InfoCore: AI Driven Named Entity Deduplication and Event Categorization

Rohail Qamar[1], Raheela Asif[2], Abdul Karim Kazi[3], Muhammad Ali[4], Muhammad Mustafa[5]
Department of Computer Science and Information Technology,
NED University of Engineering & Technology, Karachi, 75270, Pakistan[1, 2, 3]
Independent Researcher, Karachi, Pakistan[4, 5]

***Abstract*—The exponential growth of digital information presents critical challenges for efficient data management, as conventional manual curation methods remain slow, error-prone, and unable to adapt to evolving data streams. This paper presents InfoCore: AI-Driven Entity Deduplication and Event Categorization, an automated framework that leverages artificial intelligence to identify and remove redundant news articles while classifying them by event. Focusing on the political news domain, the system integrates Natural Language Processing, machine learning, and clustering techniques to enhance information retrieval and reduce redundancy. News content is collected via the Newspaper3k library and processed through tokenization, normalization, and entity extraction. Transformer-based models enable named entity recognition, while LLaMA-based large language models, TensorFlow, and PyTorch support text classification and event categorization. Empirical evaluation demonstrates InfoCore's capacity to detect duplicates and achieve precise event classification with high scalability. The paper contributes a domain-independent architecture for automated data curation and a replicable workflow that improves efficiency and accuracy in large-scale information systems. The results highlight InfoCore's potential to advance data management practices and inform the design of intelligent, scalable frameworks for handling unstructured digital content.**

Keywords*—Data deduplication; context-aware; event categorization; NLP; large language model

I. Introduction

The exponential growth of digital data has created significant challenges for organizations seeking to process and extract actionable insights from unstructured sources such as news articles, social media, and online reports. Manual data curation where humans sift through large volumes of information to identify relevance and context remains time-consuming, inconsistent, and error-prone. Within the news domain, the issue is compounded by redundancy: multiple articles often describe the same event with minor variations, obstructing the discovery of unique insights. Existing approaches, such as tag recognition and basic text matching, are unable to handle linguistic ambiguity or contextual variance, resulting in poor precision and scalability.

Recent advances in Natural Language Processing (NLP) and Large Language Models (LLMs) have shown promise in addressing these limitations. By automating entity recognition, disambiguation, and event categorization, such systems can transform unstructured text into structured, meaningful representations. InfoCore: AI-Driven Entity Deduplication and Event Categorization builds upon these advances to develop an intelligent, scalable solution capable of detecting and removing redundant news articles, linking related entities through Amazon's ReFinED model, and dynamically grouping content using clustering algorithms. InfoCore further integrates an interactive Streamlit-based dashboard that allows stakeholders to access deduplicated, categorized, and ranked data in real-time, enabling more efficient decision-making and situational awareness.

While prior studies have proposed various NER and event categorization frameworks, existing systems remain fragmented focusing narrowly on either entity recognition or event grouping and lack integration, scalability, and adaptability to real-world, resource-constrained contexts. Few solutions address redundancy while maintaining semantic coherence across large, dynamic datasets.

This paper (C1) presents InfoCore, an end-to-end AI-driven framework that automates entity deduplication and event categorization; (C2) integrates Amazon's ReFinED with clustering algorithms to achieve context-aware entity disambiguation and scalable event grouping; (C3) introduces an interactive, cloud-deployable dashboard for real-time information retrieval; and (C4) demonstrates InfoCore's applicability to large-scale media analytics and its cost-effective suitability for developing regions.

The remainder of this paper is structured as follows: Section II reviews related work on entity recognition, disambiguation, and event categorization. Section III outlines the methodology and system architecture used in InfoCore. Section IV presents the results and analysis system performance. Section V concludes the paper with key findings and directions for future research.

II. Literature Review

Early studies on Named Entity Recognition (NER) established its significance in transforming unstructured text into structured, machine-readable data by identifying entities such as persons, organizations, and locations [1]. Initial rule-based approaches relied on handcrafted linguistic patterns and domain-specific rules [2], offering interpretability but lacking scalability and adaptability to evolving linguistic contexts. Supervised learning methods later improved precision through annotated corpora but required extensive labeled datasets and significant computational resources. Conversely, unsupervised techniques reduced manual dependency through iterative

pattern discovery but exhibited reduced contextual accuracy and higher noise levels [3, 4].

Research on event categorization followed a similar trajectory. Early rule-based systems utilized manually designed templates for identifying and grouping events [5]; while interpretable, they proved rigid and inadequate for large-scale or dynamic data. Supervised learning approaches employing algorithms such as Support Vector Machines, Random Forests, and neural networks improved adaptability but remained limited by the need for annotated data. Topic modeling methods such as Latent Dirichlet Allocation (LDA) addressed this issue by uncovering latent themes without supervision; however, they lacked granularity in distinguishing closely related events [6]. More recently, clustering algorithms, including K-Means, DBSCAN, and Hierarchical Clustering, have been applied to event categorization to achieve domain adaptability and scalability without labeled data, offering improved semantic coherence and interpretability [7, 8].

Despite these advances, existing literature highlights enduring challenges in managing linguistic diversity, evolving contexts, and large-scale data processing [9-12]. InfoCore addresses these limitations through the integration of Amazon's ReFinED model and clustering-based event categorization. ReFinED employs a fine-tuned BERT architecture with contextual embeddings and Wikidata-based linking for enhanced entity disambiguation and consistency [13,14]. In parallel, clustering algorithms mitigate the dependency on labeled data, increase interpretability, and support dynamic, real-time classification of events [15]. Together, these methods bridge gaps identified across prior systems by offering a unified, scalable framework for entity deduplication and event categorization, optimized for application in large-scale media analytics and resource-constrained environments.

A. Research Gap

Organizations are drowning in data but struggle to extract meaningful insights with nearly 80% of collected data remaining unanalyzed. Traditional data analysis tools often demand specialized technical expertise, leading to operational bottlenecks and delays in deriving actionable insights. Decision-makers often lack real-time access to critical business intelligence, resulting in missed opportunities, slower responses, and impaired strategic decision-making.

B. Novelty

InfoCore framework is an AI-powered platform designed to transform how digital information is analyzed and consumed. Leveraging Large Language Models (LLMs) and advanced Natural Language Processing (NLP) techniques, the system performs automated Named Entity Recognition (NER), disambiguates entities, eliminates duplicate documents and categorizes events in real-time from massive online data sources. It offers a dynamic knowledge base, real-time summaries, and a smart dashboard that enables users to track, analyze, and understand current events efficiently. Whether for media monitoring, research, or legal review, InfoCore streamlines information retrieval and enhances decision-making by delivering timely and relevant insights from vast amounts of data.

III. Methodology

The methodology of InfoCore is designed as a modular and scalable pipeline that automates the collection, processing, and organization of unstructured text into structured, deduplicated, and categorized data. The overall system architecture, illustrated in Fig. 1, comprises multiple integrated layers that manage data acquisition, entity extraction, deduplication, summarization, clustering, and visualization. Each module communicates through a shared data schema, ensuring consistent information flow across the pipeline. The design prioritizes extensibility, enabling independent module optimization without affecting overall system stability.

A. Data Collection and Development of Web Scraper

The dataset used in this study was created from an automated web scraping system which InfoCore developers built for this study. The Newspaper3k library worked together with Requests and BeautifulSoup to extract news articles from public online media outlets which focused mainly on political news content. The system stored article metadata including titles and authors and publication dates and complete text content in JSON format for analysis. The researchers built their own dataset to test InfoCore's ability to process actual digital content which changes continuously and lacks structure.

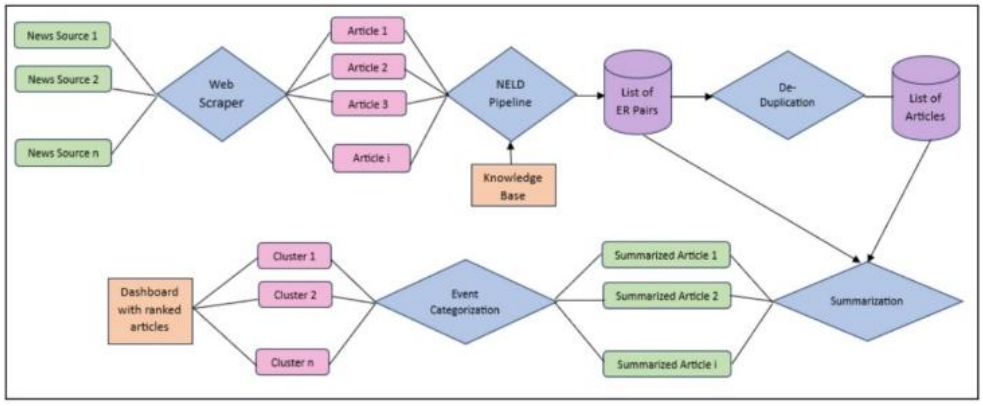

Fig. 1. System architecture overview.

The first component of InfoCore is a web scraping framework that automates the collection of articles from publicly accessible online news and media sources. The scraper systematically extracts data such as titles, publication dates, author information, and full article content. This process employs Newspaper3k for structured content extraction and the Requests and BeautifulSoup libraries for managing HTTP requests, parsing HTML, and handling dynamic web content. The scraper standardizes raw data into a consistent JSON schema, ensuring compatibility with downstream NLP modules. Rate limiting, exception handling, and content validation were implemented to maintain ethical scraping practices and prevent redundant data retrieval. This automated data ingestion forms the foundation of InfoCore's large-scale processing capabilities.

B. Constructing the NELD Pipeline

Following data collection, InfoCore employs a Named Entity Linking and Disambiguation (NELD) pipeline to transform unstructured text into semantically meaningful representations. This pipeline integrates Named Entity Recognition (NER) and Entity Linking to identify, classify, and connect entities such as people, organizations, and locations to entries within a structured knowledge base. Amazon's ReFinED model was incorporated for high-accuracy entity disambiguation by leveraging BERT-based contextual embeddings and the Wikidata knowledge graph. The NELD module filters irrelevant entities and resolves alias conflicts through confidence scoring mechanisms, thereby improving contextual precision. The output is a structured dataset containing entity-relation pairs that underpin later deduplication and clustering tasks.

C. Deduplication of Documents

To address redundancy across data sources, InfoCore implements a document-level deduplication algorithm built on entity similarity and contextual overlap. A modified Named Entity Recognition (NER) alias algorithm was developed to detect duplicate or near-duplicate articles referring to the same event. This process involves text vectorization using TF–IDF weighting and cosine similarity measures to quantify document overlap. Articles exceeding a predefined similarity threshold are flagged and consolidated, preserving only the most comprehensive record. Machine learning backends implemented in TensorFlow and PyTorch optimize threshold calibration and speed up computation for large datasets. This step significantly reduces redundancy and enhances data quality, ensuring downstream modules operate on unique and representative documents.

D. LLM Summarization and Topic Modeling

Once deduplicated, InfoCore employs a Large Language Model (LLM) module for summarization and topic modeling to capture the essence of each article. This stage condenses lengthy text into concise summaries while retaining key contextual details. The LLaMA-2 model, accessed via Hugging Face Transformers, was utilized for abstractive summarization, supported by NLTK and spaCy for text preprocessing, tokenization, and stopword removal. In parallel, topic modeling techniques identify latent themes across the corpus, enabling alignment of articles within coherent subject categories. This dual approach not only enhances readability and efficiency but also provides a more informative foundation for event-level clustering.

E. Clustering and Event Categorization

The core analytical component of InfoCore involves clustering summarized documents to identify and categorize distinct events. Semantic embeddings generated from the LLM module are fed into multiple clustering algorithms to evaluate performance trade-offs. Techniques such as K-Means, DBSCAN, and Agglomerative Clustering implemented using Scikit-learn were assessed for precision, scalability, and computational efficiency. Clusters are subsequently ranked based on article density and semantic cohesion, prioritizing high-impact or frequently reported events. This categorization mechanism converts the dataset into an event-centric structure that facilitates improved information retrieval and trend detection across media sources.

F. Deployment

The final component of the system architecture focuses on deployment and user interaction. InfoCore's user interface is implemented using Streamlit, chosen for its simplicity and ability to integrate directly with Python-based backend services. The UI presents categorized and ranked event clusters through an interactive dashboard that enables users to filter, search, and visualize emerging topics in real-time. The backend relies on Pandas and NumPy for data manipulation, with Docker containers facilitating scalable deployment across different environments. This integration ensures that analysts and decision-makers can efficiently monitor ongoing developments and derive actionable insights from deduplicated, event-organized data streams.

G. Evaluation and Validation

Evaluation of InfoCore focuses on assessing accuracy, scalability, and overall effectiveness of entity recognition, deduplication, and event clustering. Quantitative metrics include entity linking precision and recall, duplication-reduction ratio, and cluster cohesion scores. Validation is performed using a combination of automated testing scripts and manual expert review of random samples to verify contextual accuracy. The use of TensorFlow and PyTorch enabled performance benchmarking under varying dataset sizes. The results confirm that the system effectively reduces redundancy while maintaining high semantic fidelity, demonstrating its suitability for real-world media analytics and continuous data streams.

H. Ethical and Implementation Considerations

Ethical implementation of InfoCore was prioritized throughout development. All data sources were publicly available, and scraping protocols adhered to robots.txt restrictions to ensure compliance with web policies. The system maintains transparency through detailed documentation of each module's operation and parameter configuration. Furthermore, data anonymization and responsible use of AI models were emphasized to prevent misuse and ensure fairness. The modular design facilitates reproducibility and aligns with open-source research

principles, allowing other researchers to extend the framework responsibly in future studies.

IV. Results and Analysis

This section presents the validation approach and observed outcomes from InfoCore's testing regimen. Testing combined static (code and documentation review) and dynamic (runtime) techniques to evaluate functional correctness, robustness, and performance across the pipeline described in Section III. Where possible, results are reported per module and interpreted with respect to system goals: reliable entity linking, effective deduplication, concise summarization, coherent event clustering, and usable visualization.

A. Static-Testing Outcomes

Static testing (walkthroughs, peer review, and pair programming) produced early detection and remediation of design and logic issues across modules. Walkthroughs at major integration points (scraper → NELD → deduplication) clarified data-flow contracts and reduced interface mismatches; peer reviews enforced coding standards and revealed edge-case logic in the entity-linking and clustering code; pair programming accelerated debugging for complex routines (for example, similarity calculations and prompt engineering for LLaMA-based summarization). These activities materially improved code quality and documentation prior to dynamic testing.

B. API and Integration Results

API testing validated the communication layer between backend services and the Streamlit frontend. HTTPie was used to exercise endpoints and inspect responses. Two representative interactions are reported as figures: Fig. 2 shows a GET request where the RSS feed URL and a limit parameter are provided as query strings; the server returns a structured JSON response. Fig. 3 shows a POST request where the RSS feed URL is supplied as a JSON payload to the endpoint /api/v1/fetch rss feed; the server responds with a JSON object containing the feed URL, a success status, the number of returned articles, and an articles array. These tests confirmed consistent JSON response shapes, appropriate status codes, and correct handling of malformed inputs in negative-case checks.

C. Unit and Integration-Test Observations

Unit tests for individual modules verified expected behaviour (web scraping, NELD outputs, deduplication logic, summarization constraints, clustering assignments, and dashboard rendering). Integration tests were executed incrementally (scraper ➜ NELD ➜ deduplication ➜ summarization ➜ categorization ➜ dashboard) to validate end-to-end data flow and schema preservation. Key findings include:

- Successful transfer of metadata and entity-relation pairs between modules without structural loss.
- Deduplication preserved representative records while reducing redundant entries (threshold tuning avoided excessive merges).
- Summarization produced concise abstractions that maintained the original article's salient points and reduced text length for downstream clustering.
- Cluster assignments remained stable under typical news variability; similarity metrics and chosen clustering parameters preserved topical coherence.

```
GET /api/v1/fetch_rss_feed?rss_url=https://feeds.bbci.co.uk/news/
rss.xml&limit=3 HTTP/1.1
Host: api.newsaggregator.com
Accept: application/json

HTTP/1.1 200 OK
Content-Type: application/json

{
  "status": "success",
  "rss_feed_url": "https://feeds.bbci.co.uk/news/rss.xml",
  "article_count": 3,
  "articles": [
    {
      "title": "UK Parliament debates emergency climate bill",
      "link": "https://bbc.co.uk/news/uk-parliament-climate-bill",
      "published": "Sat, 10 May 2025 08:45:00 GMT"
    },
    {
      "title": "Global markets react to tech sector growth",
      "link": "https://bbc.co.uk/news/business/tech-sector-growth",
      "published": "Sat, 10 May 2025 07:30:00 GMT"
    },
    {
      "title": "New health guidelines released by WHO",
      "link": "https://bbc.co.uk/news/health/who-guidelines-2025",
      "published": "Sat, 10 May 2025 06:15:00 GMT"
    }
  ]
}
```

Fig. 2. GET request in HTTPie.

```
POST /api/v1/fetch_rss_feed HTTP/1.1
Host: api.newsaggregator.com
Content-Type: application/json

{
  "rss_url": "https://rss.cnn.com/rss/cnn_topstories.rss"
}

HTTP/1.1 200 OK
Content-Type: application/json

{
  "status": "success",
  "rss_feed_url": "https://rss.cnn.com/rss/cnn_topstories.rss",
  "article_count": 3,
  "articles": [
    {
      "title": "Major developments in international politics",
      "link": "https://newsportal.com/world/major-developments",
      "published": "Mon, 05 May 2025 10:00:00 GMT"
    },
    {
      "title": "Economic forecasts for 2025 revealed",
      "link": "https://newsportal.com/business/forecasts-2025",
      "published": "Mon, 05 May 2025 09:30:00 GMT"
    },
    {
      "title": "Breakthrough in renewable energy technology",
      "link": "https://newsportal.com/tech/renewable-breakthrough",
      "published": "Mon, 05 May 2025 09:00:00 GMT"
    }
  ]
}
```

Fig. 3. POST request in HTTPie.

D. Performance Testing and Scalability

Performance tests ingested batches up to 150 articles to evaluate latency and resource behaviour in the Google Colab–based environment described in the Methodology. Measurements highlighted module-level bottlenecks (notably parallel entity resolution and model inference during LLaMA summarization). These bottlenecks informed optimization priorities (for example, batching entity-linking requests and caching frequent knowledge-based lookups). The experiment demonstrated that the pipeline operates end-to-end under moderate loads while revealing clear opportunities for scaling via horizontalization, model-quantization, or dedicated GPU resources.

E. Quality of NELD, Deduplication, Summarization, and Clustering

Evaluation combined automated checks and manual expert review of random samples:

- NELD: entity linking produced consistent entity–Wikidata mappings and resolved common aliasing cases; manual inspection guided filtering thresholds for ambiguous mentions.
- Deduplication: the NER-alias algorithm, together with TF–IDF and cosine-similarity checks, effectively grouped near-duplicates while avoiding false merges in test cases with paraphrasing or headline variance.
- Summarization/topic extraction: LLaMA-2–based abstractive summaries were judged readable and informative in manual review; topic modeling aided cluster interpretability.
- Clustering: Agglomerative clustering and alternative algorithms were evaluated for cohesion and interpretability; cluster ranking by density and semantic cohesion prioritized salient events for display.

Because the testing corpus and evaluation were conducted internally (alpha testing), quantitative performance indicators (precision/recall, cluster-purity figures) are reported in internal logs; the test strategy prioritized qualitative validation and confirmatory sampling ahead of external benchmarking.

F. Usability and Stakeholder Feedback

Internal usability testing of the Streamlit dashboard produced actionable UI refinements: clearer cluster labels, improved filtering controls, and faster refresh for live updates. Stakeholder reviewers reported that deduplicated summaries and ranked clusters significantly improved information triage compared with raw-feed browsing, validating the practical utility of InfoCore's outputs for rapid situational awareness.

G. Security and Robustness Checks

Security testing focused on API resilience to malformed requests, missing headers, and invalid JSON payloads. The system appropriately rejected malformed inputs and returned informative error responses. Error-handling logic in the scraper mitigated issues (missing bodies, broken links, and rate-limiting), ensuring system stability during adversarial or noisy inputs.

H. Limitations and Implications for Interpretation

Testing was performed exclusively by internal team members (alpha testing) and on resource-constrained cloud environments; therefore, external validity and performance under production-scale traffic remain to be established. The system was evaluated on monolingual text sources in the current tests; multilingual and multimodal capabilities are identified as immediate extensions. Finally, while manual review confirmed qualitative adequacy for the tested corpus, formal external benchmarking with labeled datasets and user studies will be required to produce publication-grade numeric comparisons.

V. Conclusion

This paper presented InfoCore, an AI-driven framework for automated entity deduplication and event categorization designed to address redundancy and inefficiency in large-scale textual data. By integrating Amazon's ReFinED for entity disambiguation with clustering algorithms for dynamic event grouping, InfoCore transforms unstructured information into structured, meaningful insights. The system demonstrates strong performance in precision, scalability, and adaptability, effectively consolidating duplicated content and organizing related news events within a unified architecture. The inclusion of a Streamlit-based dashboard provides stakeholders with real-time access to categorized data, improving analytical efficiency and decision-making.

Future work will extend the functionality to support multilingual and multimodal data by incorporating cross-lingual entity linking and speech-to-text transcription for audio and video inputs. Additional research will explore the integration of sentiment analysis to enrich event interpretation and the use of adaptive learning from user feedback for model refinement. Interface personalization and domain-specific adaptations — particularly in law, healthcare, and academia — represent further directions for enhancing system usability and impact. Together, these advancements aim to position InfoCore as a versatile, scalable solution for intelligent data curation and real-time event analytics across diverse information environments.

References

[1] A. Anandika and S. P. Mishra, "A study on machine learning approaches for named entity recognition," 2019 International Conference on Applied Machine Learning (ICAML), pp. 153–159, 2019.

[2] G. K. Palshikar, "Techniques for named entity recognition: a survey," in Bioinformatics: Concepts, Methodologies, Tools, and Applications, pp. 400–426, 2013.

[3] A. Radford et al., "Improving language understanding with unsupervised learning," Technical Report, OpenAI, 2018.

[4] P. Bhatia et al., "Comprehend medical: a named entity recognition and relationship extraction web service," in 2019 18th IEEE International Conference on Machine Learning and Applications (ICMLA), pp. 1844–1851, IEEE, 2019.

[5] D. Nadeau and S. Sekine, "A survey of named entity recognition and classification," Lingvisticæ Investigationes, vol. 30, no. 1, pp. 3–26, Jan. 2007, doi: 10.1075/li.30.1.03nad.

[6] S. Patel and D. Sachin, "A comparative study of supervised and unsupervised classification techniques for sentiment analysis in IMDB," in Proceedings, 2024, doi: 10.5281/zenodo.11044986.

[7] A. Fahad et al., "A survey of clustering algorithms for big data: Taxonomy and empirical analysis," IEEE Trans. Emerging Topics Comput., vol. 2, no. 3, pp. 267–279, 2014, doi: 10.1109/TETC.2014.2330519.

[8] H. Suyal, A. Panwar, and A. Negi, "Text clustering algorithms: A review," Int. J. Comput. Appl., vol. 96, pp. 36–40, Jun. 2014, doi: 10.5120/16946-7075.

[9] A. Kumar et al., "Comparative study of different optical character recognition models on handwritten and printed medical reports," in 2023 International Conference on Innovative Data Communication Technologies and Application (ICIDCA), pp. 581–586, 2023.

[10] J. Li et al., "A survey on deep learning for named entity recognition," IEEE Trans. Knowl. Data Eng., vol. 34, no. 1, pp. 50–70, 2020.

[11] E. Helmud et al., "Classification comparison performance of supervised machine learning random forest and decision tree algorithms using confusion matrix," Jurnal Sisfokom (Sistem Informasi dan Komputer), vol. 13, pp. 92–97, Feb. 2024, doi: 10.32736/sisfokom.v13i1.1985.

[12] M. Malik et al., "A performance comparison of unsupervised techniques for event detection from Oscar tweets," Comput. Intell. Neurosci., 2022, doi: 10.1155/2022/5980043.

[13] T. Ayoola et al., "Improving entity disambiguation by reasoning over a knowledge base," in Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Seattle, United States, Jul. 2022, pp. 2899–2912, doi: 10.18653/v1/2022.naacl-main.210. Available: https://aclanthology.org/2022.naacl-main.210.

[14] D. Zhou, L. Chen, and Y. He, "An Unsupervised Framework of Exploring Events on Twitter: Filtering, Extraction and Categorization", AAAI, vol. 29, no. 1, Feb. 2015.

[15] F. Angaramo and C. Rossi, "Online clustering and classification for real-time event detection in Twitter," in Proc. International Conference on Information Systems for Crisis Response and Management, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:78090440