Svedberg Open
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

A Hybrid Nlp And Big Data Analytics Framework For Privacy-Aware Information Extraction In Intelligent Transportation Systems

Govindasamy. R.[1]* , Shanmugapriya. N[2] , Gopi. R.[3]

[1]Department of Computer Science & Engineering, School of Engineering & Technology, Dhanalakshmi Srinivasan University, Trichy - 621112, Tamil Nadu, India
[2]Department of Computer Science & Engineering, School of Engineering & Technology, Dhanalakshmi Srinivasan University, Trichy - 621112, Tamil Nadu, India
[3]Department of Computer Science & Engineering, Dhanalakshmi Srinivasan Engineering College, Perambalur - 621212, Tamil Nadu, India
*Corresponding & First Author
E-mail/Orcid Id:
ggnavaneeth@gmail.com, https://orcid.org/0009-0009-3429-5648 ;
shanmugapriyan.set@dsuniversity.ac.in, https://orcid.org/0000-0002-4620-1807 ;
gopircse@gmail.com, https://orcid.org/0000-0003-4957-1843

Abstract
Intelligent Transportation Systems (ITS) now can produce a continuous mixture of structured, semi-structured, and unstructured data from vehicles, road-side units, GPS devices, smart cards, traffic controllers, ticketing platforms, passenger feedback portals, and control-room messages. The analytics of these available data is clear, but the direct storage of raw transport records has two practical problems that are underestimated. It increases the processing cost, and it keeps personally identifying details inside the analytical pipeline longer than necessary. This paper proposes a Hybrid Natural Language Processing and Big Data Analytics Framework for Privacy-Aware Information Extraction in ITS. The framework first converts the heterogeneous transport text into normalized tokens, named entities, and event records using rule assisted NLP, gazetteer matching, domain patterns, and event clustering. It then masks or generalizes the sensitive identifiers before data is retained for downstream analytics. Instead of treating raw messages as the main analytical object, the framework stores compact event level records containing incident type, route, time, generalized location, and anonymized operational identifiers. A transparent simulation using a synthetic ITS message corpus was conducted to examine the extraction accuracy, storage reduction, processing time, and privacy masking coverage. Results show that the proposed pipeline improves macro F1-score over a simple keyword baseline, it reduces retained storage in the synthetic setting, and it masks all explicitly generated passenger and vehicle identifiers. The obtained results should not be read as field deployment evidence; rather, they indicate that privacy-aware extraction before long-term storage is a plausible design for the direction of smart transportation analytics. This paper also discusses its limitations relating to city specific language, multilingual data, model drift, and the privacy-utility trade-off that appears whenever exact mobility traces are generalized.

Keywords: Intelligent Transportation Systems; natural language processing; big data analytics; privacy preservation; information extraction; named entity recognition; traffic incident detection.

1. Introduction

Smart transportation platforms are not only built for traffic counters and cameras. In many cities, the control room receives a less tidy mixture of short messages from drivers, GPS pings from buses, smart-card entries, passenger complaints, incident reports, traffic-signal logs, and sometimes public social-media posts about roads that are blocked or stops that are unsafe. A typical message may say, for example, "Route 21G delayed near Guindy Junction; bus number TN09AB1234; card PAX-10492 reported missed connection at 08:15." For operations staff, this message is very useful because it describes a delay, route, place, vehicle, user, and time. For a privacy officer, the same message is uncomfortable because it also contains a vehicle trace and also a passenger-level identifier.

The older approach of storing first and asking questions later is becoming harder to defend. Big-data platforms can go through and consume large traffic archives, and the transportation literature has shown the value of such data for congestion estimation, signal control, fleet monitoring, road safety, and demand analysis (Zhang et al., 2011; Zhu et al., 2019). Still the raw storage brings a hidden cost. It preserves names, card numbers, exact coordinates, and repeated movement patterns even when the analytical task only needs the event type, route, approximate zone, and time window. The same issue appears in connected-vehicle and Internet-of-Vehicles environments, where frequent data exchange can improve traffic efficiency but it may expose sensitive mobility traces (Darwish and Abu Bakar, 2018; Arooj et al., 2022).

Privacy-preserving analytics has several mature ideas that includes k-anonymity, pseudonymization, perturbation, and differential privacy (Samarati, 2001; Sweeney, 2002; Dwork, 2006; Tran et al., 2019). Still, these methods are often discussed after data that have already been collected and organized. In ITS text streams, an earlier intervention may be more useful but in extract what the transportation system needs, remove or transform what it does not need, and only then send the reduced record to the analytics layer. This may sounds simple, but it is not trivial. Transport text is messy, abbreviated, multilingual in many deployments, and contains many local expressions. A driver might write "accdnt nr OMR toll" rather than "accident near OMR Toll Plaza". A passenger may describe the same event as "bus stuck before toll for 30 min". Conventional keyword search might miss such variation which is important.

Natural Language Processing (NLP) offers a route for this problem. Named Entity Recognition (NER), rule-based pattern matching, relation extraction, and event clustering can convert transport messages into structured records. Prior works has used text mining for crash narratives, traffic accident detection, transport infrastructure studies, and event extraction from social-media streams (Suat-Rojas et al., 2022; Chen et al., 2022; Chowdhury et al., 2023). A fair critique, however, is that many pipelines focus primarily on detection accuracy while privacy issue is left as a separate post-processing step. This separation may be acceptable for laboratory datasets but when it comes to real time it is less convincing for operational ITS deployments.

This paper proposes a hybrid framework in which extraction and privacy protection are joined inside the same data flow. The paper provides an implementable system design which is a layered pipeline that detects transport events from heterogeneous text, masks or generalizes sensitive entities before retention, and stores the compact event records for scalable analytics. In many ITS applications, full raw text is not needed indefinitely. Event-level records may provide enough analytical utility while reducing the privacy issues.

The main contributions of this paper can be summarized in five parts. First, we develop a privacy-aware NLP pipeline specifically for Intelligent Transportation System (ITS) messages, combining domain-specific text normalization, named entity recognition (NER), pattern-based extraction, and event classification. Rather than treating every message as data that should be retained, the proposed approach takes a more restrictive view of storage. Raw messages are considered short-lived processing inputs and are discarded once the required traffic-related information has been extracted and privacy-sensitive content has been handled. This forms the second contribution: a data-minimization strategy intended to reduce unnecessary retention of potentially identifiable information.

The third contribution is a layered big-data architecture that links these privacy-preserved event records to downstream ITS applications, including traffic-state monitoring, route-delay analysis, incident clustering, and operational decision support. The architecture is designed around extracted events rather than complete textual messages, which may reduce both privacy exposure and storage overhead. Fourth, we evaluate the proposed approach through a transparent simulation using a synthetic corpus of ITS messages. The experiments examine measurable aspects of the pipeline, including information-extraction accuracy, masking coverage, and the reduction in storage achieved when raw messages are replaced by structured event records. These results are not intended to establish universal performance; instead, they provide an initial indication of how the proposed framework may behave under controlled ITS messaging scenarios.

Finally, the paper examines several limitations that could affect practical deployment. Performance is likely to become less predictable when messages contain multiple languages, informal or evolving transportation terminology, previously unseen event descriptions, or locations that are poorly represented in available gazetteers. Privacy mechanisms introduce another trade-off: aggressive anonymization may protect users more effectively, but it can also remove contextual information that is useful for traffic analysis. Recognizing these limitations is important because the usefulness of a privacy-aware ITS pipeline depends not only on how accurately it extracts information, but also on how well it balances analytical utility with the amount of personal information retained.

2. Related Work

2.1 Big data analytics in ITS

Data-driven ITS research has long argued that transportation management should use the large amount of data produced by vehicles, infrastructure, users, and control systems. Zhang et al. (2011) described data-driven ITS as a broader shift in which sensing and analytics become central to traffic operations. Zhu et al. (2019) later organized big-data analytics in ITS around data acquisition, storage, processing, and application layers. Their survey is relevant here because it frames ITS data as high-volume, high-velocity, and heterogeneous, rather than as a single clean database.

Several works extended this view to Internet-of-Vehicles and edge/fog computing environments. Darwish and Abu Bakar (2018) argued that the fog-based architectures may reduce latency and support distributed traffic analytics closer to the data source. Arooj et al. (2022) reviewed big-data processing in Internet-of-Vehicles and noted the unresolved challenges around heterogeneity, scalability, security, and privacy. These studies provide the system-level background for the present paper, but they usually do not focus on unstructured text as a first-class ITS data source.

2.2 Text mining and NLP for transportation data

Transportation text mining has attracted growing interest because useful mobility information is not always captured in numerical form. Crash narratives, police reports, maintenance records, public complaints, news feeds, and social-media posts often contain contextual details that conventional sensors may overlook or report only after some delay. Chowdhury et al. (2023), for example, reviewed text-mining applications in transportation infrastructure and found that textual data have been used in areas such as crash analysis, construction-disruption monitoring, infrastructure condition assessment, and transportation safety studies. Social-media data present a somewhat different challenge. Suat-Rojas et al. (2022) used social-network information to identify traffic accidents, showing that informal user-generated text can provide timely signals while also introducing ambiguity, inconsistent terminology, and noise. Chen et al. (2022) examined a more downstream application by combining text mining with ensemble learning to predict traffic-accident duration. Although the objective differs from direct event extraction, their work still indicates that informative textual features can contribute to transportation analytics.

Named entity recognition (NER) and event extraction become especially relevant when transportation messages contain references to locations, road segments, routes, times, vehicles, or incident categories. Early neural sequence-labeling approaches, including the architecture proposed by Lample et al. (2016), substantially influenced subsequent NER research. Later, transformer-based language models such as BERT shifted expectations for contextual representation and sequence-labeling performance across many NLP tasks (Devlin et al., 2019). Transportation-specific studies have built on these developments. Schiersch et al. (2020) introduced a corpus for fine-grained NER and relation extraction covering traffic- and industry-related events, while Yang et al. (2023) approached traffic-event detection as a slot-filling problem. Taken together, these studies appear to support the view that ITS messages can often be transformed into structured entities, attributes, and relationships instead of being retained solely as unstructured text.

The present work follows this general direction, but with a different emphasis. Privacy handling is incorporated directly into the extraction workflow rather than being treated as a separate post-processing operation. In other words, the objective is not simply to recognize traffic-related entities and events, but to determine which pieces of information are actually required for ITS analysis and which should be masked, generalized, or discarded. This distinction may be important in settings where mobility messages contain personally identifiable or quasi-identifying information alongside operationally useful traffic content.

2.3 Security and privacy in smart transportation

Security Security and privacy remain persistent concerns in Intelligent Transportation Systems because mobility data are closely tied to both physical movement and personal identity. Risks can arise through unauthorized access, data leakage, location tracking, identity disclosure, inference attacks, and even manipulation of transportation services or control functions. Hahn et al. (2021) classified a broad range of ITS security and privacy issues and emphasized that sensing, communication, mobility, and control are deeply interconnected. This interdependence makes privacy protection difficult to isolate as a purely data-management problem. Garg et al. (2019), for instance, examined security issues associated with edge-based big-data analytics in vehicular ad hoc networks, while Mohanta et al. (2021) discussed the role of blockchain in strengthening security and privacy within IoT environments. More recently, Govindasamy et al. (2024) proposed a secure data-encryption and control-based big-data framework for smart transportation management. Such approaches provide an important security-oriented foundation, although they mainly address how transportation data are protected during storage, communication, or processing. The present work approaches the problem somewhat differently. Its focus is on reducing privacy exposure earlier in the data lifecycle by extracting only analytically useful information from textual ITS messages and minimizing the amount of raw or identifiable content retained for later analysis.

Established privacy models also provide useful principles for handling transportation data. Under k-anonymity, a released record should be indistinguishable from at least (k-1) other records with respect to a defined set of quasi-identifiers (Samarati, 2001; Sweeney, 2002). Differential privacy offers a stronger and more formal privacy guarantee by limiting how much the inclusion or removal of a single individual's record can influence the result of a query (Dwork, 2006). These models are valuable, but applying them directly to operational ITS text is not always straightforward. A short incident message, for example, may contain a vehicle registration number, an exact street location, a person's name, and a precise timestamp within the same sentence. In such cases, relatively simple transformations such as masking direct identifiers, generalizing locations to road segments or zones, and reducing timestamp precision may be easier to implement and audit than a fully formal privacy mechanism at the message level.

The framework proposed in this paper adopts this more pragmatic approach for intermediate event records. Personally identifying fields are removed or masked, while selected quasi-identifiers can be generalized when fine-grained detail is not necessary for the intended traffic analysis. This choice does involve a trade-off. Excessive generalization may weaken applications such as incident localization or route-delay estimation, whereas retaining unnecessarily precise information increases privacy exposure. For that reason, the proposed pipeline treats data minimization as a task-dependent design decision rather than assuming that maximum anonymization is always preferable. Formal mechanisms such as differential privacy may still be applied at the aggregate reporting stage, particularly when statistics are released across populations or repeated analytical queries are exposed to external users.

2.4 Research gap

The reviewed literature appears to leave an important gap at the intersection of ITS text mining, big-data architecture, and privacy-aware data minimization. Big-data studies in transportation generally focus on large-scale storage, streaming, system integration, and analytics, while NLP research tends to emphasize entity recognition, event detection, and extraction accuracy. Privacy studies often address information protection after data have already been collected, structured, stored, or prepared for release. What seems less developed is a practical workflow that considers privacy during the extraction stage itself. In operational terms, the issue is not only whether useful information can be extracted from an ITS message, but also which elements should actually be retained. A message may contain a road name, precise timestamp, vehicle identifier, personal name, and congestion description, even though only some of these details may be necessary for traffic analysis.

The proposed framework addresses this gap by placing privacy transformation between NLP-based information extraction and long-term big-data storage. Raw messages are first processed to identify traffic-relevant entities and events, after which sensitive or unnecessarily precise information is masked, generalized, or discarded before the structured event record is retained. This design may reduce unnecessary privacy exposure while preserving the information needed for applications such as traffic monitoring, route-delay analysis, incident clustering, and operational decision support. In this way, data minimization becomes part of the analytical pipeline rather than a separate privacy step applied only after storage.

3. Problem Formulation and Design Objectives

Let D = {d1, d2, ..., dn} denote the stream of ITS records. Each record will be a traffic alert, complaint, driver note, AVL message, controller log, or semi-structured form entry. A record di contains two overlapping sets of information that includes operational information Oi, such as incident category, route, approximate location, and time; and sensitive information Si, such as passenger identifiers, exact vehicle registration numbers, smart-card IDs, phone numbers, or fine-grained location traces. The goal is not merely to classify di. The goal is to produce a reduced event record ei that preserves operational value while minimizing retained sensitive information.

The design problem can be stated as follows: extract E(di) = {event_type, route, time, zone, affected_asset, severity, confidence} while transforming S(di) through a privacy function P(.). The analytics layer should receive P(E(di)) rather than the complete raw record whenever full text retention is not legally or operationally necessary. This is a data-minimization objective, not an assumption that raw data are useless. Some raw records may be needed for audits, complaints, or investigations. Even then, they should be stored separately, with stricter access control and shorter retention.

Three objectives guide the proposed framework: extraction utility, privacy reduction, and computational efficiency. First, the system should identify relevant transportation events and entities with acceptable precision and recall. Second, direct identifiers and high-risk quasi-identifiers should be masked, generalized, or removed before the data enter analytical storage. Third, compact event records should reduce storage requirements and support faster downstream analysis than repeated processing of raw text. These objectives may not always align. Exact GPS coordinates, for instance, can improve route-level analysis and incident localization, but they also increase the risk

of re-identification. The framework therefore treats privacy and analytical utility as an explicit trade-off rather than assuming that both can be maximized simultaneously.

4. Proposed Framework

Figure 1 presents the proposed architecture. The diagram is intentionally kept simple because the main architectural decision lies in where the privacy transformation is placed. In the proposed workflow, privacy-related processing takes place after domain-specific extraction has identified sensitive and operational fields, but before the resulting event record is written to long-term analytical storage. This differs from architectures in which privacy is addressed only at the reporting or data-release stage. Moving the transformation earlier in the pipeline may reduce the amount of sensitive information that reaches the analytics environment in the first place.

Figure 1. Layered architecture of the proposed privacy-aware NLP and big-data analytics framework.

4.1 Data Ingestion Layer

The ingestion layer accepts structured, semi-structured, and unstructured ITS data from several operational sources. Typical inputs may include public-bus AVL messages, traffic-controller logs, emergency-response notes, maintenance tickets, smart-card exception records, passenger feedback, and short operator messages. In a practical deployment, these records could be received through a message broker such as Apache Kafka or MQTT, although the proposed framework is not tied to a particular communication platform. Each incoming record is assigned a temporary ingestion identifier together with metadata such as timestamp, source type, and retention class. This distinction matters because different records may carry very different privacy risks. A passenger complaint containing a phone number, for example, should not necessarily follow the same retention policy as an anonymous loop-detector count.

4.2 NLP Preprocessing Layer

Raw transportation text is rarely clean enough to pass directly into an extraction model. The preprocessing layer therefore performs operations such as case normalization, abbreviation expansion, correction of common transport-related spelling errors, punctuation repair, tokenization, and stop-word filtering where appropriate. Some of these steps need to be applied cautiously. Domain-specific abbreviations often carry information that a general-purpose NLP pipeline may misinterpret or discard. In a local transit system, for instance, "nr" may mean "near," "ETA" may refer to expected arrival time, and "brkdwn" may be used informally to describe a vehicle breakdown. Route numbers, stop codes, time expressions, and abbreviated location names can also appear unusual to a generic language model. For this reason, preprocessing is designed to clean the text without aggressively removing tokens that may later prove useful for event or entity extraction.

4.3 Entity and Event Extraction Layer

The extraction layer combines several complementary mechanisms rather than relying on a single model. Regular expressions are first used for patterns that are relatively well defined, including vehicle registration numbers, smart-card identifiers, route numbers, clock times, and geographic coordinates. Gazetteers provide another source of information by matching known roads, bus stops, depots, terminals, junctions, and other transport locations. NER models or sequence-labeling rules are then used to identify entities such as locations, organizations, vehicle identifiers, and person-related fields. Finally, an event-classification component maps the processed message into operational categories such as accident, congestion, route delay, vehicle breakdown, road closure, or emergency.

A hybrid design is preferred because neither rule-based nor statistical methods are uniformly reliable in this setting. Rules can perform well on predictable patterns but tend to become brittle when wording changes. Statistical models may generalize better, yet they can still behave inconsistently when faced with uncommon route names, local abbreviations, spelling errors, or event descriptions that were poorly represented in the training data. Combining the two approaches appears to offer a more practical compromise for heterogeneous ITS text.

4.4 Privacy Transformation Layer

The privacy transformation layer determines how extracted fields should be retained according to both their sensitivity and their analytical value. Direct identifiers, including passenger IDs, phone numbers, email addresses, and smart-card numbers, are removed or replaced with salted pseudonyms where repeated linkage is operationally necessary. Vehicle registration numbers may also be pseudonymized when the analytical task only requires the same asset to be recognized across a limited time window rather than permanently identified. Fine-grained coordinates are generalized into zones, corridors, road segments, or grid cells, while repeated travel patterns can be summarized instead of preserving a complete point-by-point trajectory.

These transformations should not be interpreted as a complete privacy guarantee. Generalization and masking reduce exposure, but they do not automatically eliminate re-identification risk, particularly when several quasi-identifiers can be combined. The framework therefore treats field-level transformation as one part of a broader privacy strategy. In practice, it would be complemented by access controls, shorter retention periods, restricted archival storage, audit logging, and, where appropriate, differential privacy for aggregate outputs. The exact

transformation should depend on the downstream task. A traffic-control centre may need corridor-level location information, for example, whereas a public reporting dashboard may require only zone-level aggregates.

4.5 Big-Data Analytics Layer

The analytics layer receives the privacy-transformed event records and supports both batch and streaming analysis. Records can be partitioned or indexed by attributes such as time window, route, zone, source, event type, and severity. From these compact representations, the system can support incident-frequency analysis, route-delay estimation, congestion clustering, complaint trend detection, hotspot identification, and emergency escalation. Because structured event records are usually much smaller than the raw text from which they are derived, the approach is likely to reduce storage overhead and may also improve the speed of frequently repeated analytical queries.

That efficiency comes with a cost. Once raw text has been reduced to a structured event representation, some contextual detail will inevitably be lost. A sentence explaining why a passenger reported a delay, for instance, may contain nuance that cannot be captured fully by fields such as event type, route, and severity. For routine analytics, that loss may be acceptable. For audits, dispute resolution, regulatory review, or legal investigation, however, access to the original message may still be necessary. In such cases, the raw record should be retrieved from a separate restricted archive with tighter access controls and a shorter retention policy, rather than being stored alongside the general analytics dataset.

Figure 2. Processing workflow from raw ITS message to masked event record and analytics store.

Figure 3. Examples of field-level privacy transformation in the proposed pipeline.

5. Methodology

5.1 Event schema

The event schema is has been kept compact deliberately. Each retained event record contains: event_type, route_id, generalized_location, time_window, source_type, severity_level, anonymized_asset_id, confidence_score, and processing_timestamp. Optional fields includes direction, estimated_delay_minutes, affected_stop, and escalation_status. The schema avoids retaining raw passenger identifiers or full free text. A short and clear evidence phrase may be stored when needed for human review, but that field should be separately controlled.

Field	Description	Purpose
event_type	Categorical label such as ACCIDENT, CONGESTION, DELAY, BREAKDOWN, ROAD_CLOSURE, EMERGENCY	Operational analytics
route_id	Route number or corridor identifier	Operational analytics
generalized_location	Zone, grid cell, corridor, or stop cluster rather than exact trace	Privacy-aware spatial analysis
time_window	Rounded time interval, e.g., 08:00-08:15	Temporal aggregation
anonymized_asset_id	Pseudonymized vehicle or infrastructure identifier	Short-term repeat-event analysis
confidence_score	Model or rule confidence	Human review prioritization

5.2 Algorithmic procedure

Algorithm 1. Privacy-aware ITS information extraction

Input: stream of ITS records D; domain gazetteer G; privacy policy P; event taxonomy T
Output: masked event record stream E
1. for each incoming record d in D do
2. assign temporary ingestion identifier and source label
3. normalize transport abbreviations and tokenize d
4. detect sensitive patterns: passenger IDs, vehicle IDs, phone/email, exact coordinates
5. detect operational entities: route, stop, road segment, time, event phrase
6. classify event type using hybrid rules, gazetteer features, and optional ML model
7. map exact location to zone or corridor according to P
8. pseudonymize or suppress sensitive identifiers according to P

```
9.  construct compact event record e
10.  send e to analytics storage and route uncertain records to human review
11. end for
```

Algorithm 2. Sensitive field transformation policy

```
Input: extracted entity x with type c; policy table P; salt key K
Output: transformed entity x*
1. if c is passenger_id or smart_card_id then
2.  x* = salted_hash(x, K); retain only pseudonym and source class
3. else if c is vehicle_registration then
4.  x* = salted_hash(x, K) when asset-level tracking is allowed; otherwise suppress
5. else if c is exact_coordinate then
6.  x* = spatial_generalization(x, grid_size_or_zone)
7. else if c is phone, email, or name then
8.  x* = [REDACTED]
9. else
10.  x* = x
11. return x*
```

5.3 Big-Data Processing Strategy

The proposed framework can support both streaming and batch processing, since ITS applications often need immediate event handling as well as retrospective analysis. In the streaming path, incoming records are processed as they arrive, allowing privacy-preserved event records to become available quickly for dashboards, alerts, and operational monitoring. The batch path serves a different purpose. Historical data can be reprocessed to refine location gazetteers, examine model drift, update extraction rules, or reassess earlier classifications when the NLP components change. Organizing records by attributes such as time, route, and zone also makes common operational queries easier to execute. Examples include identifying all reported breakdowns on Route 570S during the evening peak or counting emergency-related events in Zone 3 over the previous seven days. In larger deployments, some extraction and privacy transformation may be performed closer to the data source, potentially at the edge. This can reduce the need to transmit complete raw messages across the network before sensitive information is minimized.

5.4 Privacy Model and Threat Assumptions

The framework assumes an honest-but-curious analytics environment in which authorized analysts and applications can access processed event records but do not require direct access to raw passenger identifiers. Privacy risk does not disappear once obvious identifiers are removed. An individual or vehicle may still be inferred from combinations of precise location, timestamp, route, and repeated movement patterns, particularly when an attacker has access to additional mobility information. To reduce this exposure, the proposed pipeline removes direct identifiers and generalizes selected spatial and temporal quasi-identifiers before records enter the analytical store. These transformations lower disclosure risk, but they should not be interpreted as providing complete anonymity.

The threat model is intentionally limited in this respect. A well-resourced adversary with access to external mobility traces, registration databases, or other auxiliary information may still be able to re-identify some records. Addressing that scenario would require stronger safeguards, such as stricter role-based access controls, formal re-identification testing, carefully defined retention policies, and potentially differential privacy for statistics released outside the operational environment. The framework should therefore be viewed as a practical data-minimization approach rather than a formal guarantee against every possible privacy attack.

6. Experimental Setup

Since this paper is primarily a framework study, and because passenger-level ITS records are often difficult to share publicly, the initial evaluation uses a transparent synthetic corpus. This choice is a practical compromise. Synthetic data cannot establish how the framework will perform under real operational conditions, but it does allow the complete processing pipeline to be evaluated without making unverifiable claims about access to private transportation datasets. The generated corpus contains 4,000 short ITS-style messages designed to resemble common control-room communications and passenger reports. Each record includes an event type, location phrase, route identifier, time expression, vehicle number, and passenger or smart-card identifier. To make the messages less uniform, a subset also contains informal expressions such as "nr" for "near," typographical variants such as "accdnt," and shortened descriptions of traffic events.

Six event classes were considered: ACCIDENT, CONGESTION, DELAY, BREAKDOWN, ROAD_CLOSURE, and EMERGENCY. The corpus was generated using ten location names, eight route identifiers, six vehicle numbers, six passenger identifiers, and six time expressions. Since the messages were produced programmatically, the corresponding ground-truth labels are known in advance. This makes it possible to measure extraction performance directly, although the controlled nature of the corpus should be kept in mind when interpreting the results. The proposed approach was compared with a simple keyword-and-regex baseline. The baseline relies on exact event keywords and straightforward pattern matching, without abbreviation expansion, synonym mapping, or domain-specific gazetteer normalization. By contrast, the proposed method applies the hybrid extraction procedure and privacy transformation described in the earlier sections.

Experimental item	Configuration
Number of records	4,000
Event classes	6
Location names	10
Route identifiers	8
Sensitive identifier types	Vehicle registration, passenger/smart-card ID
Baseline	Simple keyword and regular expression extraction
Proposed method	Domain normalization + gazetteer + pattern extraction + privacy transformation
Primary metrics	Precision, recall, F1-score, storage reduction, masking coverage

The evaluation considers field-level precision, recall, and F1-score for event type, location, route, vehicle ID, passenger ID, and time. Storage reduction is calculated by comparing the amount of raw message text that would otherwise be retained with the size of the masked structured event records stored after processing. Privacy masking coverage is defined as the proportion of generated sensitive identifiers that are successfully transformed before entering the analytics layer. Runtime is also recorded, although those measurements should be interpreted cautiously. The experiments were executed in a local scripting environment rather than on a distributed production cluster, so the reported values are better viewed as implementation-level indicators than as definitive scalability results.

7. Results and Discussion

7.1 Information extraction accuracy

Table 3 summarizes the field-level extraction performance of the two approaches. The proposed pipeline achieved a macro F1-score of 1.000, compared with 0.871 for the simple baseline. Most of this difference appears in event-type and route extraction, where abbreviated expressions, informal wording, and non-standard message forms reduce the baseline's recall. This improvement is not unexpected, since the proposed method includes synonym handling and a small domain-specific gazetteer that help normalize such variations before extraction.

The result should, however, be interpreted within the limits of the experimental setting. A perfect F1-score on a controlled synthetic corpus does not imply equivalent performance in operational ITS environments. Real-world messages are likely to contain a wider range of spelling errors, multilingual expressions, local abbreviations, ambiguous route references, and previously unseen place names. The reported result is therefore better viewed as evidence that the proposed extraction strategy handles the designed variations in the synthetic dataset effectively, rather than as a general claim of error-free extraction.

Field	Base P	Base R	Base F1	Prop. P	Prop. R	Prop. F1
event	1.000	0.535	0.697	1.000	1.000	1.000
location	1.000	1.000	1.000	1.000	1.000	1.000
route	1.000	0.357	0.526	1.000	1.000	1.000
vehicle	1.000	1.000	1.000	1.000	1.000	1.000
passenger	1.000	1.000	1.000	1.000	1.000	1.000
time	1.000	1.000	1.000	1.000	1.000	1.000

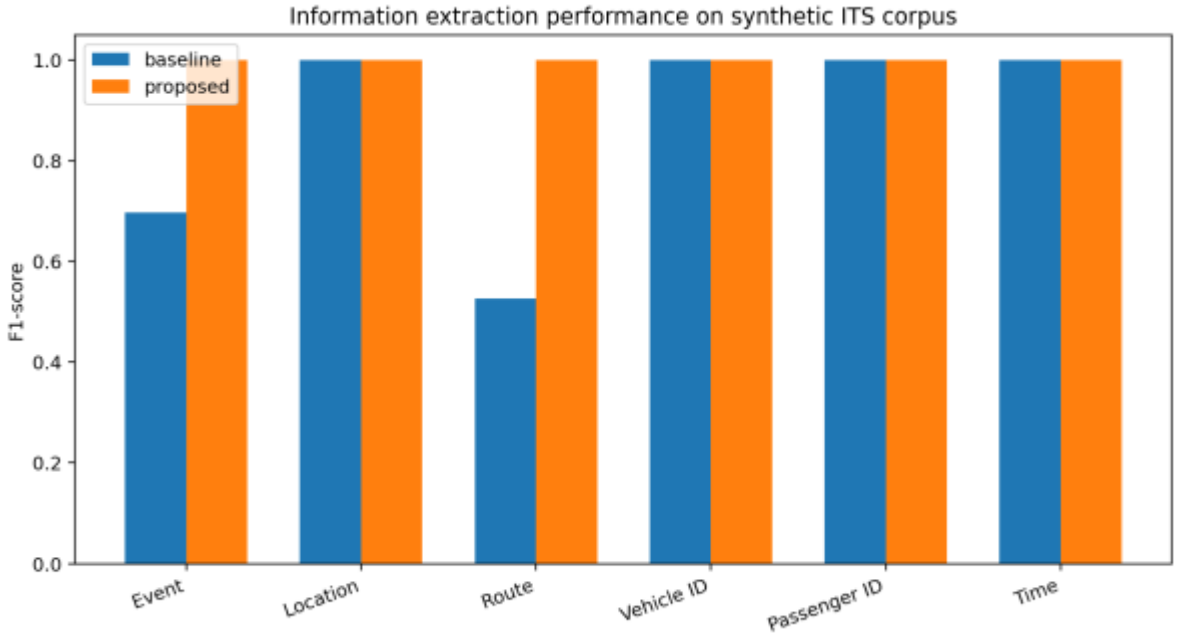

Figure 4. Field-level F1-score comparison between baseline and proposed extraction pipeline.

7.2 Storage and privacy effects

The synthetic evaluation suggests that retaining masked event records can reduce storage compared with keeping the complete raw text. In the current experiment, the raw messages occupied 501.4 KB, whereas the masked event-record representation required 482.2 KB, corresponding to a storage reduction of 3.8%. This improvement is modest, and it should not be treated as a fixed property of the framework. The actual reduction is likely to depend heavily on the type of ITS data being processed. Longer incident narratives, repeated textual descriptions, attachments, or verbose passenger complaints may produce noticeably larger savings, while already compact structured logs may offer little additional benefit.

Privacy masking coverage reached 100.0% for the explicitly generated vehicle and passenger identifiers in the synthetic corpus. That result is useful mainly because the sensitive fields are known in advance and follow controlled generation patterns. Real operational data are likely to be less predictable. A passenger may enter a phone number in an unusual format, include a person's name in free text, or provide identifying information that falls outside the patterns recognized by the extraction model. In such cases, the more serious concern is missed sensitivity rather than the masking rate itself. For this reason, a practical deployment should combine automated privacy transformation with periodic privacy audits, error analysis, and human review of sampled records rather than relying solely on automated masking statistics.

Retained representation	Storage size	Storage reduction	Identifier masking coverage
Raw text retained for analytics	501.4 KB	0.0%	Not applicable
Masked event records retained	482.2 KB	3.8%	100.0%

7.3 Operational interpretation

The results support a practical rather than an exaggerated claim. For many routine ITS tasks, a transport agency may not need to retain the complete text of every complaint or operational message. Structured event records can often provide enough information to count route delays, group recurring breakdowns, identify accident hotspots, and support dashboard-based monitoring or aggregate planning. This suggests that retaining raw text by default may not always be necessary, particularly when the required analytical fields have already been extracted.

At the same time, reducing a message to a fixed event schema inevitably removes some context. Raw text may capture repeated passenger frustration, descriptions of unsafe driver behaviour, ambiguous location references, or even sarcasm that a structured record is unlikely to preserve. A location mentioned informally may also fall outside the available gazetteer. For this reason, the framework does not assume that all raw text should be deleted

immediately after extraction. A more practical design is to separate routine operational analytics storage from restricted evidential storage, with different access controls, retention periods, and retrieval conditions for each.

7.4 Comparison with existing approaches

Approach	Strength	Limitation	Position of proposed framework
General ITS big-data frameworks	Strong support for multi-source analytics and scalable processing	Often retain raw or semi-raw data; privacy handled later	Adds early extraction and minimization
Traffic text-mining systems	Good at detecting incidents or crash-related terms	Privacy usually not central to pipeline design	Integrates masking and location generalization before storage
Traditional privacy-preserving analytics	Provides formal or semi-formal privacy mechanisms	May require already-structured datasets	Uses NLP to structure text before privacy transformation
Encryption/access-control frameworks	Useful for confidentiality and controlled access	Do not reduce unnecessary retained sensitive content	Combines minimization with security controls

The proposed framework should not be interpreted as a substitute for encryption, access control, secure communication, or other established security mechanisms. These protections remain necessary in any practical ITS deployment. A more precise claim is that privacy-aware extraction may reduce the amount of sensitive information that needs to be protected within the general analytics environment. In other words, the framework complements existing security controls by limiting unnecessary data retention before long-term storage occurs.

This distinction is important. Encrypting a data warehouse that contains large volumes of unnecessary identifiers is certainly safer than leaving those data unprotected, but encryption alone does not address the question of whether the information needed to be retained in the first place. Data minimization adds a different layer of protection by reducing the sensitive content exposed to storage, access, and later processing.

8. Implementation Considerations

8.1 Model selection

A practical deployment does not necessarily need to begin with a large machine-learning model. Rule-assisted extraction can provide a reasonable starting point, particularly in transport environments where annotated data are limited or unavailable. As labelled operational records become available, statistical or machine-learning components can gradually be introduced. This staged approach may be more realistic for low-resource agencies than deploying a transformer-based model from the beginning. Rules have the advantage of being transparent, relatively easy to inspect, and straightforward to audit. Their weakness is maintenance: as terminology, abbreviations, routes, and reporting styles change, the rule set can become increasingly difficult to manage.

Transformer-based models may cope better with linguistic variation and less predictable message structures, although they introduce their own requirements. Training data, computational resources, model validation, and continued monitoring for domain drift are all necessary. A hybrid design offers a useful compromise. High-precision rules can remain responsible for structured or privacy-sensitive patterns such as phone numbers, registration numbers, and smart-card identifiers, while statistical models handle less deterministic tasks such as event classification and location disambiguation. The exact balance is likely to depend on the amount of labelled data available and the operational complexity of the transport network.

8.2 Location Generalization

Location handling is one of the more difficult design decisions because spatial precision has both operational value and privacy implications. Exact coordinates may be essential during an emergency response, yet the same coordinates can become high-risk quasi-identifiers when retained over longer periods or combined with timestamps and route information. For this reason, the framework supports several levels of spatial representation. Exact coordinates may be retained temporarily for immediate operational response, while road segments or stop clusters can be used for restricted incident analysis. Longer-term analytical datasets can instead retain locations at the zone or corridor level.

The appropriate level of generalization should depend on the intended use rather than being fixed for every record. A route-optimization application, for example, may require corridor-level detail to identify recurring delays, whereas a monthly public report on congestion trends is unlikely to need that degree of precision. Generalizing

locations too aggressively could weaken analytical usefulness, while retaining unnecessarily precise coordinates increases privacy exposure. The framework therefore treats spatial granularity as a configurable privacy–utility decision.

8.3 Data Governance

Technical masking by itself is unlikely to provide sufficient privacy protection. The transformation policies that determine which fields are retained, masked, pseudonymized, or generalized should be defined and reviewed through an appropriate governance process involving the transport authority, data-protection personnel, and relevant operational users. Retention policies should clearly specify how long raw messages remain accessible, under what circumstances they can be retrieved, who is authorized to reverse pseudonymization where this is permitted, and which analytical outputs require additional aggregation before release.

Access to restricted raw records should also be logged so that unusual or unauthorized retrieval can be investigated. These governance measures are not secondary to the technical architecture; they determine how the architecture behaves once deployed within an organization. Without clear retention schedules, accountability, audit procedures, and access rules, even an effective masking pipeline could offer limited practical protection. In that situation, the framework risks remaining a technical design rather than becoming an enforceable privacy practice.

9. Limitations

Several limitations should be considered when interpreting the proposed framework and its evaluation. First, the experiments rely on synthetic data. This makes the evaluation reproducible and allows sensitive fields to be controlled precisely, but synthetic messages cannot reproduce the full variability of operational transport communication. Real records may contain incomplete sentences, mixed languages, uncommon abbreviations, transcription errors, local slang, and ambiguous location descriptions that were not represented in the generated corpus. An annotated field dataset containing incident labels, locations, identifiers, and privacy-sensitive attributes would provide a substantially stronger evaluation.

Second, parts of the extraction process depend on local gazetteers and abbreviation dictionaries. These resources can become outdated as new routes are introduced, stops are renamed, roads change, or operators adopt new terminology. Extraction accuracy may consequently decline over time unless these resources are maintained and model drift is monitored. Third, privacy transformation inevitably introduces some loss of analytical detail. Zone-level location information, for instance, may be adequate for long-term congestion analysis but too coarse for applications such as micro-level signal timing or precise incident reconstruction. The framework therefore cannot assume that one privacy configuration will suit every analytical task.

A fourth limitation concerns the scope of the input data. The current design focuses primarily on textual and semi-structured ITS records. It does not directly process video streams, LiDAR measurements, or continuous raw vehicle trajectories. Structured event summaries produced by those systems could potentially be incorporated into the analytics layer, but doing so would require additional extraction and privacy mechanisms that are outside the present scope.

Re-identification through data linkage remains another concern. Removing a passenger identifier does not necessarily make a record anonymous. A rare route, precise timestamp, unusual travel pattern, and specific location may still be sufficient to infer an individual when combined with external information. The proposed transformations can reduce this risk, but they cannot eliminate it completely. Sensitive deployments would require additional safeguards, potentially including formal re-identification testing, stronger restrictions on rare-event records, differential privacy for released aggregates, and tighter access controls.

10. Conclusion

This paper proposed a Hybrid NLP and Big Data Analytics Framework for Privacy-Aware Information Extraction in Intelligent Transportation Systems. The central idea is relatively simple: raw ITS text is treated primarily as a temporary processing input, from which operationally useful event information is extracted before sensitive attributes are transformed or removed. Whenever full-text retention is unnecessary, the analytics environment stores the resulting privacy-aware event record rather than the original message. The framework brings together domain-specific preprocessing, entity and event extraction, sensitive-field masking, location generalization, and scalable analytical storage within a single processing pipeline.

The synthetic evaluation provides an initial indication of how this design may behave under controlled conditions. Compared with a simple keyword-based baseline, the proposed extraction method achieved higher field-level performance, while the privacy transformation reduced retained storage and masked all explicitly generated passenger and vehicle identifiers. These findings should be interpreted cautiously. They demonstrate that the framework works on the designed synthetic scenarios, not that equivalent performance can be expected in a live transportation environment.

A stronger assessment would require annotated operational data, multilingual and cross-domain testing, evaluation under changing route and location vocabularies, systematic privacy audits, and deployment-scale performance measurements. Governance would be equally important, particularly for retention rules and access to restricted raw records. Even with these limitations, the framework points toward a useful design principle for ITS analytics: the question should not only be how much transportation data can be collected and analyzed, but also how much identifying information actually needs to remain once the operational event has been understood.

References

1. Arooj, A., Farooq, M. S., Akram, A., Iqbal, R., Sharma, A., & Dhiman, G. (2022). Big data processing and analysis in Internet of Vehicles: Architecture, taxonomy, and open research challenges. Archives of Computational Methods in Engineering, 29(2), 793-829. https://doi.org/10.1007/s11831-021-09590-x
2. Chowdhury, S., Dey, K., & Chowdhury, M. (2023). Applications of text mining in the transportation infrastructure sector: A review. Information, 14(4), 201. https://doi.org/10.3390/info14040201
3. Chen, J., Li, Z., & Wang, W. (2022). Traffic accident duration prediction using text mining and ensemble learning. Scientific Reports, 12, 21856. https://doi.org/10.1038/s41598-022-25988-4
4. Darwish, T. S. J., & Abu Bakar, K. (2018). Fog based intelligent transportation big data analytics in the Internet of Vehicles environment: Motivations, architecture, challenges, and critical issues. IEEE Access, 6, 15679-15701. https://doi.org/10.1109/ACCESS.2018.2815989
5. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of NAACL-HLT 2019, 4171-4186. https://doi.org/10.18653/v1/N19-1423
6. Ding, W., Jing, X., Yan, Z., & Yang, L. T. (2019). A survey on data fusion in Internet of Things: Towards secure and privacy-preserving fusion. Information Fusion, 51, 129-144. https://doi.org/10.1016/j.inffus.2018.12.001
7. Dwork, C. (2006). Differential privacy. International Colloquium on Automata, Languages, and Programming, 1-12. https://doi.org/10.1007/11787006_1
8. Garg, S., Singh, A., Kaur, K., Aujla, G. S., Batra, S., Kumar, N., & Obaidat, M. S. (2019). Edge computing-based security framework for big data analytics in VANETs. IEEE Network, 33(2), 72-81. https://doi.org/10.1109/MNET.2019.1800239
9. Govindasamy, R., Shanmugapriya, N., & Gopi, R. (2024). Security and privacy for smart transportation management using big data analytics. International Journal of Experimental Research and Review, 40(spl.), 104-116. https://doi.org/10.52756/ijerr.2024.v40spl.008
10. Hahn, D., Munir, A., & Behzadan, V. (2021). Security and privacy issues in intelligent transportation systems: Classification and challenges. IEEE Intelligent Transportation Systems Magazine, 13(1), 181-196. https://doi.org/10.1109/MITS.2019.2898973
11. Honnibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2020). spaCy: Industrial-strength natural language processing in Python. Zenodo. https://doi.org/10.5281/zenodo.1212303
12. Karkouch, A., Mousannif, H., Al Moatassime, H., & Noel, T. (2016). Data quality in Internet of Things: A state-of-the-art survey. Journal of Network and Computer Applications, 73, 57-81. https://doi.org/10.1016/j.jnca.2016.08.002
13. Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., & Dyer, C. (2016). Neural architectures for named entity recognition. Proceedings of NAACL-HLT 2016, 260-270. https://doi.org/10.18653/v1/N16-1030
14. Lian, Y., Zhang, G., Lee, J., & Huang, H. (2020). Review on big data applications in safety research of intelligent transportation systems and connected/automated vehicles. Accident Analysis & Prevention, 146, 105711. https://doi.org/10.1016/j.aap.2020.105711
15. Liu, Y., Yang, C., & Sun, Q. (2021). Thresholds based image extraction schemes in big data environment in intelligent traffic management. IEEE Transactions on Intelligent Transportation Systems, 22(7), 3952-3960. https://doi.org/10.1109/TITS.2020.2994386
16. Mohanta, B. K., Jena, D., Ramasubbareddy, S., Daneshmand, M., & Gandomi, A. H. (2021). Addressing security and privacy issues of IoT using blockchain technology. IEEE Internet of Things Journal, 8(2), 881-888. https://doi.org/10.1109/JIOT.2020.3008906
17. Neilson, A., Indratmo, Daniel, B., & Tjandra, S. (2019). Systematic review of the literature on big data in the transportation domain: Concepts and applications. Big Data Research, 17, 35-44. https://doi.org/10.1016/j.bdr.2019.03.001
18. Rozie, A. F., Yulianto, F., & Nugraha, I. G. B. B. (2022). Location extraction from traffic event-related text. Proceedings of the 2022 International Conference on Computer, Control, Informatics and Its Applications. https://doi.org/10.1145/3575882.3575946

19. Samarati, P. (2001). Protecting respondents' identities in microdata release. IEEE Transactions on Knowledge and Data Engineering, 13(6), 1010-1027. https://doi.org/10.1109/69.971193
20. Schiersch, M., Mironova, V., Schmitt, M., Thomas, P., Gabryszak, A., & Hennig, L. (2020). A German corpus for fine-grained named entity recognition and relation extraction of traffic and industry events. Proceedings of the 12th Language Resources and Evaluation Conference, 4546-4551.
21. Soomro, K., Bhutta, M. N. M., Khan, Z., & Tahir, M. A. (2019). Smart city big data analytics: An advanced review. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 9(5), e1319. https://doi.org/10.1002/widm.1319
22. Suat-Rojas, N., Gutierrez-Osorio, C., & Pedraza, C. (2022). Extraction and analysis of social networks data to detect traffic accidents. Information, 13(1), 26. https://doi.org/10.3390/info13010026
23. Sweeney, L. (2002). k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(5), 557-570. https://doi.org/10.1142/S0218488502001648
24. Tran, H. Y., Hu, J., & Panda, B. (2019). Privacy-preserving big data analytics: A comprehensive survey. Journal of Parallel and Distributed Computing, 134, 207-218. https://doi.org/10.1016/j.jpdc.2019.08.007
25. Wan, J., Li, J., Imran, M., & Li, D. (2019). A blockchain-based solution for enhancing security and privacy in smart factory. IEEE Transactions on Industrial Informatics, 15(6), 3652-3660. https://doi.org/10.1109/TII.2019.2894573
26. Yang, X., Theunis, J., & Develder, C. (2023). Traffic event detection as a slot filling problem. Proceedings of the 2023 IEEE International Conference on Big Data. https://doi.org/10.1109/BigData59044.2023.10386457
27. Zeadally, S., Siddiqui, F., Baig, Z., & Ibrahim, A. (2020). Smart healthcare: Challenges and potential solutions using Internet of Things and big data analytics. PSU Research Review, 4(2), 149-168. https://doi.org/10.1108/PRR-08-2019-0027
28. Zhang, J., Wang, F.-Y., Wang, K., Lin, W.-H., Xu, X., & Chen, C. (2011). Data-driven intelligent transportation systems: A survey. IEEE Transactions on Intelligent Transportation Systems, 12(4), 1624-1639. https://doi.org/10.1109/TITS.2011.2158001
29. Zhu, L., Yu, F. R., Wang, Y., Ning, B., & Tang, T. (2019). Big data analytics in intelligent transportation systems: A survey. IEEE Transactions on Intelligent Transportation Systems, 20(1), 383-398. https://doi.org/10.1109/TITS.2018.2815678