International Journal of Advances in Electrical Engineering

E-ISSN: 2708-4582
P-ISSN: 2708-4574
IJAEE 2023; 4(1): 10-17

www.electricaltechjournal.com
Received: 15-10-2022
Accepted: 18-11-2022

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

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

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

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

De-duplication avoidance in regional names using an approach based on pronunciation

Nagesh Raykar, Dr. Prashant Kumbharkar and Dr. Dand Hiren Jayatilal

DOI: https://doi.org/10.22271/27084574.2023.v4.i1a.32

Abstract
Demographic data deduplication has occurred in every field, including government, marketing, and opinion research, particularly if you work in IT and are in charge of taking backups or transferring large amounts of data. Duplication occurs both directly and indirectly when copying the same backup. As a result, there is an inherent need to proceed or remove redundant data. The term "de-duplication" refers to the removal of duplicate data. This is required for better data storage utilization. The deduplication process involves removing the duplicate copy and keeping only one copy. Deduplication includes a de-duplication process. A different user stores the same file in the same location. As a result, it increases redundancy. Many scholars have already did work on demographic data de - duplication, and one such requirement is that a specific reduction rule is useful for the deduplication algorithm in Indian demographic data. Based on the pronunciation rule, the researchers will evaluate the regional name, first name, and last name. It is necessary to test with various phonetic-based algorithms and then develop an efficient new phonetic-based algorithm. The phonetic algorithm is responsible for indexing words based on their own phonetics. The majority of phonetic algorithms have been primarily designed for English language. Demographic Information provides data on individuals based on features such as First name, Surname, age, gender, contact no, email id, and so on. Considering the Indian regional languages names scenario, we must identify an individual who has the same name but different spellings. The proposed study compares traditional regional names in the format of First name and Surname using the pronunciation rule. For the local languages, a prototype effective phonetic-based algorithm has been developed. An effort has been made to avoid redundant information in the names, and secondly, equivalent names, even with different alphabetical arrangements, have been identified in order to locate an individual in e-governance of a region or any industry. The proposed approach's findings are encouraging, and it can be used in a real - world environment.

Keywords: Phonetic algorithm, data de-duplication, demographic data, natural language processing

1. Introduction

1.1 Background study

The pronunciation-based strategies are adopted with word indexing. The majority of phonetic algorithms were made solely for the English language. They are well suited to English because it is a phonics-based language. These algorithms cannot be used to index words in all other languages. Indian languages are not phonic. These techniques designed for the English language are not always useful for indexing words in Indian regional languages such as Marathi, Hindi, Telgu, Kannada, and Bengali. Demographic Data provides information about a person by using characteristics like as firstname, lastname, date of birth, gender, contact number, email id, and address. Thus, demographic data provides statistics in web analysis. It is critical to de-duplicate the data in order to obtain the precise information required. In recent years, many investigators have contributed to the field of demographic trends. Duplicate data removal is critical when processing Demographic Data. In Indian Demographic Data, a reduction rule that really is beneficial for the deduplication algorithm is required. Trying to identify recurring patterns in files on that volume is obligated during the deduplication process. Data deduplication primarily helps to optimize without having to sacrifice data accurateness or authenticity. Identifying recurring patterns in documents on that volume during in the deduplication process. Data deduplication mainly helps to optimize without having to sacrifice data adherence or authenticity [1].

1.2 Literature Review

In the literature review, it was clear that there was a real need for this kind of research problem and improvements. The author will see the research study as both a theoretical and an experimental method. In the reviewed literature, I will learn about relevant studies that is based on phonetic pronunciation in addition to different ways to get rid of duplicate data. We are shown different ways to get rid of duplicate data that we can use in our research. I looked at various research papers for my current research. When the author did research in the past, he or she used particular reduction rules to come up with database search queries. Based on the MinDist and Accurateness criteria, the author have used Edit Distance method. It will discover the five nearest matches for the query's first and last names. It will help us cut down on process time, database space needs, and costs related to those things. This gets rid of the extra repetition. Both this givenName and this lastName should have the value "generic" [2]. The two steps of a good algorithm for removing duplicates from demographic data are "enrollment" and "de-duplication." Both stages are used to cut down on the use of generic names. This is done with the help of rules based on phonetics. During the enrollment phase, the database that will be used in the bin is actually made. The SLL (single link list) basically links each bin [3]. In this research paper, the "Top-K best matches" method is used [3]. The main problem for the researcher is looking for different kinds of data in a large, often wrong database. The author talked about the different phonetic matching-based algorithms for the dataset of North Carolina street names and English dictionary words used in this study [4]. This study used the Soundex algorithm for phonetic matching. The Caverphone phonetic technique is used to do phonetic matching. There are two versions of the Caverphone: Caverphone 1.0 and Caverphone 2.0. Most people use the Caverphone 2.0 to match based on how their voices sound. This is easy to get to and doesn't cost anything to use. This algorithm was made as a part of the Caversham Project [5]. The D-M Soundex technique is used for most of the string matching method. In this research study, this algorithm requires words as input and gives out a list of codes [6]. The D-M Soundex will be made for German and Slavic names, while this algorithm was made for English pronunciation predicated on phonetics. This document has the rules for the vocabulary of the Serbian language [6]. The Levenshtein, Jaro-Wrinkler, and D-M Soundex are all used in this research [6]. The Double Metaphone phonetic method was used by the program that checks spelling. This application made the system that suggests misspelled words better at what it can do. This study compared the edit distance technique and the Double Metaphone phonetic method [7]. The author of this study talked about how important the phonetic Metaphone algorithm is for searching and correcting textual information. To find the parts of the Brazil-ian Metaphone phonetic algorithm that don't match up [8].

1.3 Main Contribution

- To develop an efficient technique for extracting words from Indian demographic characteristics.
- To create a technique for De-duplicate data based on phonetics similarity.
- Make a comparison traditional approaches to the proposed approach.

2. Methodology

2.1 Working of already Existing algorithm

A phonetic algorithm relies on phonetics, and these methods operate with word indexing. These techniques were designed for English and might not be suitable for indexing words from local languages such as Marathi, Hindi, Telgu, and Bengali. The traditional algorithms listed below have been thoroughly investigated.

- Soundex
- Daitch-Mokotoff Soundex
- Metaphone
- Double Metaphone
- Kolner Phonetic
- Match Rating Approach
- Caverphone (I & II)

2.1.1 Soundex Algorithm

Soundex is a phonetic-based technique that encodes names according to English pronunciation by voice. This will return the name's indexing. The Soundex technique produces a four-digit sequence. The Soundex method has the benefit of avoiding or minimizing most difficulties related with misspelled words of family names. Even before misspellings of family names were common in official papers, the Soundex technique is a helpful tool in searching for ancestors [9].

2.1.2 Daitch-Mokotoff Soundex

This is the next version of the Soundex method for Easter European surnames. The method is mainly used to find nearby or near matches with Eastern European surnames, which include Russian and Jewish names. The method, like Soundex, ciphers in to the digits by expanding it to an entire 6-digit encrypted data or code. D-M Soundex conversion rules are significantly more complicated than Soundex rules because they needs categories or groups of characters for encrypting. This algorithm is designed to complement Yiddish and Slavic surnames with substantially enhanced or accurateness in the American Soundex. The D-M Soundex technique is utilized because these method surnames have similar pronunciations but spelling variance [10].

2.1.3 Cologne Phonetic

The Soundex technique is equivalent to this algorithm. It is based on the hypothesis that characters with similar sounds have matching codes assigned to them. This technique can be utilized to find similarities between character or word. So every letter of a word is compared to a digit from "0" to "8" in the Kolner phonetic method. In the choosing of a suitable digit to use in context. We had to choose the best digit for the final one adjacent character using conditions. Just few rules attempt to apply to the preliminary characters/words in a particular way. Similar sounds are supposed to assign identical code in this manner [11].

2.1.4 Caverphone

This is designed to recognize English names based on their sounds. Caversham Project invented the Caverphone method. This technique has two different versions: the first was created in 2002, and the second was created in 2004. This technique was created with the objective of presently trying to pronounce a method for presenting in the study region, New Zealand, towards to the south part of the city of Dunedin. This technique comes in two versions: Caverphone 1.0, which was established in 2002, and Caverphone 2.0, which was established in 2004. The Caverphone 2.0 Technique is better suited for common phonetic matches. This technique is now available and free to use.

2.1.5 Metaphone

This technique is utilized to index characters/words based on how they're said that in English. This is an enhanced Soundex technique that makes use of data on variance and lack of consistency in the English language. The pronunciation will consequence in definitely exact encoding that does a great job of matching characters/words, names that are of a similar type or the same. Similar sounding words will start sharing the same keys for a similar name, according to Soundex. This technique is available in a wide range of systems with built-in operators [12].

2.1.6 Double Metaphone: Philips developed a new edition of the Metaphone technique, subtitled Double Metaphone. The earliest Metaphone only supported English phonetic encoding prerequisites. Metaphone pronunciation algorithmic method is superior to Double Metaphone [13].

2.1.7 Match Rating Approach: Match Rating Approach Phonetic Technique is made for indexing and telling the difference between names that sound the same but are spelled differently. A simple set of rules for encoding is used by the method. The main method is similarity distinguishing, which determines the number of unmatched or unequal characters or words by trying to measure sequences from left to right (L to R) and then from right to left (R to L) and getting rid of words or characters that are equivalent. This number is deducted from 6, and then the difference is compared to a minimum threshold. Table A shows the minimum threshold, which is based on the length or evaluation of the strings. The name that has been hidden is also called a Personal Numeric Identifier. The name that is coded can never have more than 6 alpha-only words or characters [14].

2.2 Steps followed by the already existing algorithm

2.2.1 Phonetic Base Generic Name Deduction Rule

If a genetic deduction-based method only looks at the phonetic part, then the number of different names can be cut down by a lot. By getting rid of the method, it helps to reduce the amount of search space. This part has several functions that help translate Indian names (Firstname and Surname) into common names. The illustration names "Mitesh," "Meetesh," and "Metesh" can be reduced down to the name "Mitesh". To find the general name, each name of the string can be broken down into a few phonetic steps. Phonetics is the part of linguistics that looks at how people's voices sound. Vowels and consonants are the two types of phonetic sounds. This is an example of how the proposed reducing rule can be described: Asheesh can be changed to Asish.

2.2.2 Making of Database

As in current method, rules are used to translate any given name string into a shorter generic name, which is then added to the database. There may be more than one name string that gets turned into a single generic name. You can explain what a bin is by its generic term, which is a list of all these different types of name strings. Also, you may have observed that there could be more than one person with the same name but a different Id. So, one can act as the bin when making a Singly Linked List, where each node holds a name string with different Ids. A SLL (Single Link List) can be used to add and remove links.

2.2.3 Matching Strategy Method

Edit distance technique is used to find the distance between two characters in a string and do the fewest possible operations. Insert, delete, replace, transpose. Levenshtein distance is an example of edit distance, which measures how far apart two strings of characters are in the middle. To get the normalized distance, this interval is split up into name strings with a maximum length of two. Name strings that are associated with bins and have a standard edit distance from the name string of the query data that is less than the acceptable range are thought to be candidates for multiple copies.

2.2.4 Working Flow of Existing algorithm

The existing demographic algorithm used this approach is as in Figure 1

In experiments, the average size of a bin is 1.278. Given Name has a standard deviation of 0.808 and a reduction percentage of 90.94%. Think about the number 27184 for the last name, Bins. The smallest bin is one and the largest is fifteen. Based on experiments, the average size of a Bin is 1.288. Given Name has a standard deviation of 0.797 and a reduction percentage of 94.40%.

Here are the latest findings of the system's Top Matches search queries for the name "Isha Arora" based on Min Dist as well as His Score.

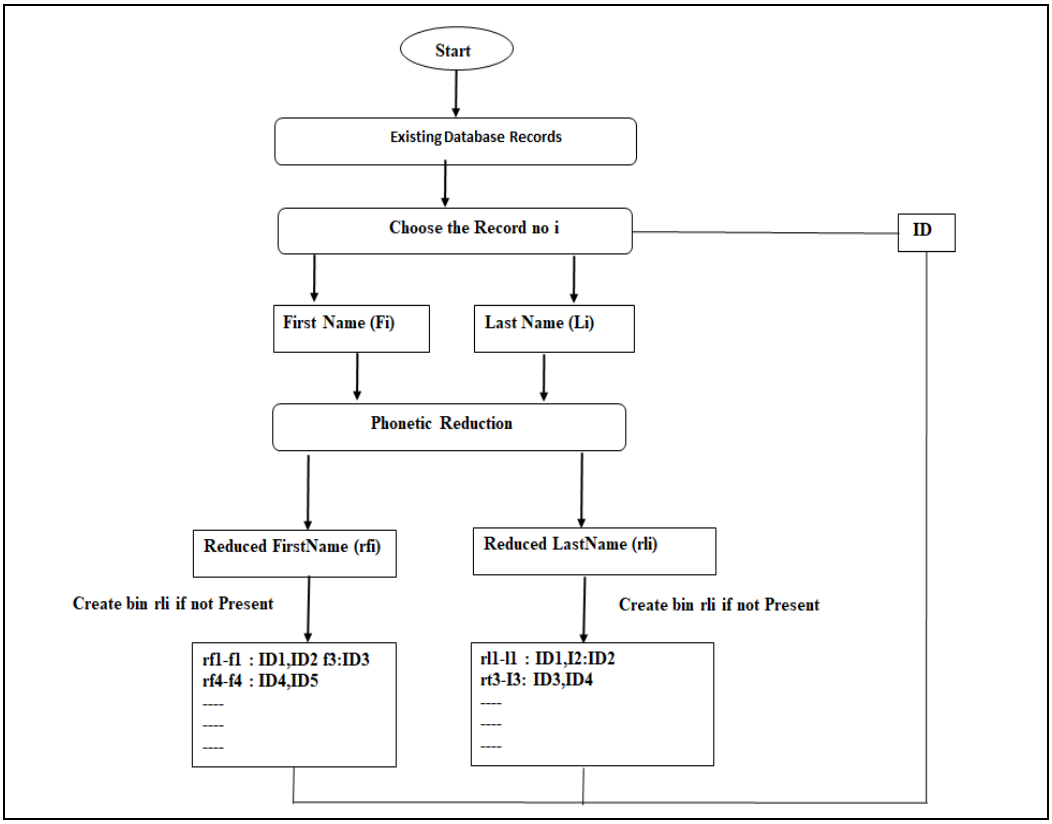

Fig 1: Flow of existing algorithm

Table 2: Finds the Name, Minimum Distance, and Score.

Isha Arora, Min Dist '0' and the Score is '0.0'
Arora Isha Ajay, Min Dist '0' and the Score is '0.06'
Arora Abhi Ravi, Min Dist '0' and the Score is '0.08'
Arora Alok Prashant, Min Dist '0' and the Score is '0.09'

With the current method, the 94% is cut by replacing the given name and last name with a generic name.

3. Proposed methodology

The proposed methodology is broken down into four stages.

The steps are as follows:
Step I: Data Classification
Step II: Suggested algorithm
Step III: Use Machine Learning Processing to Convert the Ciphered Key
Step IV: Classifying
Each step will be discussed in depth in the sections that follow.

Step I – Classification Alignment

Figure 2 illustrates the data classification phase as follows:

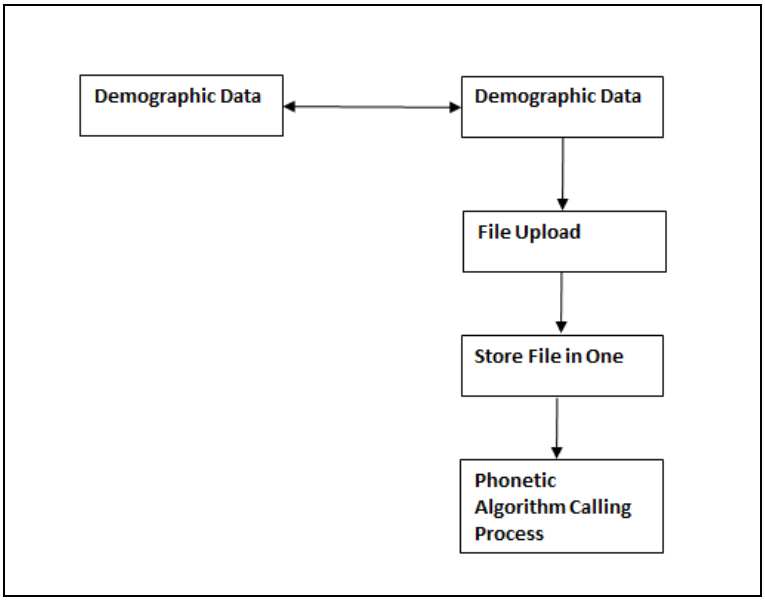

Fig 2: Phase I – Data Classification

In Step I, the data is put into groups. The information about the people is given in.xlsx or.csv format. A strategy for sampling is chosen. The file has been sent. This information is sent to the next step, which is called Step-II.

Step II- Suggested algorithm.

Suggested phonetic based algorithm is showing in below figure 3:

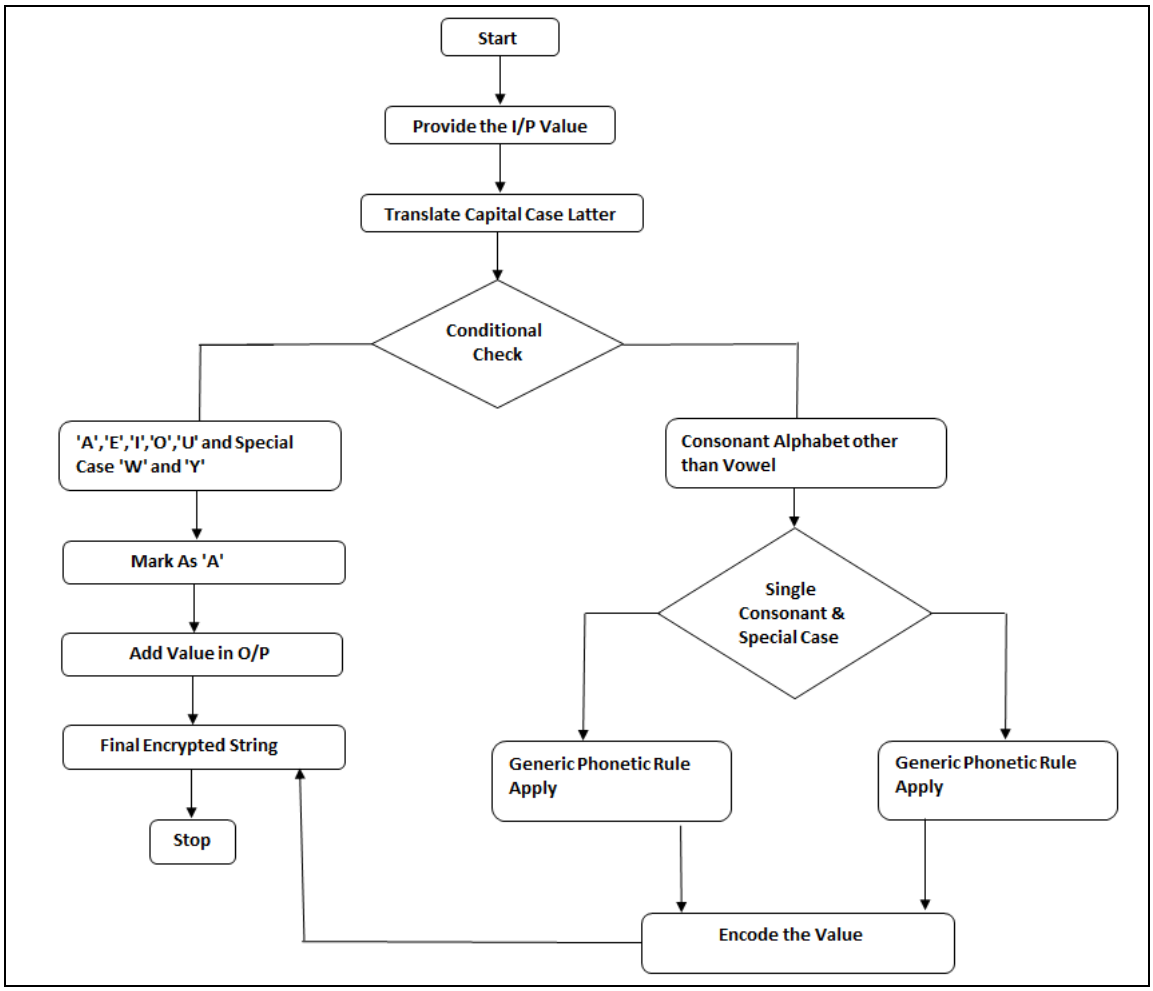

Fig 3: Suggested Phonetic Technique

The working of this suggested phonetic algorithm is as below.

- Inside this supply input, this same practice of adding a single input or a large number of inputs is executed.
- The inputs given are required to be changed to all capital letters. Names like "Dileep" will now be changed to "DILEEP."
- Remove words that are used more than once: once the name is changed to all capital letters. Its words that are used over and over again need to be taken out, and one letter in each of those words is highlighted.
- Phonetic rules will be applicable on the basic of character is alphabet or consonant.
- If the character or alphabet is vowel marked as ‘A’.
- If the same character is repeated then marked single character.
- In Some special case ‘W’ and ‘Y’ consider as the vowel and marked as ‘A’.
- If the alphabet is consonant, then it will proceed to this part with consonants. Consonant signifies a certain sound that isn't a vowel in that kind of speech.
- In Non-Vocalize pronoun, each consonant was mapped to an encipher value that was the same. This will give each consonant its own encoded value.
- If consonant is repetitive then treated as a Single character. As per the pronunciation the alphabet is translated into the encoded value.
- In the encoded alphabet scenario or special cases According to a unique consonant requirement, the alphabet will be encoded. Examples include quiet and the double consonant pronunciation. So that the situation will be addressed appropriately.
- Make the encrypted code according to consonants.
- In the anticipated output, show the name-wise encrypted data.

The Machine Learning processing technique gets the vowel/alphabet encrypted value as just an input in step III.

Step III: Use Machine Learning Processing to Convert the Crypto Key.

Step III is made up of two smaller parts

Section I: In this step, the Machine Learning technique is put to use at the time this step, the Crypto Key Conversion is done.

Section II: In this phase conversion of Crypto Key is carried out.

Step III Section I

Figure 4 shows the steps that are taken in the Machine Learning technique.

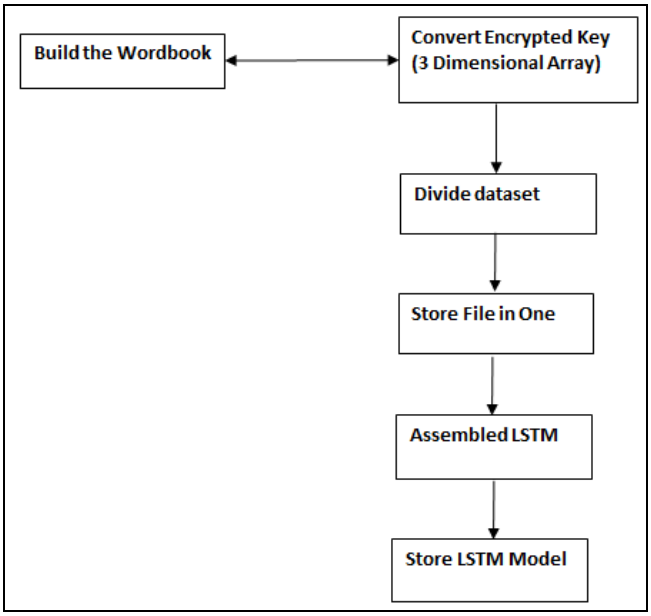

Fig 4: Step III, Machine Learning Processing Approach

Create a wordbook based on the following information. It is made up of encrypted key & values.

Step III Section II: Convert Crypto Key

In Step III, which is about deciphering the key, the key value combinations of names that were found in Step II are used as input. When these encrypted keys are changed, the keys are turned into the three-dimensional array. This can be shown with the aid of an example. Assess name "DALAP".

For the conversion, the following steps are taken. Dividing a Set of data & dividing the input values that were produced with the help of a cluster. Constructed LSTM-Based Computation: The constructed LSTM is used to measure the process clusters. LSTM tends to help us separate inputs by clusters.

Input: array of three dimensions of Crypto Key

Output: p * p (p : number unique Crypto keys in the cluster)

Store LSTM Design: LSTM Prototype is eventually being saved.

Phase IV: Classifying Method

The last step of the suggested technique is Classifying method.

The Classifying method is showing in Figure: 5

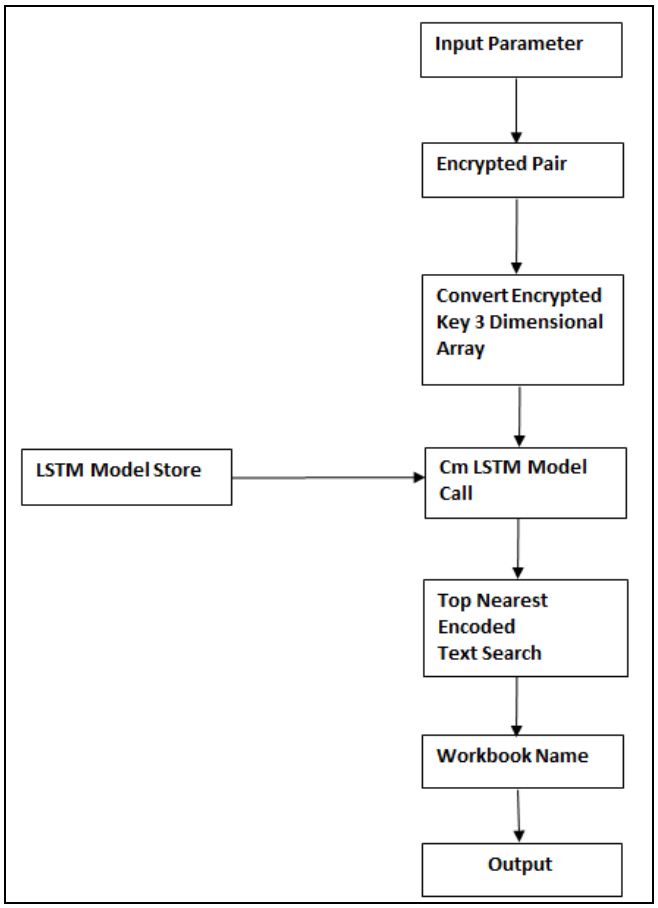

Fig 5: Classifying Method

To put the names into groups, the following steps are taken.

- The detection model for just a workbook's input data uses the classify model. It will find the match in accordance with the classification.
- Give the Inputs: It gives the value of the input.
- Encrypted pair: The template of the input data must be encrypted pair.
- Conversion in Crypto Key: During the procedure, these encrypted keys are converted into a three-dimensional array. Suppose us about string “DALAP ".
- Find nearby Groupings: In the Classify method, once the encrypted pair is found, it will discover the group

member who is closest to it.

- The Cm LSTM Model has been loaded. Input data for that model come from the Cached LSTM Model. That once designed to match is done, the next procedures, which is 5 Nearest Encrypted Text, will be called.
- This step is where the best 10 encrypted matches are shown to the final user.
- In the details of the workbook, we get a name with a list of feasible combinations.
- The final output is the Name of Arrays with the Top K Matches

4. Experimental Results, Analysis and Discussion:

With the 300k Sample Accuracy data, we had also compared the outcomes to the algorithms that are already in use. The proposed technique has been evaluated by comparing to Soundex, Daitch-Mokotoff Soundex, Caverphone1, Caverphone2, Kolner Phonetic, Metaphone, Double Metaphone, and Match Rating Approach Encoder. All of methods' accuracy, precision, recall, time complexity, and space complexity have been worked out.

Accuracy Comparison for All Traditional Phonetic Technique vs. Proposed algorithm shown in Figure -6

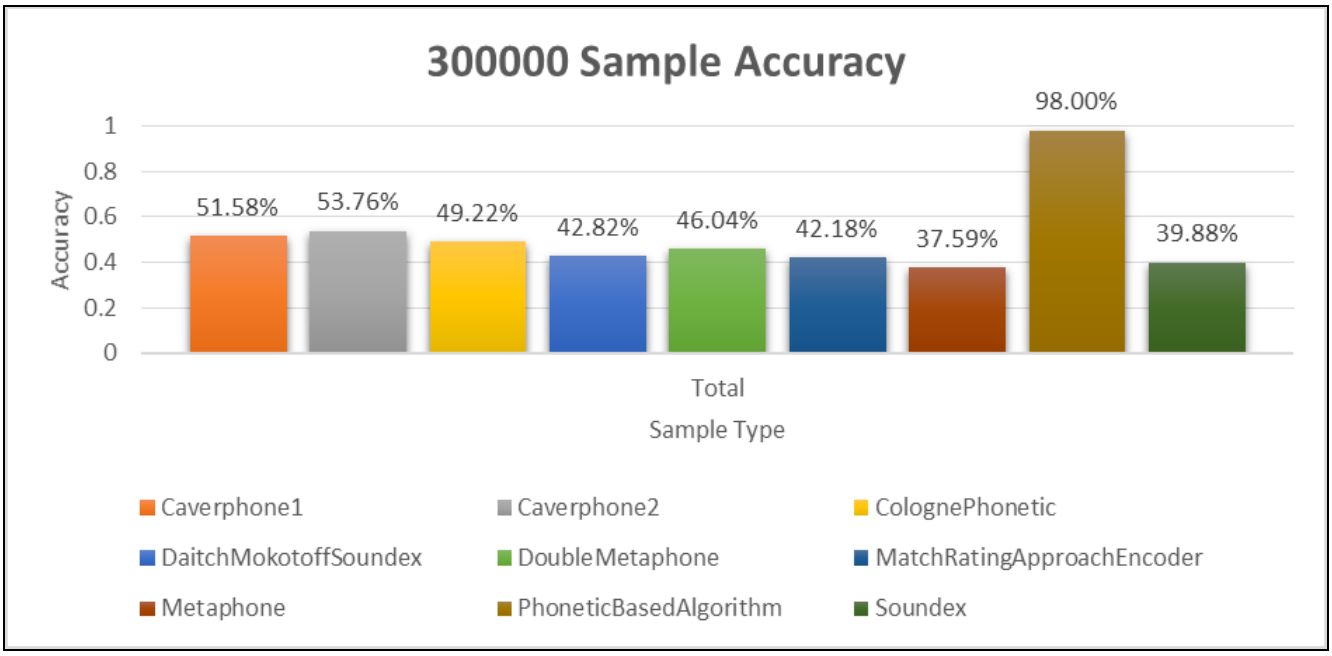

Fig 6: Accuracy Comparison of the proposed algorithm with existing algorithms

Maximum accuracy has been shown by the proposed phonetic-based algorithm. The next highest accuracy is shown by Caverphone2.

Precision has been measured to compare how well the proposed approach and the ones already in use work. Below table depicts the findings.

4.1 Precision

Positive predictive value is measured by precision. It has been computed for both algorithm.

Table 2: Precision Comparison for the Current and the Suggested algorithm

Algorithm Name	Evaluation (%)
Soundex	7.4%
Daitch Mokotoff Soundex Algorithm	7.6%
Caverphone 1	8.2%
Caverphone 2	8.4%
Metaphone	8.5%
Double Metaphone	8.6%
Match Rating Approach	8.2%
Proposed phonetic algorithm	9.4%

Soundex, D-M Soundex, Caverphone 1 and Caverphone 2, and Metaphone, Double Metaphone, Kolner Phonetic, Match Rating Approach

4.2 Recall

Table 3: Precision Comparison for the Current and the Suggested algorithm

Algorithm Name	Evaluation (%)
Soundex	7.1%
Daitch Mokotoff Soundex Algorithm	7.4%
Caverphone 1	8.1%
Caverphone 2	7.9%
Metaphone	7.2%
Double Metaphone	7.9%
Match Rating Approach	8.6%
Proposed phonetic algorithm	9.3%

The suggested phonetic-based technique shows the highest recall of 9.4 the second highest recall is shown by Match Rating Approach.

5. Conclusion

The newly made phonetic algorithm is utilized to use Indian demographic information to pull out the names firstName and surname. This method was compared to a number of many other phonetic techniques, such as Soundex, D-M Soundex, Caverphone 1 and Caverphone 2, and Metaphone, Double Metaphone, Kolner Phonetic, Match Rating Approach encoded with the parameter. The new technique for Indian names is more accurate than the previous algorithm, which was based on phonetics. It says that it is

right 98.0% of the time. This technique works best with Indian names. With a precision of 9.4, the suggested technique has shown to be the most accurate. We've come to the conclusion that despite the existence of numerous phonetic-based techniques already, they don't work very well when trying to compare Indian names. The suggested methodology works better at finding the Indian names that sound the same but have different spellings. In this work, a new, productive phonetic-based technique for the local language has been suggested.

6. References

1. Prathilothamai M, Nair PS. De-duplication of passports using Aadhaar, 2017 International Conference on Computer Communication and Informatics (ICCCI); c2017. p. 1-5. doi: 10.1109/ICCCI.2017.8117744.
2. Kaushik VD, Bendale A, Nigam A, Gupta P. Certain Reduction Rules Useful for De-Duplication Algorithm of Indian Demographic Data, 2014 Fourth International Conference on Advanced Computing & Communication Technologies; c2014. p. 79-84, doi: 10.1109/ACCT.2014.85.
3. Dixit, Vandana, Bendale, Amit, Nigam, Aditya, Gupta, Phalguni. An Efficient Algorithm for De-duplication of Demographic Data; c2012. 10.1007/978-3-642-31588-6_77.
4. Koneru K, Pulla V, Varol C. Performance Evaluation of Phonetic Matching Algorithms on English Words and Street Names - Comparison and Correlation. DOI: 10.5220/0005926300570064 In Proceedings of the 5th International Conference on Data Management Technologies and Applications (DATA ISBN: 978-989-758-193-9 Copyright c 2016 by SCITEPRESS – Science and Technology Publications, Lda. ; c2016. p. 57-64
5. Caversham Project Occasional Technical Paper, Code Number: CTP060902, Author: David Hood, caversham@otago.ac.nz
6. Xvii Conference on Applied Mathematics D. Herceg. H. Zarinarin.eds, Adaptation and Application of Daitch-Mokotoff Soundex Algorithm on Serbian Names, Peter Rajkovic, Dragan Jankovic, Department of Mathematics and Infomatics Novi Sad; c2007. p. 192-204.
7. UzZaman N, Khan M. A Double Metaphone encoding for Bangla and its application in spelling checker, 2005 International Conference on Natural Language Processing and Knowledge Engineering; c2005. p. 705-710, doi: 10.1109/NLPKE.2005.1598827.
8. Jordão CC, Rosa JLG. Metaphone-pt_BR: The Phonetic Importance on Search and Correction of Textual Information. In: Gelbukh, A. (eds) Computational Linguistics and Intelligent Text Processing. CICLing 2012. Lecture Notes in Computer Science. Springer, Berlin, Heidelberg; c2012; p. 7182. https://doi.org/10.1007/978-3-642-28601-8_25
9. del Pilar Angeles M, Espino-Gamez A, Gil-Moncada J. Comparison of a Modified Spanish Phonetic, Soundex, and Phonex coding functions during data matching process, 2015 International Conference on Informatics, Electronics & Vision (ICIEV); c2015 p. 1-5. doi: 10.1109/ICIEV.2015.7334028.
10. Draganov IR, Popova AA, Ivanov LL. Multilingual Names Database Searching Enhancement, 2008 IEEE International Symposium on Signal Processing and Information Technology; c2008. p. 474-479. doi: 10.1109/ISSPIT.2008.4775648.
11. Yu, Tzu-Yang *et al*. HetCast: Cooperative data delivery on cellular and road side network. 2017 IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC); c2017. p. 1-6.
12. Li-dong, Ma. Metaphone Phonetic Matching Algorithm and Its Application; c2010.
13. 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.
14. Long S, Feng Q, Chen W. A Novel Approach to Automatic Rating of Subjective Answers Based on Semantic Matching of Keywords, 2016 12th International Conference on Computational Intelligence and Security (CIS); c2016. p. 87-90. doi: 10.1109/CIS.2016.0028