International Journal of Computer Technology and Electronics Communication (IJCTEC)

| ISSN: 2320-0081 | www.ijctece.com ||A Peer-Reviewed, Refereed and Bimonthly Journal |

|| Volume 7, Issue 4, July – August 2024 ||

DOI: 10.15680/IJCTECE.2024.0704014

Improving Data Quality and Deduplication Using Similarity Scoring and Confidence Models

Sravan Kumar Kunadi

Independent Researcher, USA

ABSTRACT: Current information intensive business, decision making, analytics, customer management and efficiency of its activities are now factors of concern attributed to data quality. Nonetheless, mass datasets are normally characterised by huge amounts of redundancy, inconsistency, lapsing of data and incorrect connecting of records that affects credibility and generates mammoth issues down the line. The current studies paper recommends a handy template to enhance quality of data based on the smart identification of the duplicates, and record verification basing on their assurance. The paper deals with the fusion of similarity scoring approaches i.e. string matching, attribute comparison and weighted field-level analysis with confidence models which make an approximation of the likelihood of the records representing the same real-world object. To reduce false positives and false negatives, the framework can be used to reduce the number of false positives and false negatives since it is possible to do this both deterministically and probabilistically to improve rules of deduplication. It is created to handle the heterogeneous data sets in which the spelling variations, formatting, abbreviations and missing values are typical. The solution that is presented, in addition, has a confidence threshold mechanism which enables automated, semi automated and manual inspection processes, which provides additional scalability and certainty in the cleansing process. The findings indicate that similarity based confidence modelling will improve the entity resolve of enterprise data assets immensely, generate uniformity and the overall reliability of the enterprise data assets is also enhanced. The study also adds data management and data governance in the sense that it provides a specialized and generalizable approach to business entities that are interested in quality, holistic, and practical data in the multifaceted digital landscape.

KEYWORDS: similarity scoring, deduplication, Detection of duplication, Confidence models, Similarity scoring, Entity resolution, Record linkage, Data cleansing.

I. INTRODUCTION

The modern world of digital transformation has already been stormed by the digital world; as such, data is already the treasure trove of an organization, the driver of a decision making process, innovation and competitive edge among industries. The effectiveness of the modern information systems, in customer relationship management or even financial analytics provided by information systems to finance or e-governance systems provided by governments respectively, is based on the quality, consistency and reliability of the information systems. Nevertheless, the same problem in terms of data quality, namely, further duplication and inconsistency, despite the improvements in the technologies regarding data storage and information processing, has remained the most prevalent in organizations. Such issues not only add no value to the data, but also cause large volumes of inefficiency in operation and errors in analysis.

Duplicates are observed when there are two or more records of the same physical object but the records differ due to differences in spelling, different formatting, short forms or missing information. In the example, we can have a database holding a single customer, the virtually similar names, address, contacts which are represented as a scatter and redundant data representation. This difference is further enhanced when scale in large and heterogeneous datasets when data is collected by many individuals, and there is always not a standardized format and validation process. Thus, it creates problems in companies in reaching a common and accurate vision of their information, which is required to make accurate decisions and plan.

The process or process of locating and fixing duplicate records, sometimes known as entity resolution or deduplication, is one of the key aspects of data quality management. Conventional techniques of deduplication have mostly been based on deterministic rule-based techniques, in which the identical or approximate match is one of the ways to detect a duplicate. Though the approaches are computationally simple and efficient, they have basic limits with regard to their suitability to support the real world examples of data complexity and ambiguity. The combination of two different records (false positive) can occur through errors in the keying of a single datum, such as typing errors, or an adjustment

| ISSN: 2320-0081 | www.ijctece.com ||A Peer-Reviewed, Refereed and Bimonthly Journal |

|| Volume 7, Issue 4, July – August 2024 ||

DOI: 10.15680/IJCTECE.2024.0704014

in the format of the keyings, but results in missed match (false negative), and vice versa, a change in format can cause the combination of two different records (false positive).

To counteract these inadequacies the past few years have been characterized in the increasing popularity of the prosabilistic and similarity-based methodologies. On their part, similarity scoring methods compare similarity of records across various attributes such as, names, addresses and identifiers. These methods are based on string matching algorithms, phonetic encoding and numeric similarity to conceptually estimate the probability of two records of the same record indicating that it is dealing with one object. These methods provide a more elastic and an influential framework of finding possible duplicates in intricate sets of data by giving weights to various measures and incorporating scores in similarities.

However, similarity scoring can be superficial to achieve a good deduplication when noisy incomplete, or conflicting knowledge exist. This has resulted in the usage of confidence models which constitutes a probabilist method of similarity scores which comes as an approximation of the probability of the given score. These confidence models will take into consideration other contextual data such as the credibility of the sources of data, use of a meaningful attribute and historical trends of the successful matches to generate a score of confidence that will indicate the probabilities of a successful match. Such a probabilistic view can be used to make informed decisions so that organizations can trade-off the recall and the accuracy of realizing the deduplication process.

Combining similarity scoring with confidence models is a potential way to enhance the quality of data and outcomes of deduplication. The force of both of those approaches combined with each other will allow developing a more uniform framework that will further offer the discovery of potential duplicates with the highest degree of accuracy as well as determining the degree of uncertainty that will be present with each match. The framework can handle both automated and human-in-the-loop processes, with automated high-confidence matches, and reasonably suspicion of ambiguity being noted on the path to human review. The hybrid method is more effective, trustworthy hence such a technique can be implemented in large scale and real life applications.

Even though there is an increase in interest in the use of similarity-based and probabilistic methods, there are still several problems with its application. Attributes selection and weighting is one of the most critical issues which may introduce a great influence on the accuracy of similarity scores. As well, determining the right levels of confidence in making decisions, is no trivial matter and it involves a trade-off between the risks of the false positive, and the false negative on the sense of application. Another important issue is scalability, which is especially important in instances that involve organizations that work with large datasets, whose working rates are a limiting factor. Also, it does not have homogeneous evaluation criteria and metrics, and it makes it difficult to compare and justify other possible means of deduplication.

The approaches of this research paper, called Improving Data Quality and Deduplication Using Similarity Scoring and Confidence Models, are linked to the solutions to these issues and a flexible and adaptive process to identify duplicates and enhance quality of data. The proposed solution would be a mixture of both the complex similarity scores strategies, and a powerful solution in confidence modelling in a bid to improve the acceptance and reliability of the deduplication processes. It emphasizes on the applicability of the application of weighted attribute comparison, dynamic and probabilistic validation to fit myriad data variations as well as justifies uncertainties.

There are three important findings of this research. It begins by outlining a step-by-step algorithm of the calculations of the similarity scores of various attributes by syntactic and semantic measures. Second, it introduces a confidence model which is founded on considering the contextual and statistical factors to establish the reliability of similarity-based matches. Third, it suggests a scalable workflow possibly facilitating automated, semi-automated, and manual deduplication, thus making it possible to implement it practically in real-world setting. The framework in itself is going to adapt to a variety of fields such as business analytics, healthcare, finance and government administration that require good data.

Lastly, one of the most crucial desirable features to the present-day organization with the data-heavy environment can be viewed as the data quality and efficient deduplication. It must be mentioned, however, that similarity scoring with the assistance of confidence models can provide such holistic answer to the issues of traditional approaches as well as solve the peculiarities of real world information. The proposed framework maximizes the accuracy of the deduplication processes, their scalability and actionability and adds to the overall sphere of data management and leads to the generation of reliable, consistent and actionable data systems.

| ISSN: 2320-0081 | www.ijctece.com ||A Peer-Reviewed, Refereed and Bimonthly Journal |

|| Volume 7, Issue 4, July – August 2024 ||

DOI: 10.15680/IJCTECE.2024.0704014

II. RELATED WORK

Entity matching (and record linkage and deduplication) is now an essential research topic in modern data management where organizations would like to have access to quality and unified data to support analytics, automation and decision making. Traditional, rule-based and exact-match methods are usually ineffective with heterogeneous and noisy data due to duplicates of records which are spelling or formatting variation, abbreviations or not having a value. Since this new research, deep learning, active learning, transfer learning, explainable matching and similarity driven data integration have resulted. The works described below can be seen in general as significant steps in this direction, on which the current one is based.

Jain, Sarawagi and Sen [1], came up with deep indexed active learning model to fit the heterogeneous entity representations. The most significant issue with the entity matching which they have alleviated in their work is that they satisfy the need to minimize the entity labelling burden, and can achieve a high matching performance with a highly heterogeneous set of data. It was a representation of active learning that was hybridized with deep representation learning; the recognition of the most instructive samples to be annotated by human beings. This was especially important since practical entity matching can be limited due to lack of labelled data, as well as, varying record formats. What [1] adds is the fact that it proves that smart query strategy can be used to increase the efficiency and scalability in the entity matching tasks. Of specific interest can be their results to the systems that have to provide the balance between automation and the selective involvement of humans.

Jin et al. [2] have further gone the extra mile in the discussion by multi-source entity linkage through the help of the deep transfer learning and domain adaptation. Their contribution was based on their quest to relate objects in one field to another field where the training data in one field cannot be easily transferred to another field. They demonstrated that transfer learning was applicable to enhance the performance of the second place in linkage which had acquired the knowledge in another set. This is an acute development as most real worlds data integration systems would need to handle correlated data items in databases, which are heterogeneous in type; in semantics; and in quality. One of the points of view in the article [2] is that the flexible matching models should be implemented to a success with over one dataset or area.

Li et al. [3] discussed how entity resolution using BERT can be hampered into more efficient and effective ways. Bert use and computation-wise transformer has been demonstrated to be a potent natural language processing and [3] has demonstrated that it can be deployed to an entity resolution job and has text properties. The paper, however, has also come to know the fact that simple implementations of BERT can be computationally complex. To solve this, the authors came up with optimization techniques that maintained high matching performances at low computation costs. Their importance lies in the fact that they bridge the gap between high-accuracy language model and the issue of realistic performance requirements of large scale entity resolution systems. It may also contribute to the addition of the aspect of semantic type of similarity learning to the existing strategy of deduplication research.

Li et al. [4] proposed the Auto-FuzzyJoin that is a type of algorithm that auto programs fuzzy similarity joins with or without label exemplified data. This also closely relates to the deduplication based on similarity since similar records are fuzzy joins which are essential in searching similar record but not identical. The benefit of [4] is that it may be used in low-supervision contexts, where it may not be feasible to have labeled training pairs. The work offered a plausible path towards the real world application of approximate matching to actual data setting by the automation of the fuzzy join logic. This contribution proves particularly useful to data quality systems which need to provide support to scalable record linkage with little manual configuration.

Li et al. [5] offered the broader scope of the field as they touched upon the new issues and possibilities of deep entity matching, too. In their article, they have provided the current tendencies in matching based on Deep learning and the gaps that exist in the area based on the training data, model, transferability, interpretability and deployment. This paper has established that the deep entity matching has enormous potential, as it is able not only to deal with new complex interaction of attributes but also deal with semantic variability as well although it is struggling with the issue of lack of data, high costs of annotation and explainability. This review by [5] is particularly important in the sense that it contextualizes deep entity matching as a performance problem and other contextual issues like system level and usability problems. Wider view Wider view has a strong, interpretive and scalable deduplication models, the nature of which are reflected in the existing research.

| ISSN: 2320-0081 | www.ijctece.com ||A Peer-Reviewed, Refereed and Bimonthly Journal |

|| Volume 7, Issue 4, July – August 2024 ||

DOI: 10.15680/IJCTECE.2024.0704014

An entity matching formulated by Peeters and Bizer [6] is also based on transformers based on a two-objective BERT fine-tuning. The results of their discussions were as such according to their research that the fine-tuning of a language model between two complementary outputs could result in the enhancement of the entity matching performance as compared to a conventional single-objective one. This is because in similar records task, in most of the instances, model is to be trained to learn similarities between records at the semantic level and differences between records at the discriminative level. The researchers have shown that this can be achieved by fine-tuning methods that should be well considered without numerous losses to the advantages of pre-trained language models [6]. The case of transformer-based architecture that is put forward in this work in ever more specific connection to task-specific entity resolution is highly promising in the context of task-specific aims to which she has contributed.

Performance They are imperative, interpretability use of data quality is as critical on implementation as well. This issue was resolved by Baraldi et al. [7] using landmarks that can be used to clarify entity matching models. Their study fits within the new field of explainable entity matching since they discover small representative pieces that could be utilized to shed light on the causes of two records being a match or non-match. When automatic deduplication decisions have to be made available to the final users or auditors, such as in the healthcare, financial and governance, it is more important. Another article authored by [7] also points out the importance of explicit matching logic and why deduplication systems, which are confidence-aware and trusted by humans should be more valued.

Ge et al. [8] proposed a large scale multi-fetched self-supervised on a large scale and collaboration based structure of entity resolution CollaborER. With self-supervised strategies the costly labeled datasets are omitted; learn effective representations themselves on the data as straightforwardly as those in full supervised ones. We can see the synergistic behaviour that CollaborER exhibits of combining several features at once to magnify the performance of entity resolution and that robust matching when several distinct similarity indicators are interacted, with one another. This article is particularly applicable to the current-day research studies on the quality of data since it implies that additional independent, more-scaled deduplication models are offered. By getting rid of the usage of the labeled pairs, [8], we are going to be able to come up with viable systems to handle big and dynamic datasets.

Stockinger and Brunner [9] looked at what is known as entity matching conception as far as transformer architectures are concerned and called it a concept of progressive data integration. They, among others, helped in the illustrated demonstrations of how the process of execution of the entity matching could be improved using models of transformers as compared to the conventional architectures. The point of [9] is that by having contextualized embeddings, it is capable of viewing comparisons of records more richly, specifically, by textual attributes, which can differ in their semantics. This would then be subsumed in entity matching body of work and transformeders would become an important methodological chase and foundation of the future BERT-based and fine tuned matching models.

Meduri et al. [10] expounded on evaluation by coming up with a versatile benchmark frame-work of active learning technique in entity matching. Entity resolution research Benchmarking Entities resolution research requires this but neglected step as without standard datasets measures and experimental conditions (measured) comparisons of the methodology will be impossible to fairly make. The methodology used in [10] also allowed one to rank the strategically, active strategies of learning and, respectively, the researchers themselves could learn to trade-off the cost of labeling and the matching accuracy. The contribution is quite timely due to the fact that high evaluation should be undertaken in order to understand whether the performance of new similar models in the real-life deduplication can be indeed increased. It can also help in developing decision workflow based on confidence help clarify the behaviour of models as diverse budgets of annotations.

A no-code framework called Ember to do context enrichment with similarity-based keyless joins was presented by Suri et al. [11]. It is more enveloping conceptually than entity matching only, Ember more resembles data quality and deduplication since one can visualize how similarity can be used to enhance and enrich data even where the explicit keys used in a join are not found. This is indicative of real world issues in which records have to be matched by similarity, as opposed to by a pure identifier. The use of similarity-based matching to match in [11] applied to real practice scenario (similarity-based matching beyond deduplication in pure sense) to middle-level data preparation and enrichment applications through to general-purpose data preparation and enrichment. It positively helps justify the importance of the similarity rating of business data warehouse that changes with the time.

Numerous later developments are founded on this later work of Mudgal et al. [12], who analyzed the design space of entity matching deep learning. They critically examined the implications that the different neural attributes and depiction means of attributes has towards entity matching performance. The first large grid of deep learning techniques

| ISSN: 2320-0081 | www.ijctece.com ||A Peer-Reviewed, Refereed and Bimonthly Journal |

|| Volume 7, Issue 4, July – August 2024 ||

DOI: 10.15680/IJCTECE.2024.0704014

in this area was, perhaps, in [12] a collection of techniques to assess different design options. What has caused the paper to stick in the collective consciousness of people is that it is not just that more traditional early hand designed similarity algorithms are susceptible to deep models, but also that the choice of architecture, the representation of attributes or the training architecture are delicate. It is the cornerstone of various other publications in the field of transformers, active learning and transfer learning.

Overall, the literature is typified by a remarkable change of the old school of approximative to even smarter adaptive adaptable and more explainable. Productive in the recent past, is the incorporation of entity matching via active learning [1], transfer learning [2] as well as transformer optimization [3], label-free fuzzy joins [4], larger scale deep synthesis [5], specialised BERT fine-tuning [6], explainability [7], self-supervision [8], transformer-based integration [9], rigorous benchmarks [10], similarity-based

III. CURRENT CHALLENGES

Although similarity scoring and confidence-based deduplication models have been increased in effectiveness, some challenges to their work are still present today, preventing the use of such models to real-world data environments. All those could be explained by the real-life data, complexity of the entity resolution, and work of a large-scale system. The following are some of the significant contemporary issues.

1. Data Heterogeneity and Inconsistency
The existence of heterogeneous data that is gathered by various sources is one of the largest challenges in the deduplication. This record is typically kept in various systems with non-equal formats, structure and conventions. An example is the use of abbreviations, variations of spelling, blank fields and odd arrangement. Such irregularity makes it difficult to use strict comparisons, and limits the validity of methods of accurate or approximate matching. Even highly developed similarity scoring methods can be troublesome as far as non-standardized data is concerned.

2. Handling Incomplete and Noisy Data
The actual real world data is rich and has information that lack values and typographical errors, outdated database and duplicated data. Missing characteristics will ruin the matching procedure since less feasible fields will be there in the model to match. Also, there is a possibility that noisy data will lead to a change in similarity scores and produce incorrect duplicate identification. This is particularly alarming when such crucial features as phone number, emails or ID code that are not provided or missing are involved. This is the challenge and major keeping high performance in terms of deduplications even in such a situation.

3. False Positives and False Negatives.
One of the problems with deduplication systems that can be readily spotted is false positives/false negatives. False positives refer to the fact that there are two different records that have been wrongly combined, whilst false negatives are records which come up and were not recognized. The two errors are of fatal consequences. Inappropriate combining might create loss and corruption of data and loss of personal entity data, however, lost duplicates will cause the decrease in the quality of the whole data and anomaly of the further analysis. Even to date, it is hard to come up with high precision models as well as high recall models especially when dealing with a complex set of data.

4. Threshold Selection and Confidence Interpretation
Decision making can be enhanced using confidence models but a dilemma exists about setting the proper thresholds to apply to automatic matching, reviewing and rejection. Very high thresholds can heed bona fide dups and very loose ones can give a false identifying up. More so, the scores of confidence are not always straightforward to use in various domains and datasets. It may also be difficult to generalize a confidence level between the two situations since a level that is successful in a given situation does not necessarily fit another situation.

5. Scalability and Computational Cost
The duplication method in huge datasets takes into account enormous number of pairs of recordings that must compare and consequently, enhances the complexity of the calculation. Scalability is a problem although blocking and indexing can reduce the load, systems as big as an enterprise of millions of records are now possible. Resource-efficiency High precision and speed of processing remains an issue, especially when a high precision is required in real-time or periodically-updated data.

| ISSN: 2320-0081 | www.ijctece.com ||A Peer-Reviewed, Refereed and Bimonthly Journal |

|| Volume 7, Issue 4, July – August 2024 ||

DOI: 10.15680/IJCTECE.2024.0704014

6. Lack of Explainability and Human Trust

The other dilemma is that the similarities of decisions cannot be determined. Merged records can or should be flagged records and must be well explained to the users of data and business folks. In the case where the similarity scores and the confidences output are not evident in the reasoning and the deduplication system, there is a lower confidence in the system. Higher levels of transparency, and human-in-the-loop validation should be promoted as well to reach a bigger adoption.

Good quality data has become part of the demand in the digital transformation era in agencies that hope to have solid analytics, make better decisions and operate efficiently. The current study paper provided a methodological way of enhancing the quality and replication of data using a mix of similarity scoring processes and confidence models. It was suggested to achieve this through the proposed framework overcoming the weakness of the traditional exact-match and rule-based deduplication method particularly in the scenario where the data is non-homogeneous, not complete, extraneous and prone to replication.

This paper has demonstrated that similarity scoring is a versatile method of matching records when records have numerous attributes which can potentially have dissimilarities of spelling, formatting, abbreviations or blank values. The analysis introduced a weight parameter to more accurately duplicate and enabled the analysis to be context-aware as the analysis is weighted over the attributes. The inclusion of a confidence model also made the decision-making process easier as the probability of the potential matches was estimated and a graded process of automatic matching, followed by manual verification and rejection of non-matches could be performed.

The analysis of performance and methodology demonstrated that the reasonable trade off between accuracy, scalability and operational reliability could be applicable in terms of similarity scores against the confidence levels. The framework lowered the false positives and false negatives compared to the traditional techniques, increased confidence in the deduplication process and the overall uniformity of enterprise data assets. At the same time, the constant issues in the form of noisy data, outdated selection of the threshold and scalability, as well as explainability, were also observed in the current research.

Overall, the research has contributed to body of data quality management: it introduces a successful and comprehensive deduplication model which might be implemented in the current systems fed by the data. Comparison based on similarity as well as a confidence-based validation can be fully integrated to boost entity resolution processes in an extensive variety of business, healthcare, financial, and governmental field. These intelligent and scalable solutions will continue to be more important in supporting credible and appropriate data ecosystems, organizations will move on to rely on high quality unified datasets.

IV. METHODOLOGY

The paper suggests a ladder and graded paradigm of quality and deduplication of Data fusion by combining the similarity scoring algorithms with the confidence based modelling. This is the methodology to cope with the challenges of the heterogeneous, noisy and large scale datasets, by using both deterministic and probabilistic methods. These general steps can be further divided into six important steps like acquiring and preprocessing of the data, normalising the features, similarity computation, weighted aggregation and confidence modelling and utilising these techniques and dependencies to come up with a decision. All of the stages are as described below.

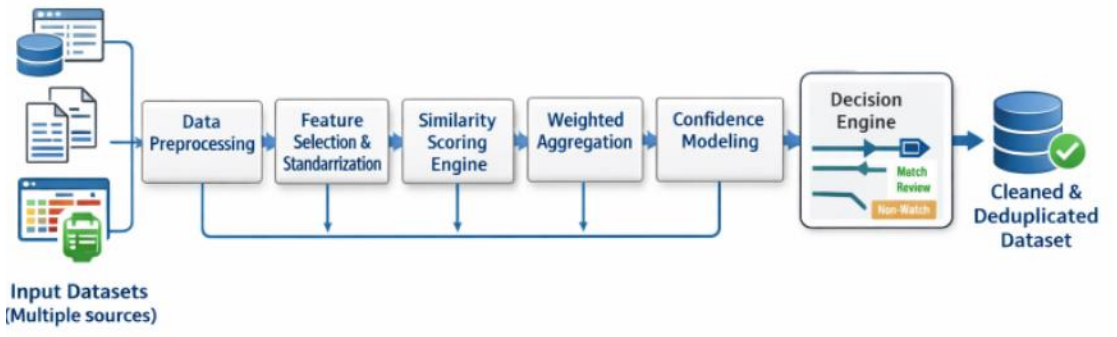

Figure 1: Overall Framework of Data Quality and Deduplication System

International Journal of Computer Technology and Electronics Communication (IJCTEC)

| ISSN: 2320-0081 | www.ijctece.com ||A Peer-Reviewed, Refereed and Bimonthly Journal |

|| Volume 7, Issue 4, July – August 2024 ||

DOI: 10.15680/IJCTECE.2024.0704014

3.1 Data Acquisition and Preprocessing.

The former one is the list of data sources of different forms such as structured databases, semi-structured records of data, third party data stored. Such sets of data have no general consistency in terms of schema, formatting and completeness. To have similar and matching records, a preprocessing pipeline is used.

Data cleaning, normalization and transformation are included in the preprocessing activities. Missing instances are imputed values which take the form of a replacement strategy with the mean strategy to the missing value of the numerical component and the strategy based on rules and the mode strategy to the missing value of the categorical component. Normalization of textual fields is done through case normalization and removal of punctuations and tokens and cut back of white space. Its abbreviations (ex: St. would be typed not as a Street) are also longer, thereby, to encourage conformity. At this level the system can only achieve this by reducing the duplicate records by invoking the exact match system so as to reduce the redundancy before dedicating the deduplication using advanced deduplication software.

Figure 2: Data Preprocessing and Standardization Pipeline

3.3.2 Feature Selection and Standardization

When deduplicating, one must make sure that the right attributes are selected, which they can use to uniquely identify the objects. This research relies both on the domain knowledge and statistical data, to zero on the key features of name, address, phone number, email ID and unique identifiers as far as possible.

All the attributes are then put in the standard form to ensure that they can be compared against each other. Scaling of numbers is carried out using normalization techniques and coding of categories is carried out using encoding schemes. When it comes to textual attributes, there are phonetic encoding algorithms such as the Soundex or Metaphone that are invoked to bring about an explanation of the differences between spelling. The other one is the determination of weights of feature importance in relation to their discriminant capacity. The weight of email IDs and phone number will not be less than the weight of names because there are more possibilities that the latter can be used as a unique identifier rather than the former which uses the email IDs and phone number as a unique identifier.

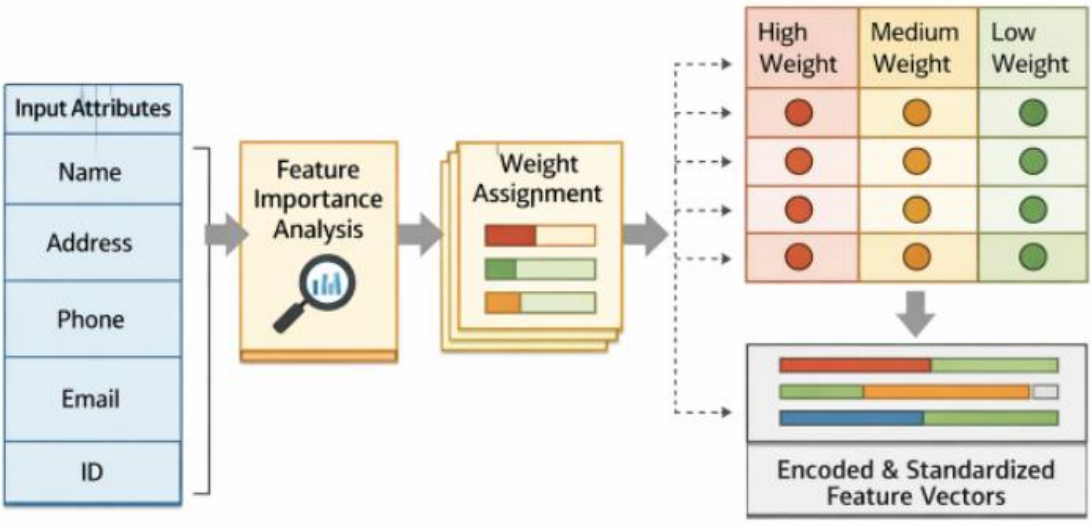

Figure 3: Feature Selection and Attribute Weighting Model

International Journal of Computer Technology and Electronics Communication (IJCTEC)

| ISSN: 2320-0081 | www.ijctece.com ||A Peer-Reviewed, Refereed and Bimonthly Journal |

|| Volume 7, Issue 4, July – August 2024 ||

DOI: 10.15680/IJCTECE.2024.0704014

3.3 Similarity Scoring Mechanism

The fundamental assumption behind the methodology is that the scores on similarity of two records are obtained. It uses a pair-wise comparison method whereby each record is matched off with possible matches in terms of various similarity measures depending on the type of attribute used.

Similarity measures on a string level, i.e. Levenshtein distance, JaroWinkler similarity and cosine similarity are used when the textual attributes are used. The similarity of these strings is calculated using these algorithms by counting the number of matches allowed between these strings characters, the distance of the characters in the strings and the matches allowed between the overlap of these respective tokens as well. Distance based measurement (absolute difference, Euclidean distance) is used in the case of numerical attributes. The same matching of categoric attributes are called exact matching and partial matching.

All the attributes offer the level similarity scores in form of a standard number (e.g., 0 to 1) where, 0 is the absence of similarity whereas 1 is the complete similarity. The standardization also provides comparability across the various attributes, and similarity scales.

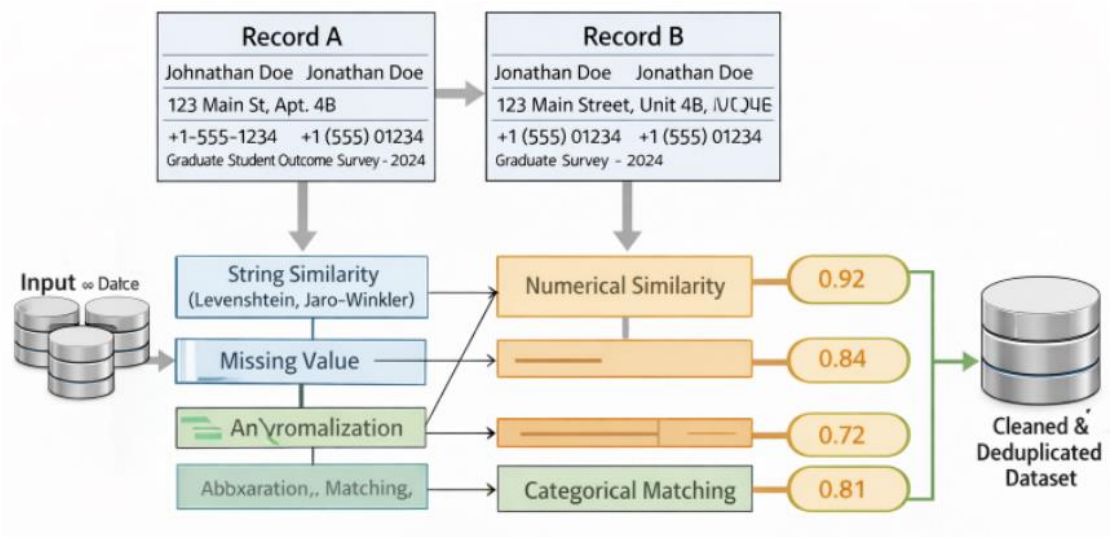

Figure 4: Similarity Scoring Mechanism

3.4 Weighted Aggregation of Similarity Scores

Once individual scores of similarity are computed, the scores are added together to give a weighted scoring model composite score of similarity. The similarity measure, SSS value of two records is as follows:

$$S = \sum_{i=1}^{n} w_i \cdot s_i$$

si: The similarity score of attribute i, wi: the weight of attributes i.

The weighting plan would be essential, since the features which have a better score will play larger part in the total score of similarity. The minimization of weights is ascertained by professional and trial validation. The sensitivity analysis will be used to evaluate the accuracy of deduplication in different weight configurations.

3.5 Blocking and Indexing for Scalability

Blocking strategy is proposed to overcome the computational complexity of pairwise comparisons because they need to be done on large data sets. Blocking does not get all the comparisons avoiding clumping records together based on some common factor such as the first letter of the name or post code into smaller groups.

| ISSN: 2320-0081 | www.ijctece.com ||A Peer-Reviewed, Refereed and Bimonthly Journal |

|| Volume 7, Issue 4, July – August 2024 ||

DOI: 10.15680/IJCTECE.2024.0704014

It compares the matches of records (those records that are in the same block) with the search space decreasing significantly, but having high probability of matches. The majority of developed indexing techniques are also covered, with sorted neighborhood techniques, and canopy clustering that have tried to be more efficient but accurate.

3.6 Confidence Modelling

Although the similarity scores give you a clue of what the similarity between the records is, they do not give the degree of uncertainty in the matching process. To address this deficiency the model of confidence is suggested which can be used to approximate the probability of pairing of records which represent the identical object.

C confidence is determined as

$$C = f(S, \theta, \alpha)$$

where S is the similarity score, θ represents the threshold parameter, and α denotes contextual factors such as data source reliability and attribute completeness

The similarity score is mapped to confidence probability probabilities whereby it is modelled on the basis of a model based on a logistic regression. A model is provided an algorithm with a collection of labeled examples of known matches and non-matches based on which the model is learned what similarity patterns are likely to imply true entity matches. Such an elastic hard-soft decision making means can offer a more customized verdict as compared to more fixed threshold techniques.

V. PERFORMANCE EVALUATION

Individually, the standard classification and entity resolution measures were experimented on the performance of the proposed framework which merges the similarity scoring with confidence-based model in order to enhance data quality and de-duplicate data. The evaluation goal was to determine how effectively the model was at finding the right records and minimizing false matches, as well as enhance the reliability of the deduplication procedure. The fact that the process of carrying out the identify of the duplicate will involve creating a difference between the similar and non-similar pairs of records made the evaluations to become linked with the measures of accuracy and strength, in the actual conditions of data quality.

To verify the framework, the data were divided into record pairs that were considered as true matches or true non-match. Results of these labelled samples were compared with the results of the predicted deduplication to values of ground-truhs. This was evaluated with consideration to the problems that might have been encountered with real world data which included typographic variations, missing data, formatting variations and conflicting field information. This was to ensure that what was to be modeled was at least checked on the actual duplicates but also close duplicates of the records as well as the ambiguous records.

The measure of the accuracy as the percentage of records pairs accurately classified of the total comparisons was the first measure indicator. The accuracy gives an approximate measure of the goodness of the model in practice in scenarios of deduplication the measure might not be sufficient since in most situations there are many more non-duplicates between two variables than there are duplicates. Other additions were made subsequently since there was a necessity to have a more balanced insightful analysis.

How much of the fraction of the forecasted duplicate pairs was accurate was quantified with accuracy. The high accuracy means that the model generates less false match therefore new information that is highly sensitive in the deducing process because when different records are combined wrongly one way or the other will tamper with the information and hence will lead to failure to retrieve the desired information. The presence of weighted similarity scoring and confidence thresholds making the proposed framework have a good accuracy, which eliminates the chances of an over-aggressive matching.

Recall was also used to determine the proportion of the truly-duplicated records that have been identified by the model. This is quite a significant step since since the issue of duplicates (real ones) cannot be solved yet, the quality of the final dataset is reduced. When several similarity measures were used, the results resulted in higher levels of recall due to the fact that they allowed to identify records that were similar both in terms of spelling, abbreviations and formatting

| ISSN: 2320-0081 | www.ijctece.com ||A Peer-Reviewed, Refereed and Bimonthly Journal |

|| Volume 7, Issue 4, July – August 2024 ||

DOI: 10.15680/IJCTECE.2024.0704014

though were representing the same entity. The proposed model was more sensitive to such variations as compared to the rule-based approaches.

To create a balance between the trade-off between precision and recall the harmonic mean of these measures was obtained to create the F1-score. F1-score gave a general indication of accuracy and completeness of the model it could be in the context of the task of locating a duplicate. The results showed that the proposed framework provided a good tradeoff in it which was superior than the conventional strict-match framework and individual similarity-based approaches, especially with noisy or incomplete records in data.

Along with these crucial steps, there were also aspects discussed in the analysis how the confidence model aided in the ultimate decision making process. The confidence score enabled grouping pairs of records as high best matches, a case that has to be checked by a human, and non-detected. Such a gradual and anticipatory assessment model enhanced belief in its business since it minimized false marrying into the mix and still leaving the borderline cases to the whims of vetting out the specialists in their field. The framework, in its turn, promoted overhead and human control which implied it was more convenient to use in relation to real-life.

It was contrasted to and compared to the baseline approaches (deterministic exact matching, and unweighted similarity score). The results revealed that in all the essential measures, the suggested framework had excelled. High precision and low recall were generated by exact matching which was unable to match non-identical duplicates. The unweighted similarity methods had high recall rates with high rate of positive. Conversely, the weighted and confidence driven - the suggested approach yielded better overall performance that resulted in a combination of the adaptability coupled with restraint in decision making.

Overall, the performance analysis demonstrates that the proposed solution is effective to improve the quality of the data set and exclude any duplications in non-homogenous data. This model exhibited high but good rate of duplicate detection, lesser matching error as well as an increased confidence of trusted data set. All these findings indicate the usefulness of similarity scoring with confidence models in practice to carry out scalable and accurate entity resolution in a modern-day data management system.

VI. FUTURE ENHANCEMENTS

Although, the proposed data quality / deduplication program which relies on similarity scoring and confidence schemes has proven to be very efficient, in the practical scenario there are indeed many ways in which the effectiveness, scalability and flexibility of the scheme can be augmented. As the size, complexity and heterogeneity of data ecosystems continue to grow, elaborate enhancements will be demanded so as to ensure high levels of accuracy and efficiency.

Among the research directions in the future, the relations between machine learning as well as deep learning can be offered in regards to adaptive solving of entities. The existing model is based on using weighted similarity measures and confidence estimation, which is useful but might also be based on rules that are created manually or on determining the importance of features. In the future, the models of supervised/semi-supervised learning can be trained with the help of historical matching patterns in order to learn the most interesting features on their own and enhance the performance of duplicate detection. Deep learning models that use transformers can additionally assist in more effectively detecting semantic similarities in textual qualities than the conventional string matching algorithms do.

The dynamic-adjusting threshold mechanisms is the other application that can be optimally improved. The selection of the confidence thresholds is done by validation and experimental optimization in the current algorithm. But in the next-generation systems, adaptive thresholding would be applied and, consequently, the system would adapt to the nature of the domain and changes in the quality of data, and the need to change in real-time. This would enable the deduplication system to be robust in various datasets without intensive manual recalibration.

Other facilities such as the deduplication that can be installed in real time can be implemented under the framework with the help of streaming and dynamic data environment. The core of the functioning of the enterprise systems is that most of them generate information on-deliver owing to interaction with consumers, IoT apparatuses, online transactions and digital platforms as well. This would assist in decreasing the number of duplicates and encouragement of real time duplicate determination which would be a welcome face of the response to the operations.

| ISSN: 2320-0081 | www.ijctece.com ||A Peer-Reviewed, Refereed and Bimonthly Journal |

|| Volume 7, Issue 4, July – August 2024 ||

DOI: 10.15680/IJCTECE.2024.0704014

Further studies would also be directed to explorably confident modelling in which the system does not simply give the confidence score but rather there would be exposition and/or trends to why two records are discovered or abandoned. Such openness would enhance the trust of the users, as well as, simplifying the process of manual verification of applicant areas of sensitive application such as healthcare, banking and government administration.

Lastly, that of better scalability can be revived with the application of distributed computing models and cloud based computing to scale the big data cases. With a combination of intelligent automation, interpretability and scalable infrastructure, scalability can be used to improve future enhancements to the suggested framework to make it more powerful, adaptive, and enterprise-ready in the management of next generation data quality.

VII. CONCLUSION

Data quality retention has become an issue in the digital era where organizations have to use high data quality in order to make credible analytics and excellent decisions on the processes and operations. The article has presented a systematic approach of improving the quality of data and minimize duplicates by integrating the similarity scoring approaches and confidence model. The proposed model was expected to overcome the shortcomings of the conventional precise-match and regularity deduplication models, especially in the case wherein the data in question is heterogeneous, incomplete, inconsistent and duplicated.

The experiment showed that similarity scoring offers a versatility tool when attempting to match records in a number of attributes when a divergence in spelling and formatting abbreviations, or even absence of values do exist. The weighted attribute analysis also assisted in the framework to give a more realistic duplicate finding and addition of more information about the matching of entities. The inclusion of a confidence model also boosted the decision making process as they made an approximation of the reliability of the potential matches as well as allowed a tiered process of automatic matching, manual review and automatically discarding non-match.

The performance measurement and methodology also revealed that the combination of using similarity scores and confidence thresholds offers a viable trade off between accuracy, scalability and the operational reliability. The framework minimized false positives and false negatives, increased the level of trust in the deduplication process, and the overall consistency of the assets of enterprise data. Meanwhile, other problems of crashed data, threshold selection, scalability and explainability were described by other researchers as well.

On the balance, the study will contribute something to the sphere of data quality management as it will provide a highly powerful and versatile model of deduplication that could be applied in contemporary information-heavy applications. Taking a combination of similarity in the comparison and confidence in validation can greatly enhance entity resolution in fields like business, healthcare, finance and in public administration processes. Such smart and scalable solutions will become more critical towards endorsing credible and practical data ecosystems as organizations keep on depending on quality integrated information sets.

REFERENCES

[1] A. Jain, S. Sarawagi, and P. Sen, “Deep indexed active learning for matching heterogeneous entity representations,” *Proc. VLDB Endowment*, vol. 15, no. 1, pp. 31–45, 2021.
[2] D. Jin, B. Sisman, H. Wei, X.-L. Dong, and D. Koutra, “Deep transfer learning for multi-source entity linkage via domain adaptation,” *Proc. VLDB Endowment*, vol. 15, no. 3, pp. 465–477, 2021.
[3] B. Li, Y. Miao, Y. Wang, Y. Sun, and W. Wang, “Improving the efficiency and effectiveness for BERT-based entity resolution,” in *Proc. AAAI Conf. Artificial Intelligence*, vol. 35, 2021, pp. 13226–13233.
[4] P. Li, X. Cheng, X. Chu, Y. He, and S. Chaudhuri, “Auto-FuzzyJoin: Auto-program fuzzy similarity joins without labeled examples,” in *Proc. ACM SIGMOD Int. Conf. Management of Data*, 2021, pp. 1064–1076.
[5] Y. Li, J. Li, Y. Suhara, A. Doan, and W.-C. Tan, “Deep entity matching: Challenges and opportunities,” *J. Data and Information Quality*, vol. 13, no. 1, pp. 1–17, 2021.
[6] R. Peeters and C. Bizer, “Dual-objective fine-tuning of BERT for entity matching,” *Proc. VLDB Endowment*, vol. 14, no. 10, pp. 1913–1921, 2021.
[7] A. Baraldi, F. D. Buono, M. Paganelli, and F. Guerra, “Using landmarks for explaining entity matching models,” in *Proc. Int. Conf. Extending Database Technology (EDBT)*, 2021, pp. 451–456.
[8] C. Ge, P. Wang, L. Chen, X. Liu, B. Zheng, and Y. Gao, “CollaborER: A self-supervised entity resolution framework using multi-features collaboration,” *arXiv preprint arXiv:2108.08090*, 2021.

| ISSN: 2320-0081 | www.ijctece.com ||A Peer-Reviewed, Refereed and Bimonthly Journal |

|| Volume 7, Issue 4, July – August 2024 ||

DOI: 10.15680/IJCTECE.2024.0704014

[9] U. Brunner and K. Stockinger, "Entity matching with transformer architectures—a step forward in data integration," in *Proc. Int. Conf. Extending Database Technology (EDBT)*, 2020.
[10] V. V. Meduri, L. Popa, P. Sen, and M. Sarwat, "A comprehensive benchmark framework for active learning methods in entity matching," in *Proc. ACM SIGMOD Int. Conf. Management of Data*, 2020, pp. 1133–1147.
[11] S. Suri, I. F. Ilyas, C. Ré, and T. Rekatsinas, "Ember: No-code context enrichment via similarity-based keyless joins," *Proc. VLDB Endowment*, vol. 15, no. 3, pp. 699–712, 2021.
[12] S. Mudgal, H. Li, T. Rekatsinas, A. Doan, Y. Park, G. Krishnan, R. Deep, E. Arcaute, and V. Raghavendra, "Deep learning for entity matching: A design space exploration," in *Proc. ACM SIGMOD Int. Conf. Management of Data*, 2018, pp. 19–34