iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** III **Month of publication:** March 2024

DOI: https://doi.org/10.22214/ijraset.2024.58764

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue III Mar 2024- Available at www.ijraset.com

Survey on Context Based Identification of Customer Name Variations using ML Techniques

Vaishnavi Katkar[1], Sandhya Rakhunde[2], Kaveri Raut[3], Simran Godhwani[4]
Department of Computer Engineering, SCTR's Pune Institute of Computer Technology, Pune, India

Abstract: ***The demand for automated validation of new customer company accounts, with a specific emphasis on resolving name variations within large databases, is steadily increasing. This surge in demand is a direct response to the significant scale at which data for new customers is being input into CRM systems, necessitating manual validation and correction processes. The primary objective of this survey paper is to identify machine learning methods capable of addressing the challenge of resolving name variations in extensive databases. To achieve this goal, we propose a multi-step approach. Firstly, we employ an approximate string matching algorithm with high computational speed to identify matches with a very high percentage of accuracy. In cases where such matches are not found in the system, we utilize named entity recognition techniques to annotate company names and their suffixes, such as INC, PVT, LTD, CORP, and CO, which may appear in various forms. To resolve abbreviation disambiguity, we explore the application of machine learning algorithms, including the naïve Bayes classifier, decision trees, and Support Vector Machines. In this survey paper, we conclude by presenting potential approximate string matching algorithms, a named entity recognition method, and a model for resolving abbreviation disambiguity. Our review not only provides a comprehensive overview of the current state of research in this area but also highlights gaps in the existing knowledge, offering valuable insights for future research and practical application.***

Keywords: Machine Learning Models, Approximate and Exact String-Matching Algorithms, Abbreviation Disambiguation, Fuzzy String Matching.

I. INTRODUCTION

In vast databases, customer names can have multiple variations, such as "ABC CO LTD," "ABC Co. Ltd," "ABC Company Ltd," and more. Identifying these variations manually in millions of records is time-consuming and challenging. To automate this process, we explore potential solutions. We begin by using approximate string matching algorithms to find the closest matches with a high percentage of accuracy. However, this may not address cases where variations in suffixes like "CO," "COMPANY," "CORP," or "CORPORATION" point to the same company but are not identified by approximate string matching alone.

To tackle this, we turn to named entity recognition to classify the company name and its suffix. Since company names remain relatively consistent, we focus on identifying and handling suffix variations. Importantly, we cannot directly remove suffixes because different countries may use different suffixes, and a suffix can even serve as a company name in some cases. To resolve abbreviation disambiguation in suffixes, we leverage supervised and unsupervised machine learning algorithms to identify the correct full form of the suffix.

This paper presents a comprehensive approach to automatically finding variations of the same company within extensive databases. We start by discussing potential approximate string matching algorithms, then delve into named entity recognition models and abbreviation disambiguation machine learning models. Ultimately, we propose the most effective algorithms and techniques to streamline this crucial task.

II. LITERATURE REVIEW

[1] HONGFANG LIU *et al.,* stated in the research paper "A Multi-aspect Comparison Study of Supervised Word Sense Disambiguation", about word sense disambiguation (WSD) using machine learning algorithms. Three approaches are examined: naïve Bayes (NBL), decision lists (TDLL), their adaptation of decision lists (ODLL), and mixed supervised learning (MSL). The research emphasizes the effectiveness of WSD when a sufficient number of tagged instances are available for training.

To implement these machine learning techniques, various tools and libraries are utilized, including machine learning libraries like scikit-learn, TensorFlow, and PyTorch, as well as natural language processing (NLP) libraries like NLTK and spaCy. The research draws from three datasets: a biomedical abbreviation dataset (ABBR), a general biomedical term dataset (MED), and a general English dataset (ENG) to conduct comprehensive comparisons.

One of the significant advantages of the study is that supervised WSD methods tend to outperform other approaches, making them highly valuable for various natural language processing (NLP) tasks. However, the research highlights a substantial limitation: the availability of large, broad-coverage, sense-tagged corpora is limited, which poses a challenge to the development of highly accurate WSD systems.

In conclusion, the mixed supervised learning approach is identified as stable and generally superior to other methods across datasets. This research provides valuable insights into enhancing the accuracy of WSD classifiers for handling ambiguous terms in various NLP applications.

[2] MAHESH JOSHI *et al.*, proposed in research paper "A Comparative Study of Supervised Learning as Applied to Acronym Expansion in Clinical Reports" the challenging task of disambiguating acronyms in medical records to determine their correct meanings. This study employs three machine learning algorithms: the naïve Bayes classifier, decision trees, and Support Vector Machines (SVMs) to tackle this issue, with the implementation utilizing software tools like Weka.

The research paper primarily draws its data from clinical notes at the Mayo Clinic, forming the database for the study. The notable advantage of this approach is its remarkable accuracy, consistently exceeding 90% through supervised machine learning with feature combinations. It provides a valuable benchmark for acronym disambiguation, which is particularly significant in the medical domain. However, it's important to note that fully supervised approaches require substantial manual annotation efforts, and the study's scope is primarily limited to the medical field. To address these limitations, the paper suggests exploring semi-supervised and unsupervised methods to reduce the manual annotation effort.

In conclusion, this research emphasizes the extent of acronym ambiguity in clinical texts, revealing a substantial percentage of acronyms with multiple meanings. The outcomes of the study demonstrate that all three methods, namely naïve Bayes, decision trees, and SVMs, consistently achieve high accuracy levels, typically surpassing 90%. This research contributes to improving the understanding of clinical reports and advancing NLP applications in the medical domain.

[3] SERGUEI PAKHOMOV et al., explored the disambiguation of acronyms and abbreviations in clinical discourse in the research paper "Abbreviation and Acronym Disambiguation in Clinical Discourse". This study employed both fully-supervised and semi-supervised approaches for acronym sense disambiguation. In the fully-supervised approach, they utilized two established machine learning algorithms, Maximum Entropy and C5.0 Decision Trees. The semi-supervised approach involved several steps: Sense Inventory, Data Collection, Data Merging, and Context Vectors Generation. Their research contributes to a better understanding of acronym sense disambiguation techniques in the clinical domain.

Maximum Entropy models are versatile and excel in handling complex, high-dimensional feature spaces, making them effective for various natural language processing tasks. They aim to find the probability distribution that is the least biased given a set of constraints, making them valuable for classification and language modeling.

Semi-supervised learning combines labeled and unlabeled data to build models. In acronym disambiguation, it leverages a sense inventory, diverse data sources, data merging, and context vector generation. This approach is useful when obtaining fully labeled data is challenging, as it can make the most of available resources to enhance model performance.

The semi-supervised approach leverages publicly available data from the internet, potentially reducing reliance on confidential clinical reports. The study explores using the vector space model for disambiguation, which shows promise in comparison to traditional classification methods.

This method doesn't usually use a singular formula as it involves representing documents and terms in a multi-dimensional space. But the similarity between two vectors (documents) is commonly calculated using the cosine similarity:

$$[\text{Cosine Similarity} = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|}]$$

Where, $(\mathbf{A}) and (\mathbf{B})$ are vectors representing two documents.

Naïve Bayes Classifier: The study applied the Naïve Bayes classifier to disambiguate acronyms in clinical reports, using probability calculations to select the most likely meaning.

The fundamental formula for Naïve Bayes is the Bayes theorem:

$$[P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}]$$

Where:
- (P(A|B)) is the posterior probability of class (target) given predictor (attribute).
- (P(B|A)) is the likelihood which is the probability of predictor given class.
- (P(A)) is the prior probability of class.
- (P(B)) is the prior probability of the predictor.

Decision Trees: Decision trees were used to disambiguate acronyms by making attribute-based decisions, mapping them to their correct meanings in clinical reports.
Decision trees often utilize entropy or Gini impurity to decide on the best split.
Support Vector Machines (SVMs): Support Vector Machines (SVMs) were employed to create a clear boundary between different acronym meanings, facilitating accurate disambiguation in clinical reports.

The primary objective of SVM is to find the optimal hyperplane that distinctly classifies the data points in n-dimensional space. The equation for a hyperplane is:

$$[\mathbf{w} \cdot \mathbf{x} + b = 0]$$

Where:
- (**w**) is the weight vector
- (**x**) is the input vector
- (**b**) is the bias

[4] The research paper " String Matching Algorithms" by Mukku Bhagya Sri, Rachita Bhavsar, Preeti Narooka focuses on string matching algorithms in document analysis, exploring established ones like Brute Force, Knuth-Morris-Pratt (KMP), Boyer Moore, and Rabin Karp. It introduces three novel algorithms: Enhanced Boyer Moore, Enhanced Rabin Karp, and Enhanced KMP. The experimentation involves .txt and .docx documents, evaluating search time, iterations, and accuracy.
The Knuth-Morris-Pratt (KMP) algorithm stands out for accuracy in existing algorithms, while the Enhanced KMP algorithm excels among proposed enhancements. The Enhanced Rabin-Karp and Enhanced Knuth-Morris-Pratt algorithms are also presented. The Enhanced Rabin-Karp improves on its predecessor, leveraging hashing functions, and the Enhanced KMP builds on the KMP algorithm's efficiency by avoiding reexamination of previously matched characters.
The research concludes that the Enhanced KMP algorithm demonstrates superior accuracy, emphasizing its significance in existing solutions and potential future enhancements. This algorithm holds promise for applications requiring precise and efficient string matching in various input scenarios. Future research could further optimize and explore additional factors for improving the overall efficiency of string matching algorithms. Following algorithms are studied in this paper:

- *Brute Force*

Simple method: On a mismatch, the pattern shifts one step. Used for exact string matching problems.

- *Knuth-Morris-Pratt (KMP)*

Reduces comparisons, achieving linear-time matching (O(n)). Superior accuracy among existing algorithms.

- *Boyer Moore*

Efficiently handles mismatches by shifting the pattern with a precomputed table.

- *Rabin Karp*

Utilizes hash functions for pattern identification. Average and best-case time complexity: O(n + m).

- *Enhanced Rabin Karp*

Improvement on Rabin-Karp, using hashing functions. Best-case running time: O(N+M), suitable for various applications.

- *Enhanced Knuth-Morris-Pratt (KMP)*

Enhances KMP algorithm efficiency. Avoids reexamination of previously matched characters.

[5] The research paper "A Survey of String Matching Algorithms" by Koloud Al-Khamaiseh and Shadi ALShagarin thoroughly explores various string matching algorithms, addressing the complexities associated with aligning patterns with text beginnings and utilizing a sliding window mechanism. The paper emphasizes the importance of a survey to structure the field, avoiding excessive classification and focusing on key features for researchers. Instead of proposing specific algorithms, the paper outlines vulnerabilities and existing problems, directing attention towards improvement solutions. Sections include a survey of string matching algorithms, related work overview, usage guidance, and conclusion.

The concise survey extends to well-known and recently updated hybrid string matching algorithms, categorized into exact and approximate matching. The exploration covers classical methods relying on character comparisons, hashing methods, and approximate matching approaches. The paper categorizes algorithms under each method, encompassing classical, counting, and filtering techniques. Recent updates and hybrid algorithms showcase novel adaptations and combinations, providing a comprehensive resource for researchers.

The research delves into exact string matching algorithms, covering classical methods such as the Brute-Force Algorithm, Knuth-Morris (KMP) Algorithm, Boyer-Moore (BM) Algorithm, and Boyer-Moore-Horspool (BMH) Algorithm. Hashing methods, including the Karp-Rabin (KR) Algorithm, are also discussed. The approximate string matching section explores classical/dynamic programming methods, counting methods, and deterministic finite automata methods. Notable algorithms like Sellers Algorithm, Diagonal Transition Algorithm, Chang–Lampe (CL) Algorithm, and others are outlined.

The paper concludes by presenting recent updated and hybrid string matching algorithms, highlighting adaptations and combinations of existing approaches.

[6] In Lisna Zahrotun's research paper, "Comparison Jaccard similarity, Cosine Similarity and Combined Both of the Data Clustering With Shared Nearest Neighbor Method", the paper investigates the efficacy of different similarity metrics, specifically Jaccard similarity, cosine similarity, and their combination, in the context of text mining and clustering. The objective is to determine the most optimal similarity value for clustering titles of practical work documents from the Department of Informatics Engineering at Universitas Ahmad Dahlan. Preprocessing techniques such as tokenization and stopword removal are applied prior to similarity calculation. The clustering method utilized is Shared Nearest Neighbor (SNN). Results indicate that cosine similarity yields the highest similarity values compared to Jaccard similarity and their combination. Additionally, the choice of the parameter Eps in SNN significantly influences cluster formation, with larger values resulting in fewer clusters. This study underscores the importance of selecting appropriate similarity metrics and clustering parameters to enhance the effectiveness of text mining and clustering tasks.

[7] In the research paper "Review on String-Matching Algorithm" by Zhaoyang Zhang, offers a detailed exploration of four key string-matching algorithms: Knuth-Morris-Pratt (KMP), Boyer-Moore (BM), Bitap, and Backward Non-Deterministic DAWG Matching (BNDM). Each algorithm addresses the challenge of exact pattern matching in strings, offering unique strategies for optimizing search efficiency. KMP stands out for its prefix-matching approach, reducing unnecessary comparisons through preprocessing. BM algorithm, on the other hand, introduces heuristic rules like bad character and good suffix to skip irrelevant comparisons, showcasing superior performance in various scenarios. Bitap algorithm introduces bit-parallelism, leveraging bitwise operations for accelerated computation despite its theoretical time complexity. BNDM algorithm merges the strengths of Bitap and BM, employing bit parallelism and substring search techniques to achieve high-speed matching with lower memory usage. The paper underscores the importance of ongoing research to enhance algorithm performance while reducing resource consumption, paving the way for future advancements in string-matching technology.

[8] In the research paper "DeezyMatch: A Flexible Deep Learning Approach to Fuzzy String Matching" by Kasra Hosseini, Federico Nanni, and Mariona Coll Ardanuy, the authors present DeezyMatch, a Python-based open-source library tailored for fuzzy string matching and candidate ranking. The research explores diverse deep neural network architectures for pair classification and candidate ranking tasks, providing flexibility and superior performance in entity linking. DeezyMatch's algorithmic summaries encompass a Pair Classifier, utilizing Siamese Deep Neural Networks and supporting various architectures, and a Candidate Ranker, employing vector representations and an adaptive searching algorithm. These algorithms collectively enable DeezyMatch to accurately classify query-candidate pairs and efficiently rank candidates within knowledge bases, demonstrating its suitability for diverse applications in Natural Language Processing (NLP) and beyond.

[9] The research paper, "Semantics-Based String Matching: A Review of Machine Learning Models" by Shaik Asha and Sajja Tulasi Krishna, explores the significance of string matching algorithms in diverse fields. Traditional methods like KMP, Rabin-Karp, and Aho-Corasick face limitations in handling real-world variability and semantics.

Recent AI and machine learning advancements offer adaptable solutions, with neural networks, reinforcement learning, and genetic algorithms addressing these limitations. The key takeaways emphasize the shortcomings of traditional algorithms and the advantages of AI-based techniques, enabling fuzzy matching and dynamic adaptation. Neural network architectures, including Siamese Networks, CNNs, and RNNs, are leveraged for their respective strengths. Reinforcement learning optimizes matching policies, and genetic algorithms provide flexibility with semantic matching capability, though facing computational challenges.

In conclusion, the paper highlights the superiority of AI-driven techniques over traditional methods, paving the way for more flexible and semantics-based string matching. Future research directions include scaling techniques, improved interpretability, domain knowledge incorporation, and versatile applications across various domains.

[10] NARENDRA KUMAR et al., delved into the field of approximate string matching in their research and proposed a research paper on “Approximate string matching Algorithm”. They examined various aspects of this topic, including string similarity measures, phonetic coding methods, and coarse search schemes. Their emphasis was on the n-gram technique, which leverages character substring indexing for efficient searches in extensive databases. These tools and techniques are pivotal for identifying similar strings in situations where exact matches are unattainable.

The n-gram technique is a valuable tool in approximate string matching, particularly for efficiently identifying similar strings in large databases. It involves breaking strings into smaller, overlapping segments of 'n' characters, where 'n' is typically a small integer. By indexing and comparing these n-grams, the technique allows for a more flexible and robust search. It can capture partial matches and variations, making it effective in scenarios where exact string matches are challenging to achieve. The n-gram technique is especially useful in tasks like spell checking, auto-suggestion, and text similarity comparisons, where identifying closely related strings is essential.

[11] YOUNGJUN KIM et al., tackled the issue of acronym and abbreviation ambiguity in clinical notes in their research paper “Acronyms and Abbreviations Ambiguity in a Diverse Set of Clinical Notes”. They developed and evaluated an abbreviation disambiguation system for clinical text, which incorporated a pipeline comprising sentence detection, tokenization, dictionary lookup, and Support Vector Machine (SVM) disambiguation to resolve ambiguous abbreviations. Their work harnessed the LRABR Lexicon and diverse features, including word uni-grams, bi-grams, word characteristics, and parts-of-speech, to construct an SVM-based classification module for addressing the ambiguity of clinical abbreviations.

Support Vector Machine (SVM) is a powerful machine learning method for data classification. It identifies the optimal hyperplane that maximally separates data points of different classes, while minimizing errors. SVM is versatile, as it can handle non-linear data by mapping it into higher-dimensional spaces. This technique is used in various applications like text classification and image recognition due to its effectiveness in finding complex decision boundaries. The paper's abbreviation disambiguation approach significantly improves the accuracy of natural language processing (NLP) tasks in healthcare. It systematically resolves ambiguity in clinical abbreviations, enhancing the understanding of clinical notes and ensuring safer and more accurate healthcare data analysis.

[12] In the research paper "Approximate string-matching with q-grams and maximal matches." by ESKO UKKONEN, the focus is on addressing the problem of approximate string matching, specifically for finding approximate occurrences of a pattern string P within a text string T. The paper introduces alternative string distance functions, namely, q-grams and maximal common substrings, which can be computed in linear time, offering advantages over the traditional edit distance. These advantages include linear time complexity, faster solutions for approximate string matching, and the development of efficient hybrid methods for solving the "k differences problem." The paper also suggests future research directions for generalizing the approach to edit distances with a broader range of editing operations and associated costs.

q-grams are fixed-length substrings extracted from a string, used to capture local patterns and similarities efficiently. They are valuable in various applications, including DNA sequence analysis and text matching. Maximal matches represent the longest common substrings between two strings. They are crucial for identifying shared patterns and sequences, particularly in fields like bioinformatics and plagiarism detection. Maximal matches provide insights into structural and content relationships between strings.

The advantages of the research approach presented in the paper include: 1)Linear time complexity for the alternative distance functions (q-grams and maximal common substrings), which is faster than the edit distance. 2)Faster solutions for the approximate string-matching problem compared to traditional edit distance-based methods

[13] The collaborative work by Maryan Rizinski, Andrej Jankov, Vignesh Sankaradas, Eugene Pinsky, Igor Mishkovski, and Dimitar Trajanov delves into a comprehensive comparative analysis of Natural Language Processing (NLP)-based models for company classification. Focusing on efficiency in similarity search, the paper highlights the Lipschitz Embeddings and Ball Partitioning algorithms for their effectiveness in navigating large datasets to identify similar strings, particularly valuable for recognizing customer name variations in extensive account databases.

The discussion emphasizes the significance of approximate string matching algorithms, designed to handle errors and variations in input strings, reducing manual efforts in correcting account information. Algorithms like Lipschitz Embeddings and Ball Partitioning prove instrumental in automating processes, aligning with the project's goal of an 80% reduction in manual effort. The paper showcases the adaptability and scalability of these algorithms to large datasets, catering to the demands of Sales Operations. It introduces the use of unit-cost edit distance as a metric for measuring similarity, emphasizing potential automation and the creation of a comprehensive dataset using advanced NLP techniques. Leveraging pretrained transformer models and employing dimensionality reduction techniques further contribute to the efficiency and adaptability of the classification models, providing insights for selecting appropriate approaches based on accuracy, efficiency, and scalability.

III. GAPS IDENTIFIED

Identifying gaps is a crucial step as it helps in understanding the specific challenges that the project aims to address. Here are some potential gaps that our existing customer name identification model might address:

A. Accuracy and False Positives

Machine learning models may not always provide 100% accuracy. False positives and negatives can occur, leading to incorrect identifications or missing variations.

B. Handling New Variations

ML model may struggle with identifying completely novel or unexpected variations that were not encountered during training. They need to be continuously updated to adapt to evolving variations.

C. Country wise and Cultural Variations

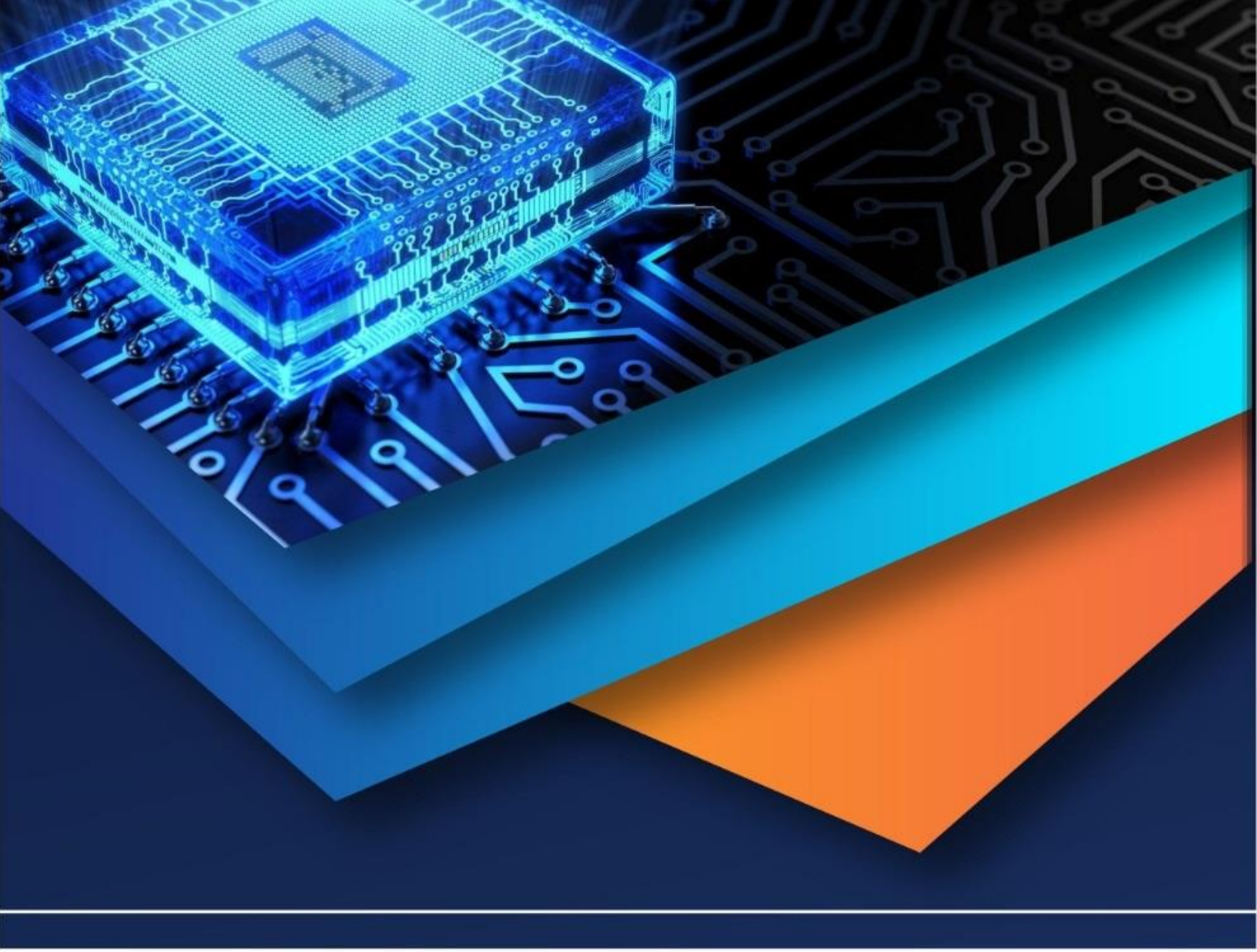
Customer names can vary significantly across countries and cultures. Machine learning models may not perform well when faced with diverse linguistic, country wise and cultural nuances.

IV. CONCLUSION

In conclusion, this survey paper has unveiled significant challenges in managing variations within customer or company names in extensive databases when employing machine learning models. These challenges include the struggle to identify entirely novel variations, the difficulty in accommodating multilingual and cultural nuances, and the necessity for context-based abbreviation identification and disambiguation. However, the proposed techniques address these gaps by introducing dynamic adaptability for continuous updates, accounting for multilingual and cultural variations, and employing context-aware strategies for accurate abbreviation identification. Overall, the outlined methodologies offer a robust and adaptable framework to effectively manage and resolve variations in customer and company names within large databases, ensuring precision and efficiency in diverse linguistic landscapes and evolving data scenarios.

REFERENCES

[1] H. Liu, V. Teller, and C. Friedman, "A Multi-aspect Comparison Study of Supervised Word Sense Disambiguation."
[2] M. Joshi, S. Pakhomov, T. Pedersen, and C. G. Chute, "A Comparative Study of Supervised Learning as Applied to Acronym Expansion in Clinical Reports."
[3] S. Pakhomov, T. Pedersen, and C. Chute, "Abbreviation and Acronym Disambiguation in Clinical Discourse."
[4] Mukku Bhagya Sri, Rachita Bhavsar, Preeti Narooka, "String Matching Algorithms." International Journal Of Engineering And Computer Science ISSN:2319-7242 Volume 7 Issue 3 March 2018, Page No. 23769-23772 Index Copernicus Value (2015): 58.10, 76.25 (2016) DOI: 10.18535/ijecs/v7i3.19
[5] Koloud Al-Khamaiseh, Shadi ALShagarin, "A Survey of String Matching Algorithms", Koloud Al-Khamaiseh Int. Journal of Engineering Research and Applications ISSN : 2248-9622, Vol. 4, Issue 7(Version 2), July 2014, pp.144-156
[6] Lisna Zahrotun, "Comparison Jaccard similarity, Cosine Similarity and Combined Both of the Data Clustering With Shared Nearest Neighbor Method.", Computer Engineering and Applications Vol. 5, No. 1, February 2016
[7] Zhaoyang Zhang, "Review on String-Matching Algorithm.", SHS Web of Conferences 144, 03018 (2022)
[8] Kasra Hosseini, Federico Nanni, Mariona Coll Ardanuy, "DeezyMatch: A Flexible Deep Learning Approach to Fuzzy String Matching."
[9] Shaik Asha1, Sajja Tulasi Krishna, "Semantics-Based String Matching: A Review of Machine Learning Models."
[10] Narendra Kumar, Vimal Bibhu Mohammad Islam, Shashank Bhardwaj, " Approximate string matching Algorithm."
[11] Y. Kim, J. F. Hurdle, and S. M. Meystre, "Acronyms and Abbreviations Ambiguity in a Diverse Set of Clinical Notes."
[12] E. Ukkonen, "Approximate string-matching with q-grams and maximal matches."
[13] Maryan Rizinski, Andrej Jankov, Vignesh Sankaradas, Eugene Pinsky, Igor Mishkovski and Dimitar Trajanov, " Comparative Analysis of NLP-Based Models for Company Classification."



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY