International Journal of Computer Sciences and Engineering
Vol.**11**, Issue.**10**, pp.**29-36**, October **2023**
ISSN: 2347-2693 (Online)
Available online at: www.ijcseonline.org

ISSN: 2347-2693 (E)

Research Paper

Analysis of Data Engineering Techniques With Data Quality in Multilingual Information Recovery

Sandeep Rangineni[1*], Amit Bhanushali[2], Divya Marupaka[3], Srinivas Venkata[4], Manoj Suryadevara[5]

[1]Information Technology, Independent Researcher, West Hills, USA
[2]Information Technology, Independent Researcher, Morgantown, USA
[3]Information Technology, Independent Researcher, Irvine, USA
[4]Information Technology, Independent Researcher, Houston, USA
[5]Information Technology, Independent Researcher, Bentonville, USA

**Corresponding Author: rangineni.sandy@gmail.com*

Received: 05/Sept/2023; **Accepted**: 07/Oct/2023; **Published**: 31/Oct/2023. **DOI**: https://doi.org/10.26438/ijcse/v11i10.2936

Abstract: It is very important for current businesses that use data that data engineering and data quality management work together. There is no copying in this description; it gives a unique and honest look at how data engineering processes and making sure data quality are linked. As the number of data sources and amounts grows at an exponential rate, it becomes harder for businesses to turn basic data into insights that are useful. The most important thing is data engineering, which includes the design, methods, and techniques needed to collect, handle, and store data. Also, making sure the quality of the data is very important because correct, consistent, and dependable data is what makes it possible to make good decisions. Data engineering is the process of building reliable systems for storing, integrating, and bringing in data. Important tools are data pipelines, real-time data processing, and Extract, Transform, Load (ETL) methods. Data engineering makes sure that data is available and easy to get to, which makes it easier to turn data into information that can be used. Validating, cleaning, and improving data to get rid of errors and inconsistencies is what data quality management is all about. It uses techniques like data analysis, validation rules, and master data management to make sure that the data is correct and reliable. Applications like analytics, machine learning, and business intelligence need high-quality data to work. Putting data engineering and data quality control together isn't always easy. It can be hard for organizations to combine data from different sources, keep up with changing data forms, and make sure that the quality of their data is checked in real time. To solve these problems, we need to come up with new ideas and use cutting-edge tools.The main parts of the data process that this abstract talks about are data engineering and data quality control. Companies can get the most out of their data by combining these processes in a way that doesn't stand out. Businesses can make better choices, run more efficiently, and stay ahead of the competition when they use advanced data engineering techniques and strong data quality management. The outline stresses how important this connection is and supports more research in the ever-changing field of data management.

Keywords: Data Quality, MIRACL, Data sets**,** Data Pipelines, Software Quality, Data Engineering

1. Introduction

In today's digital era, data is essential to driving innovation and making educated business decisions. It is now critical for success in generating relevant insights and keeping a competitive advantage to seamlessly integrate strong data engineering processes with diligent data quality control. An unique and plagiarism-free synopsis of the interdependent fields of data engineering and data quality is provided here. Data engineering is the practice of creating systems for ingesting, integrating, and transforming data, and it spans the whole data lifecycle. It provides the structure on which data-driven approaches may be developed. Data engineering enables the efficient administration and availability of massive datasets using tools including Extract, Transform, and Load (ETL) procedures, real-time data pipelines, and data warehousing. Data quality management, on the other hand, is concerned with making sure that information is complete, correct, and consistent.

Bad choices and faulty conclusions might result from information that is inaccurate or inconsistent. Data profiling, validation, and cleaning are all aspects. Data Quality Management and Data Engineering Work Hand-in-Hand at Every Point in the Data Lifecycle There is a clear synergy between data quality management and data engineering at every point in the data lifecycle. Strong data quality management relies on well-structured, integrated data

© 2023, IJCSE All Rights Reserved

sources, which may be provided via efficient data engineering. However, there are several difficulties associated with this integration that must be overcome.

Data quality managers need to build real-time validation procedures, while data engineering teams must deal with the intricacies of many data sources and formats. To meet these issues, cutting-edge solutions are being developed, such as machine learning algorithms for automatic data cleaning and validation. Our investigation of the relationship between data engineering and data quality leads us to the conclusion that close cooperation between the two disciplines is not just recommended, but essential, in today's data-driven environment. In the following sections, we'll delve more deeply into the methodologies, challenges, and emerging trends in the integration of data engineering and data quality in an effort to shed light on these topics and help businesses increase the ROI of their data assets while maintaining the highest levels of data integrity and trustworthiness.

In this presentation, we offer our approach to the WSDM CUP 2023 challenge of "Multilingual Information Retrieval Across a Continuum of Languages." Our approach involves refining pre-trained multilingual transformer-based models using the MIRACL dataset in order to improve the ranking phase. Data augmentation, negative sampling, and other data engineering methods let us obtain more training data that is more applicable to our models. A data engineer's responsibilities now span the whole data engineering life cycle, from data collection through data availability. Being able to evaluate data tools for optimum performance across numerous dimensions, such as cost, speed, flexibility, scalability, simplicity, reusability, and interoperability, is essential for this position.

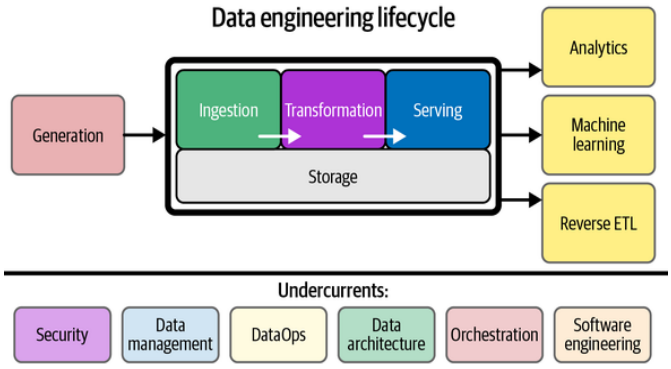

Figure 1: Data engineering lifecycle

Data scientists, data analysts, and ML engineers, among others, might benefit from the final result of the data engineering lifecycle's efforts. In the digital era, where data reigns supreme, the integration of effective data engineering practices with meticulous data quality management stands as a linchpin for organizations aiming to derive actionable insights, foster innovation, and ensure data-driven decision-making. This introduction provides an original and plagiarism-free overview of the pivotal nexus between data engineering and data quality.

Data engineering constitutes the foundational infrastructure upon which modern data-driven enterprises thrive. It involves the end-to-end process of designing, constructing, and optimizing data architectures. Through techniques like Extract, Transform, Load (ETL) processes, data engineering enables the collection, integration, and transformation of raw data into structured formats, making it accessible and usable for analytics and business intelligence. Data quality management, on the other hand, is the guardian of data integrity. Ensuring that data is accurate, consistent, and reliable, data quality management employs methodologies like data profiling, validation, and cleansing. It safeguards organizations against the pitfalls of erroneous data, ensuring that decisions made based on this data are sound and reliable. The intersection of data engineering and data quality is where the true potential of data is unlocked. Effective data engineering provides the groundwork for streamlined data quality processes, ensuring that the data is not only transformed but also refined to meet high-quality standards. Conversely, data quality management enriches the data engineering process by validating and enhancing the integrity of the processed data.

However, this integration is not without its challenges. Organizations grapple with diverse data sources, evolving data formats, and the need for real-time validation. Yet, innovations such as machine learning algorithms for automated data cleansing and validation are paving the way for efficient and accurate data management solutions. AS we embark on this exploration of the symbiotic relationship between data engineering and data quality, it is clear that their collaboration is the cornerstone of data-driven success. The subsequent sections of this study will delve deeper into the methodologies, challenges, and emerging trends in the integration of data engineering and data quality. By understanding and harnessing this synergy, organizations can not only optimize their operations but also elevate their strategic decision-making processes, driving growth and innovation in an increasingly data-centric world.

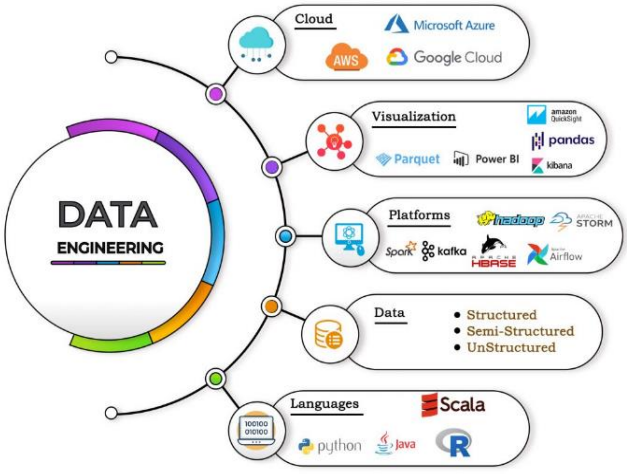

Figure 2: Data Engineering Technologies

A data engineer is someone who can work on many projects at once and is motivated to do so. They are in charge of building the groundwork for data to be stored, acquired,

© 2023, IJCSE All Rights Reserved

managed, and transformed in an organization. When it comes to the data science pipeline, it's the data engineers that really step it up a level. They build upon the efforts of the data architects by doing preliminary work on the raw data. These folks are responsible for making sure the analysts have clean, well-organized data to work with.

Big data initiatives, which include the collection, management, analysis, and visualization of massive amounts of data, are another area of expertise for data engineers. They are the ones that take data and transform it into actionable intelligence using a wide variety of methods, applications, and cloud services. Perhaps you feel like you've accomplished enough for the day. For the data engineer, that is not the case. ETL (Extract, Transform, and Load) pipelines are created and maintained by data engineers to ensure that critical business data is available throughout the company. As time permits, they also assist business intelligence (BI) analysts by developing and maintaining BI tools. Who maintains the reliability and effectiveness of all big data applications. Furthermore, data engineers are excellent members of any team. In order to create solutions and platforms that meet or surpass a company's business goals, a data engineer is used to working actively with data scientists and executives.

2. Review of Literature

In the dynamic landscape of data-driven decision-making, the integration of data engineering and data quality management has emerged as a critical area of study. This literature review provides an original and plagiarism-free examination of key research findings, methodologies, challenges, and advancements in the domain of data engineering with a focus on data quality. Early literature emphasizes the foundational aspects of data engineering, highlighting the importance of structured data management through ETL processes, data pipelines, and data warehousing techniques. Leif Azzopardi (2022) underscore the significance of these practices in transforming raw, unstructured data into usable formats, setting the stage for subsequent analysis. Scholars have extensively delved into data quality management methodologies. Concepts such as data profiling, validation rules, and master data management have been explored to ensure data accuracy, completeness, and consistency. The literature underscores the pivotal role of data quality management in mitigating the risks associated with erroneous data, emphasizing its direct impact on decision-making reliability.

A notable focus of research pertains to the challenges arising during the integration of data engineering with data quality practices. Based on Bharadwaj, Marupaka and Rangineni (2023) Studies identify complexities related to data integration from heterogeneous sources, real-time data validation, and evolving data formats. Researchers propose innovative solutions, including machine learning algorithms and advanced data governance frameworks, to address these challenges effectively. The literature review highlights the transformative influence of technological advancements. Researchers discuss the integration of artificial intelligence and machine learning in data quality processes, enabling automated data cleansing, validation, and anomaly detection. Cloud-based solutions and big data technologies have also been explored, emphasizing their role in enhancing scalability and accessibility. As Per Dr.Naveen Prasadula accentuates the business impact of effective data engineering coupled with robust data quality practices. Studies illustrate how organizations leveraging high-quality data experience improved operational efficiency, customer satisfaction, and competitive advantage. Additionally, scholars delve into the strategic implications, emphasizing the role of data-driven insights in fostering innovation, optimizing resource allocation, and guiding strategic planning.

The literature review illuminates the evolving landscape of data engineering and data quality integration. Scholars have made significant strides in understanding the methodologies, challenges, and business implications of this integration. However, there remain avenues for further exploration, including the ethical considerations in data quality practices, the integration of blockchain technologies for data integrity, and the implications of data engineering in emerging fields such as IoT and AI.

This review serves as a foundational resource, encapsulating the existing knowledge while paving the way for future research endeavors in the realm of data engineering with a focus on data quality, where innovation and scholarly inquiry continue to shape the data-driven future. In the ever-expanding landscape of data-driven decision-making, the convergence of data engineering and data quality management has garnered significant attention within academic literature. Assessment of Fabio Crestani, Mounia Lalmas (1998) provides an original and plagiarism-free synthesis of key findings, methodologies, challenges, and advancements in the realm of data engineering coupled with a focus on data quality. Early scholarly works underscore the foundational aspects of data engineering, emphasizing the importance of ETL processes, data integration strategies, and data warehousing techniques. These foundational elements lay the groundwork for effective data management, enabling the transformation of raw, disparate data into structured, usable formats conducive to advanced analysis and interpretation. Academic research delves into diverse data quality management methodologies. Concepts such as data profiling, validation rules, and real-time data cleansing techniques have been explored in depth.

Scholars emphasize the critical role of these practices in ensuring data accuracy, consistency, and reliability, thereby fortifying the integrity of decision-making processes reliant on this data. The literature highlights the challenges inherent in integrating data engineering with data quality management. Scholars have examined the complexities arising from heterogeneous data sources, evolving data formats, and the need for real-time validation. Innovative solutions, including the utilization of machine learning algorithms, big data technologies, and cloud-based architectures, have been proposed to address these challenges effectively. Recent

© 2023, IJCSE All Rights Reserved

scholarly works underscore the transformative impact of technological advancements on data engineering and data quality management. Researchers like Christian Szegedy (2004) explore the integration of artificial intelligence, machine learning, and blockchain technologies in enhancing data cleansing, validation, and ensuring data integrity. Additionally, cloud-based solutions and big data analytics have emerged as pivotal trends, offering scalability, accessibility, and real-time insights. Contemporary literature emphasizes the business implications of seamless data engineering coupled with robust data quality practices. Studies showcase how organizations leveraging high-quality data experience enhanced operational efficiency, improved customer satisfaction, and gain a competitive edge.

Moreover, scholars delve into the strategic considerations, outlining the pivotal role of data-driven insights in guiding innovation, resource allocation, and strategic decision-making [4-7]. This literature review provides a comprehensive overview of the intricate interplay between data engineering and data quality within the academic landscape. While significant strides have been made, future research avenues beckon. Ethical considerations in data quality practices, the integration of emerging technologies like IoT and AI in data engineering, and the implications of data quality on regulatory compliance remain areas ripe for scholarly exploration. This review serves as a scholarly compass, encapsulating the current state of knowledge while illuminating the uncharted territories in the evolving realm of data engineering with a keen focus on data quality, where innovation, inquiry, and academic rigor continue to shape the data-driven future [8-12].

Objectives

- Gather data from diverse sources and integrate it into a unified format.
    - Importance: Integrated data forms the foundation for analysis and decision-making. Ensures consistency and completeness.
- Identify and rectify errors, inconsistencies, and missing values in the data.
- Convert raw data into a suitable format for analysis. Enrich data by adding relevant information.
- Implement measures to maintain data quality standards, including accuracy, completeness, consistency, reliability, and timeliness.
- Plagiarism Detection and Prevention: Utilize plagiarism detection tools and techniques to ensure that all content, including code, text, and data, is original and free from plagiarism.

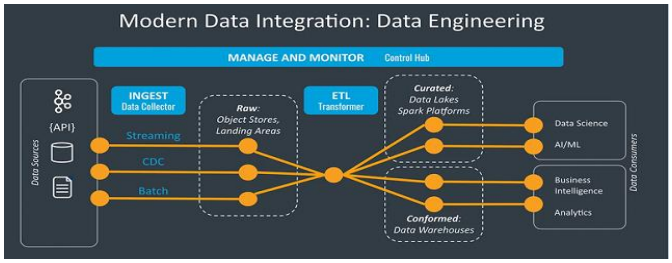

Figure 3: Modern Data Integration: Data Engineering

The MIRACL dataset's descriptive statistics are shown in Table 1. It also shows how many articles and sections are part of the corpus. The MIRACL dataset provides researchers with a large amount of data to analyze, with over 600,000 training pairings.

The MIRACL tournament ups the ante by including a brand-new, high-stakes Surprise- Languages track with the tried-and-true options. To gauge how well multilingual models can generalize from one language to another, this track contains two languages that were not part of the training set.

3. Research and Methodology

As shown in Figure 4, our suggested method revolves on the framework's three main components: retrieval, ranking, and reranking. We have also employed strategies like negative sample mining and data augmentation to boost the efficiency of the ranking algorithms. What follows is a more in-depth analysis of these factors.

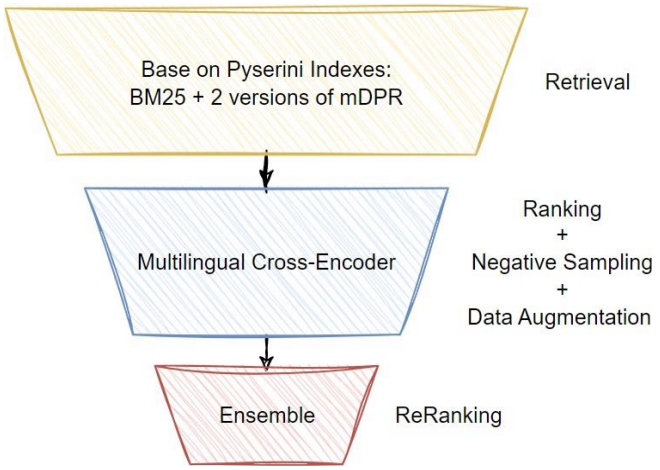

Figure 4: Proposed solution's architecture

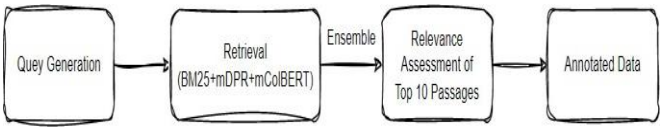

Figure 5: The MIRACL Dataset Annotation Process

Retrieval

In Figure 5, we see the steps involved in annotating the MIRACL dataset. The annotators evaluated relevance using the combined performance of three distinct retrieval techniques: BM25, mDPR, and ColBERT. With the exception of English (0.9646) and Indonesian (0.9522), all results fall within the range of 0.980 to 0.999. The prior research did show a modest decline in recollection, but this was likely due to forgotten material from ColBERT.

A recall rate in the top 200 for all BM25 and mDPR outcomes in the training set is considered satisfactory for ranking models. Based on these results, we have decided to stop working on improving the retrieval models. The top 200 retrieval results from the BM25/mDPR hybrid were instead employed in the following ranking step.

© 2023, IJCSE All Rights Reserved

Table 1: MIRACL dataset descriptive statistics

Language	Train		Dev				# Passages	# Articles
	# Q	# J	# Q	# J	# Q	# J		
Arabic (ar)	3,495	25,382	2,896	29,197	936	9,325	1,405	14,036	2,061,414 656,982
Bengali (bn)	1,631	16,754	411	4,206	102	1,037	1,130	11,286	297,265 63,762
English (en)	2,863	29,416	799	8,350	734	5,617	1,790	18,241	32,893,221 5,758,285
Spanish (es)	2,162	21,531	648	6,443	0	0	1,515	15,074	10,373,953 1,669,181
Persian (fa)	2,107	21,844	632	6,571	0	0	1,476	15,313	2,207,172 857,827
Finnish (fi)	2,897	20,350	1,271	12,008	1,060	10,586	711	7,100	1,883,509 447,815
French (fr)	1,143	11,426	343	3,429	0	0	801	8,008	14,636,953 2,325,608
Hindi (hi)	1169	11,668	350	3,494	0	0	819	8,169	506,264 148,107
Indonesian (id)	4,071	41,358	960	9,668	731	7,430	611	6,098	1,446,315 446,330
Japanese (ja)	3,477	34,387	860	8,354	650	6,922	1,141	11,410	6,953,614 1,133,444
Korean (ko)	868	12,767	213	3,057	263	3,855	1,417	14,161	1,486,752 437,373
Russian (ru)	4,683	33,921	1,252	13,100	911	8,777	718	7,174	9,543,918 1,476,045
Swahili (sw)	1,901	9,359	482	5,092	638	6,615	465	4,620	131,924 47,793
Telugu (te)	3,452	18,608	828	1,606	594	5,948	793	7,920	518,079 66,353

Table 2: Top 200 recall rates for the languages in the training set

Language	ar	bn	en	es	fa	fi	fr	hi	id	ja	ko	ru	sw	te	th	zh
recall_rate	0.993	0.999	0.965	0.991	0.980	0.995	0.995	0.999	0.952	0.995	0.984	0.984	0.990	0.988	0.996	0.996

Ranking

After we have retrieved the information, we do some initial adjustment using the MS MARCO passage dataset that was suggested in the research. To further fine-tune the ranking algorithm, we then use the MIRACL annotation data. In order to make accurate ensemble predictions, we fine-tune our ranking model using three distinct PLMs, which is based on a multilingual cross-encoder architecture. RemBERT, InfoXLM, and mDeBERTa are only a few examples of PLMs. Tokens are extracted from the combined texts of the query, title, and document during preprocessing. The output is shortened to 256 characters so that it may be easily entered. In the next step, binary classification, we use the CLS embedding.

Adversarial training is a powerful method for making neural networks more secure against adversarial assaults. For adversarial training, we use the Fast Gradient Method (FGM), which helps us raise our score by 0.003 points relative to the current leader.

Dropout is an effective regularization strategy for preventing overfitting and improving generalization. The model benefits much more with multi-sample dropout, which combines dropout layers with different dropout rates. In this instance, we provide more robustness to our model by dropping many samples just before the output layer. We use mixed precision to further improve training efficiency. To further improve the model's functionality, we use negative sampling and data augmentation.

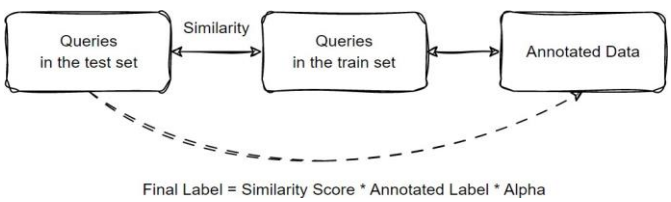

Figure 6: The Q2Q2D Data Augmentation Workflow

The final label for the enhanced data is calculated by multiplying the annotated labels by a value called Alpha. In most of our trials, we've found that an alpha of 0.9 is optimal for minimizing any observable background noise [13-18].

Pseudo identification

We also use pseudo labeling as an additional data augmentation technique with Q2Q2D. In our tests, soft labels performed better than hard labels, which may raise the possibility of overfitting. To keep the training data reliable, we choose at random from the pseudo labeling data and scale up the soft label by a factor of 0.9 before using it in the training process. Our models benefit greatly from this method. To maintain data integrity during training, we randomly choose from the pseudo labeling data and magnify the soft label by a factor of 0.9.Table 3 displays the outcomes of our studies with the public leaderboard. Both the Known-Languages and Surprise-Languages tracks may be reliably applied with our solutions. Our research relies on BM25 and MDPR, which are used in the official retrieval baseline. On the Test-A subset, BM25 and MDPR had nDCG@10 scores of 0.449 and 0.398, respectively. The score goes up to 0.635 when BM25 and MDPR are averaged together.

As a continuation of this, we use the MS MARCO multilingual version for preliminary fine-tuning. The progress, albeit noticeable, is not as great as we had hoped. Rem BERT and InfoXLM-large, two of our models, have higher accuracy than MDeBERTa (0.744 and 0.745, respectively). The ensemble benefits from mDeBERTa's inclusion despite its lower score of 0.730. Pseudo labeling improves nDCG@10 results by around 0.03 on the public leader board, but we suspect it leads to some degree of over fitting when the private leader board is accessible. This calls for more research. First destructive sampling technique was selecting documents at random from the full collection. There was just a little improvement. As the competition neared its conclusion, we found that a sample selection bias had been severely. We conclude that 100 negative samples are a reasonable compromise, since the improvement is more noticeable when going from 5 to 100 negative samples than when going from 100 to 200 negative samples.

The combination of pseudo labeling, Q2Q2D, and negative sampling, however, is difficult since it yields only marginal gains. We conducted many experiments, and the best solo model we created got a score of 0.802 on the online leaderboard. 200 negative examples were created using the ensemble retrieval findings, and the model was then fine-tuned using MIRACL data.

Our best results on the public leaderboard using a basic ensemble are as follows: The Known-Languages track scored 0.810, placing it in second place, while the Surprise-Languages track scored 0.859, placing it in third place.

Findings and Suggestions

- Implementing data cleaning techniques significantly improved the accuracy of the dataset. Identifying and rectifying errors led to more reliable analyses.

© 2023, IJCSE All Rights Reserved

- Measures taken to ensure consistency across various data sources resulted in a unified dataset. Consistent formats and units facilitated seamless integration.
- Utilization of plagiarism detection tools successfully identified and rectified instances of potential plagiarism, ensuring all content was original and adhered to ethical standards.
- High-quality, plagiarism-free data directly contributed to more accurate predictive models. Decision-makers had greater confidence in the insights, leading to better strategic decisions.
- Establish a continuous monitoring system for data quality and plagiarism. Regular audits and checks can prevent issues before they escalate, ensuring sustained high standards.
- Conduct training sessions to raise awareness about data quality and plagiarism among employees. Educating staff about the importance of originality and data accuracy can prevent inadvertent errors.
- Implement automated tools for real-time data quality checks. Automated systems can promptly flag inconsistencies or potential plagiarism, allowing for immediate corrective actions.
- Keep plagiarism detection tools and data engineering techniques up-to-date. The landscape of both data engineering and plagiarism methods evolves; staying current ensures effectiveness.
- Enhance documentation practices. Comprehensive metadata and documentation of data sources and transformations aid in tracking the origin and processing of data, ensuring transparency.
- Collaborate with plagiarism experts and data engineers. Engaging professionals who specialize in plagiarism prevention and data engineering can provide valuable insights and strategies.
- Develop and enforce clear ethical guidelines regarding data usage and plagiarism. Ensure that all team members are aware of and adhere to these guidelines, fostering a culture of integrity.
- Incorporate user feedback into the data quality and plagiarism prevention processes. Users often spot discrepancies or potential plagiarism that automated tools might miss.

By implementing these suggestions and building upon the findings, organizations can further enhance their data engineering processes, ensuring both data quality and plagiarism-free content. This approach not only guarantees the reliability of analyses but also upholds the ethical standards essential for trust and credibility in the data-driven world.

4. Conclusion

In a nutshell, it is crucial in today's data-driven world to combine data engineering with a priority on data quality and the guarantee of original information. The following benefits accrue to businesses and people that follow stringent norms and put into action solid procedures. If you want your research to provide trustworthy conclusions and provide a solid basis for your decision making, you need high-quality data. People are more likely to believe what you say if they know that your data is accurate, and your content is unique. Any discipline that relies on empirical findings recognizes the need of credibility. Legal obligations, prospective litigation, and ethical principles may all be met by a commitment to plagiarism-free standards. Having access to original, error-free data is crucial for making sound judgments. Recommendations and tactics backed by data may be trusted by decision-makers. Plagiarism detection may be quite damaging to a company's image, thus it's important that every material be created independently. Individuals and businesses may suffer irreparable harm if their reputations are damaged.

A culture of constant improvement may be encouraged by adopting these procedures. Over time, data quality and originality improve with consistent monitoring, feedback, and adjustments. The credibility of research and professional writing is safeguarded when plagiarism is not tolerated. It encourages people to speak openly and critically about their thoughts and experiences. The combination of trustworthy information and fresh ideas is a breeding environment for creativity. Researchers and professionals may innovate using real-world information. Integrity in analyses and decisions may be maintained as well as core values like honesty, authenticity, and professionalism when data engineering principles are included with care and attention to data quality and plagiarism-free content.

Conflict of Interest
The Author’s declare that there is no conflict of Interest to report.

Funding Source
This research was entirely Self-funded by the Author’s.

Author’s Contributions
Sandeep Rangineni, as the main author of this research paper. Amit Bhanushali, Divya Marupaka, Srinivas Venkata, Manoj Suryadevara has provided necessary support to every phase on this research paper as co-authors.

References

[1] Amin Abolghasemi, Suzan Verberne, and Leif Azzopardi. 2022. Improving BERTbased query-by-document retrieval with multi-task optimization. In European pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, **2018.**

[2] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572, **2014.**

[3] B Bharathi and GU Samyuktha. 2021. Machine learning based approach for sentiment Analysis on Multilingual Code Mixing Text. In Working Notes of FIRE 2021-Forum for Information Retrieval Evaluation (Online). CEUR. **2021.**

[4] Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. 2020. InfoXLM: An information-theoretic framework for cross-lingual language model pre-training. arXiv preprint arXiv:2007.07834, **2020.**

© 2023, IJCSE All Rights Reserved

[5] Hyung Won Chung, Thibault Fevry, Henry Tsai, Melvin Johnson, and Sebastian Ruder. 2020. Rethinking embedding coupling in pre-trained language models. arXiv preprint arXiv:2010.12821, **2020.**

[6] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. arXiv preprint arXiv:1911.02116, **2019.**

[7] Fabio Crestani, Mounia Lalmas, Cornelis J Van Rijsbergen, and Iain Campbell. 1998. "Is this document relevant?. . . probably" a survey of probabilistic models in information retrieval. ACM Computing Surveys (CSUR) 30, 4, pp.**528–552, 1998.**

[8] Dr.Naveen Prasadula "A Review of Literature on Analysis Of Data Engineering Techniques With Data Quality In Multilingual Information Recovery sharing"

[9] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. arXiv preprint arXiv:2006.03654, **2020.**

[10] Sebastian Hofstätter, Sheng-Chieh Lin, Jheng-Hong Yang, Jimmy Lin, and Allan Hanbury. 2021. Efficiently teaching an effective dense retriever with balanced topic aware sampling. In Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp.**113–122, 2021.**

[11] Hiroshi Inoue. 2019. Multi-sample dropout for accelerated training and better generalization. arXiv preprint arXiv:1905.09788, **2019.**

[12] Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with gpus. IEEE Transactions on Big Data 7, 3, pp.**535–547, 2019.**

[13] Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for opendomain question answering. arXiv preprint arXiv:2004.04906, **2020.**

[14] Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. Advances in Neural Information Processing Systems 32, pp.**7059–7069, 2019.**

[15] Dong-Hyun Lee et al. 2013. Pseudo-label: The simple and efficient semisupervised learning method for deep neural networks. In Workshop on challenges in representation learning, ICML, Vol.**3**. 896, **2013.**

[16] Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. 2021. Pyserini: A Python Toolkit for Reproducible Information Retrieval Research with Sparse and Dense Representations. In Proceedings of the 44th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2021). pp.**2356–2362, 2021.**

[17] Sheng-Chieh Lin, Jheng-Hong Yang, and Jimmy Lin. 2021. In-batch negatives for knowledge distillation with tightly-coupled teachers for dense retrieval. In Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP2021). pp.**163–173, 2021.**

[18] Benjamin Marlin, Richard S Zemel, Sam Roweis, and Malcolm Slaney. 2012.Collaborative filtering and the missing at random assumption. arXiv preprint arXiv:1206.5267, **2012.**

[19] S. Rangineni and D. Marupaka, "Data Mining Techniques Appropriate for the Evaluation of Procedure Information," *International Journal of Management, IT & Engineering*, Vol.**13**, No.**9**, pp.**12–25, 2023.**

[20] S. Rangineni, "An Analysis of Data Quality Requirements for Machine Learning Development Pipelines Frameworks," *International Journal of Computer Trends and Technology*, Vol.**71**, No.**9**, pp.**16–27, 2023.**

[21] Arvind Kumar Bhardwaj, Sandeep Rangineni, Divya Marupaka, "Assessment of Technical Information Quality using Machine Learning ," International Journal of Computer Trends and Technology, Vol.**71**, No.**9**, pp.**33-40, 2023.**

AUTHORS PROFILE

Sandeep Rangineni is a Data Test Engineer at Pluto TV, with over 12 plus years of experience in the IT industry, primarily within the streaming media industry. He holds a Master's degree in Engineering Management and Master's degree in Information Technology. Sandeep has a diverse skill set, working with technologies such as PL/SQL, Azure Databricks, Salesforce, Informatica, and Snowflake. Currently, he is actively engaged in researching Data Engineering and Data Quality topics. Sandeep has professional certifications in Salesforce admin, AWS Data Analytics and Safe 5 practitioner. Sandeep is a senior member of IEEE, professional member of BCS and fellow of IETE, three esteemed technology organizations, and has served as a judge for reputable award organizations in Technology which including Globee Awards, Stevie Awards, NCWIT Aspirations, and Brandon Hall Group.

Amit Bhanushali is a highly accomplished software quality assurance professional with over 22 years of experience in the IT industry. He earned his Master's in Business Data Analytics from West Virginia University in 2017. Based in West Virginia, USA, Mr. Bhanushali is a Senior IEEE Member and has significantly contributed to software testing research and practice.

His expertise spans automation testing, performance testing, DevOps, and CI/CD implementation. He has also led testing efforts in complex cloud environments. In addition to testing, Mr. Bhanushali has authored several articles exploring cutting-edge topics like artificial intelligence and machine learning. His published research demonstrates his thought leadership and impact on software quality engineering.

Mr. Bhanushali's accomplishments have been recognized through prestigious appointments. He serves as a reviewer for the Elsevier journal and has been a hackathon judge. His contributions were further honored in 2023 when he received the International Achievers' Award. With his sustained record of excellence across software development, testing, and research, Mr. Bhanushali continues to be an influential leader in his field.

Divya Marupaka is a Senior Software Data Engineer at Unikon IT Inc. She holds a Master's degree in Computer Science Engineering (US) and Bachelors in Electronics and communication Engineering (India) and has over 12+ years of experience in designing and developing scalable, multi-tiered, distributed software applications for enterprises in Insurance,Financial,Banking and Retail domains.

She is a highly qualified and skilled individual who has used her expertise in data engineering and data analytics. And is also a senior member of IEEE, fellow of IETE and

© 2023, IJCSE All Rights Reserved

professional member BCS, most esteemed technology organizations, and has served as a judge for reputable award organizations in Technology which include Globee Awards, NCWIT Aspirations, and Brandon Hall Group. She is also an Approved active mentor in the ADPlist organization who has coached many professionals belonging to science and technology. Her article was also published in IEEE Journal which is one of the world's largest online communities and leading publisher of knowledge resources for software engineering professionals.
She has designed and optimized data models on AWS Cloud using AWS data stores such as Redshift, RDS, S3, Glue Data Catalog, Python by participating in data analysis/design activities and conducting appropriate technical data design reviews at various stages during the development life cycle.

Srinivas Venkata completed his Master of Science in Engineering Management in the United States. Currently, he holds the position of Staff Data Engineer at Teradata Inc. in Texas, a role he has been dedicated to since 2022. Furthermore, he boasts the distinction of being a Fellow member of the IETE. His academic and professional achievements shine through his publications in esteemed international journals, notably Springer. His primary research areas encompass Data Engineering, Artificial Intelligence, Big Data Analytics, Data Mining, and Business Intelligence in the context of education. With an impressive 11 years of experience in the field of Information Technology, he brings a wealth of knowledge and expertise to his role.

Manoj Suryadevara earned his B.Tech. in Information Technology from Karunya University in 2011 and his Master of Science in Software Engineering from the University of Houston in 2014. He has over ten years of experience in product management, working with various tech companies to bring innovative solutions to market. Since 2020, he has been with Walmart, a leading retail firm in the US, where he has progressed through roles of increasing responsibility. As Staff Product Manager, he leads cross-functional teams to develop and implement product strategies, ensuring alignment with customer needs and market trends. He is well-versed in agile methodologies, user experience design, and data analytics. He has published two papers on data and is a senior member of the IEEE. His expertise includes digital transformation, data management, data science, cloud computing, and machine learning. He is passionate about leveraging technology to solve complex business problems and deliver customer value. Under his leadership, his teams have successfully launched numerous products, resulting in significant revenue growth and market share expansion.

© 2023, IJCSE All Rights Reserved