognition; c2014. https://doi.org/10.48550/arXiv.1402.1128
7. Mane, Madhuri, Ghorpade, Vijay. Comparison of Data Duplication Algorithms; c2015.www.ijcst.com.6.
8. Xumin SNa L, Yong G. Research on k-means Clustering Algorithm: An Improved K-means Clustering Algorithm," 2010 Third International Symposium on Intelligent Information Technology and Security Informatics; c2010. p. 63-67. doi:10.1109/IITSI.2010.74.
9. Zhang, Xiaodong, Sun, Xu, Wang, Houfeng. Duplicate Question Identification by Integrating FrameNet With Neural Networks. Proceedings of the AAAI Conf. on Artificial Intel; c2018. P. 32. 10.1609/aaai.v32i1.12036.
10. Rajib Rana, Julien Epps, Raja Jurdak, Xue Li, Roland Goecke, Margot Breretonk, *et al*. Gated Recurrent Unit (GRU) for Emotion Classification from Noisy Speech; c2016. https://www.adweek.com/category/socialtimes/smartphones/480485
11. Esteves RM, Hacker T, Rong C. Competitive K-Means, a New Accurate and Distributed K-Means Algorithm for Large Datasets, 2013 IEEE 5th International Conference on Cloud Computing Technology and Science; c2013. p. 17-24. doi:10.1109/CloudCom.2013.89.
12. Cai J, Wei C, Tang XL, Xue C, Chang Q. The Motor Imagination EEG Recognition Combined with Convolution Neural Network and Gated Recurrent Unit. 2018 37th Chinese Control Conference (CCC); c2018. p. 9598-9602. doi:10.23919/ChiCC.2018.8484033.
13. Istiake Sunny MA, Maswood MMS, Alharbi AG. Deep Learning-Based Stock Price Prediction Using LSTM and Bi-Directional LSTM Model, 2020 2nd Novel Intelligent and Leading Emerging Sciences Conference (NILES); c2020. p. 87-92. doi:10.1109/NILES50944.2020.9257950.