Politecnico di Torino

Data Science and Engineering

A.a. 2024/2025

Enhancing a machine learning-based social media threat intelligence automotive framework

Supervisors:

Stefano Di Carlo
Alessandro Savino
Nicola Scarano
Luca Mannella

Candidate:

Yan Xu

Abstract

Open Source Cyber Threat Intelligence (OSCTI) is critical for addressing the rapidly evolving automotive cybersecurity threat landscape. Platforms such as YouTube and Reddit have increasingly become informal yet highly influential ecosystems where users exchange technical knowledge, demonstrate operational procedures, and disseminate tools related to vehicle modifications, Engine Control Unit (ECU) reprogramming, and emissions system tampering. Consequently, extracting actionable insights from such unstructured social media content is essential for supporting cybersecurity decision-making in the automotive domain. Despite the considerable potential of social media data, leveraging it for OSCTI remains challenging. User-generated posts are heterogeneous, noisy, and highly context-dependent, often intertwining legitimate maintenance discussions with illicit modification activities, incomplete descriptions, or ambiguous intent. Traditional keyword-based or rule-based approaches fail to capture these nuances, underscoring the need for methods capable of deeper semantic understanding. To address these challenges, this thesis work constructs an analytical framework that integrates two detecting approaches: embedding-based clustering and instruction-tuned LLM direct classification. Experiments show that the embedding-based clustering path does not yield meaningful semantic separation between tampering and non-tampering content, indicating that unsupervised embeddings alone are insufficient for this detection task. We then evaluated the instruction-tuned LLM direct classification approach using three open-source models. These models demonstrated a clearer ability to identify tampering-related semantics than clustering, showing emerging sensitivity to tampering intent despite their imperfect performance. During the evaluation phase, we explored a multi-dimensional tampering-intent scoring approach that aggregated weighted dimension scores into final tampering decisions. However, the method exhibited instability and limited discriminative power on the golden dataset, making it unsuitable as a dependable validation mechanism. We subsequently applied OpenAI's GPT-4o-mini model as an external evaluator. While GPT-4o-mini exhibited strong sensitivity to tampering signals, its behavior remained inconsistent for non-tamper and uncertain cases. Overall, these findings highlight both the potential and limitations of current LLM-based approaches in supporting OSCTI workflows within the automotive cybersecurity domain.

Acknowledgements

A big thank you goes to Professor Stefano Di Carlo for his guidance and encouragement throughout this research, and to Nicola Scarano for his constant support, valuable discussions, and collaboration. Thanks to the members of the research group for their assistance with computational resources and experiments. Thanks to my family for the fundamental support to my studies.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Automotive Cybersecurity Context

Over the past decade, automotive manufacturers have undergone continuous technological and design innovations to meet consumer demands and enhance market competitiveness. Modern vehicles have consequently evolved from purely mechanical systems into highly interconnected cyber-physical systems. Contemporary vehicles integrate a large number of networked components, such as sensors, Electronic Control Units (ECUs), buses, actuators, and various other electronic modules used to monitor and control vehicle behavior[1]. This transformation marks a shift from mechanical mechanisms based on gears and drive shafts toward complex electronic architectures that communicate through multiple protocols. However, with the continuous increase in the number of electronic components and the growing interconnectivity among subsystems, the overall complexity of modern vehicles has risen dramatically. The denser dependencies and interaction paths between systems have not only enhanced vehicle intelligence but also significantly expanded the potential attack surface. Attacks that previously required physical access can now be triggered remotely through wireless interfaces, in-vehicle network buses, or even cloud services. Consequently, vehicles are exposed to a wider variety of cybersecurity threats with more complex attack chains—ranging from localized attacks targeting individual ECUs to deep intrusions that penetrate entire vehicle networks across bus systems. This has led to an unprecedented proliferation of security risks.

Recent studies indicate that modern vehicles, due to their high connectivity and software-driven design, are becoming increasingly vulnerable to software-based attacks, which may result in severe consequences. A well-known study from the University of Washington and the University of California, San Diego demonstrated that by injecting malicious code into ECUs. Either via physical access or wireless communication interfaces, an attacker could compromise critical vehicle functions such as braking[1]. Unlike attacks on traditional IT systems, intrusions targeting vehicles directly endanger human safety, making automotive cybersecurity a crucial field of research.

Among the various attack vectors, one of the most concerning phenomena is *tampering*. Tampering refers to unauthorized modifications to a vehicle's software or hardware, such as ECUs remapping or the removal of Diesel Particulate Filters (DPF) and Exhaust Gas Recirculation (EGR) systems. While these practices are often motivated by performance enhancement, cost savings, or regulatory avoidance, they undermine vehicle reliability, environmental compliance, and overall road safety.

1.2 The Role of Social Media

Traditionally, vehicle tampering activities were mostly confined to relatively closed real world environments like repair shops or private garages. However, in recent years, tampering-related knowledge and tools have increasingly been disseminated through social media platforms such as YouTube and Reddit. With more than 2.5 billion active users [2], YouTube has become a major channel for tutorials on ECUs remapping, DPF removal, EGR disabling, and similar procedures. Similarly, active communities on online forums like Reddit share experiences, discuss modification outcomes, and trade devices used to bypass safety mechanisms. While these communities facilitate technical learning and knowledge exchange, they also inadvertently accelerate the spread of unsafe or illegal vehicle modifications, making tampering more prevalent and increasing regulatory challenges. Beyond serving as a medium for sharing knowledge and tools, social media is also a valuable source of Open Source Intelligence (OSINT). Previous research, such as the SOCMATI framework [3], has demonstrated that structured collection and analysis of publicly available posts, comments, and videos can provide meaningful insights into trends within automotive cybersecurity. Thus, vehicle tampering-related content represents a rich yet underutilized information stream that reflects emerging technical practices, potential regulatory risks and offers opportunities for informing cybersecurity protection strategies.

At the same time, relying on social media for such analysis presents challenges. User-generated content is highly heterogeneous, informal, and often incomplete. Legitimate maintenance discussions are frequently interwoven with unsafe tampering advice, and intent may be ambiguously expressed. These characteristics make social media both a valuable resource for understanding tampering behaviors and a complex medium that is difficult to analyze systematically. Extracting reliable intelligence from these platforms requires methods capable of handling ambiguous language, inconsistent details, and the highly variable nature of user-generated content.

1.3 Research Objectives and Analytical Strategy

The objective of this work is to investigate systematic and scalable approaches for analyzing vehicle tampering-related discussions on social media to support OSCTI in the automotive domain. As previously discussed, posts and videos concerning vehicle modifications are often highly informal, semantically ambiguous, and heterogeneous in content, making it challenging to distinguish legitimate maintenance discussions from content promoting unsafe or illegal modifications. Within this context, this study aims to identify which analytical strategies can effectively detect tampering-related information from large unstructured datasets such as YouTube and Reddit.

To achieve this goal, this thesis work adopts a dual-path analytical strategy combining unsupervised exploration with LLM-based classification. The study begins by filtering and organizing raw data using keywords related to emission systems and ECUs, constructing a structured research dataset. A subset of this data is then manually annotated according to clearly defined guidelines to form a gold-standard dataset for subsequent evaluation benchmarks. The first analytical path focuses on unsupervised methods, exploring whether embedding models and clustering techniques can reveal underlying patterns in the data. Each data instance is encoded into semantic vectors using modern sentence embedding models, and the resulting vector space is examined to determine whether tampering and non-tampering content naturally form separable structures. Simultaneously, the second path employs instruction-tuned large language models (LLMs) for supervised-style classification. Through carefully designed prompts, the models generate deterministic classification labels (tamper, non-tamper, or uncertain) for each data instance. Given the limited scale of the gold-standard dataset, OpenAI's deterministic classifier serves as an external reference to evaluate the consistency and reliability of open-source model outputs across the complete dataset, providing a scalable validation approach without requiring extensive manual annotation.

The rationale for this analytical design is twofold. First, exploring unsupervised embedding clustering helps determine the feasibility of such methods in this domain, which is particularly valuable for handling rapidly evolving social media ecosystems. Second, instruction-tuned LLMs offer a practical alternative for large-scale classification tasks, though their outputs require rigorous validation before deployment. By integrating both analytical paths within a unified framework, this thesis work evaluates the strengths and limitations of different approaches and identifies which techniques can provide reliable support for automotive tampering intelligence analysis.

Chapter 2

Background

2.1 Automotive Cybersecurity and Tampering Risks

With the widespread integration of networked ECUs, Advanced Driver-Assistance Systems (ADAS), and Vehicle-to-Everything (V2X) communication interfaces, automobiles are rapidly evolving from traditional mechanical devices into highly complex cyber-physical systems [1, 4]. This transformation has significantly enhanced vehicle intelligence and connectivity, but has simultaneously introduced unprecedented security risks. When critical vehicle functions rely on software and digital communication, vulnerabilities in code, firmware, or protocols can directly translate into real-world consequences such as brake failure, steering anomalies, or controlled powertrain manipulation. Multiple studies have confirmed that attackers can manipulate critical systems without physical access through remote code execution or unauthorized firmware modifications [5]. Consequently, cybersecurity has become a core requirement in vehicle design, development, and homologation processes, leading to the establishment of standards such as ISO/SAE 21434 and UNECE WP.29 [6].

Among various threat vectors, *tampering* is receiving increased attention as a long-standing yet increasingly prevalent phenomenon. Tampering is generally defined as proactive modification of vehicle software or hardware configurations, often aimed at circumventing emissions regulations, enhancing performance, or bypassing safety restrictions [3]. While some users claim motivations of personalized tuning or optimizing driving experience, these practices frequently compromise environmental compliance and may weaken system protections for critical functions. As shown by Koscher et al. [7], attackers can overwrite or inject code into ECUs, causing the vehicle to diverge from its intended behavior. This discrepancy effectively breaks the trust model and integrity of the vehicle's control system. Therefore, identifying and analyzing tampering activities concerns not only regulatory enforcement but also the security and trustworthiness of the entire connected vehicle ecosystem.

2.2 Social Media Resource

OSINT refers to the overall process in which anyone can collect and analyze information based on open-source information and create useful information[8]. In the field of cybersecurity, research has already established the integration of OSINT into analysis workflows[9, 10]. By mining user-generated content, researchers can detect critical signals at an earlier stage. Similar research directions are now emerging in automotive cybersecurity. Existing study demonstrates that knowledge related to vehicle tampering is rapidly spreading through social media platforms, creating an easily accessible ecosystem for modification knowledge[3]. Consequently, integrating OSINT monitoring and data-driven behavioral analysis into the automotive cybersecurity lifecycle is becoming increasingly crucial. This approach is essential for understanding how illicit modification behaviors evolve and propagate through cyberspace.

OSINT encompasses information gathered from publicly available sources such as blogs, technical forums, online communities, and social-media platforms. Among these sources, social media has recently attracted significant research interest: Scarano *et al.* [3] demonstrated that social platforms such as YouTube and Reddit contain abundant material related to automotive modification and system manipulation. Their SOCMATI framework formalized a pipeline for transforming unstructured social-media content into analyzable cybersecurity intelligence, classifying posts into categories such as tampering guidance, tool advertisement, and security bypass discussion. This approach established social media as a legitimate data source for understanding the social diffusion of security-critical practices.

However, the exploitation of social-media data for OSINT introduces several methodological challenges. First, the data are linguistically heterogeneous, with informal syntax, domain-specific jargon, and frequent multimedia elements that complicate standard text processing [3]. Second, the high volume and dynamic nature of user-generated content require scalable computational approaches that can generalize across contexts without extensive human annotation [11]. These constraints motivate the adoption of advanced Natural Language Processing (NLP) and semantic modeling techniques, which can extract latent patterns and infer intent from noisy textual environments.

2.3 Semantic Representations

Semantic representation models aim to encode textual information into dense numerical vectors that capture both syntactic and contextual relationships among linguistic units. Early representations such as TF–IDF and Word2Vec relied on surface-level co-occurrence statistics [12], which were limited in their ability to model long-range dependencies or semantic compositionality. The introduction of transformer architectures [13] enabled contextualized embeddings that dynamically adjust token representations according to their surrounding context. Subsequent pre-trained models such as BERT [14] and its sentence-level extensions (e.g., Sentence-BERT, E5) [15, 16] improved semantic alignment by training on large-scale text similarity objectives. These embeddings serve as fundamental representations for a wide range of downstream NLP tasks such as clustering, semantic search, and intent detection, where distance metrics between vectors reflect degrees of semantic relatedness [17, 11]. More recently, embeddings derived from instruction-tuned LLMs (e.g., Llama 3, Mistral) have shown superior generalization across domains, as they encode semantic relations aligned with human intent rather than task-specific objectives [18, 19]. Since modern text understanding tasks rely on vector representations that encode contextual and syntactic information, transforming textual data into embeddings provides a mathematically grounded way to analyze semantic similarity.

2.4 Unsupervised Learning and Clustering

Unsupervised learning, particularly clustering, has long been employed to uncover latent structures in text corpora without the need for labeled data [20, 21]. Clustering algorithms aim to group semantically or statistically similar data points together, facilitating tasks such as topic discovery, anomaly detection, and knowledge organization across large text corpora. Unlike supervised learning, which relies on predefined categories, clustering infers structure directly from data distributions, making it suitable for exploratory analysis in domains where labeled datasets are unavailable or incomplete. Recent research has demonstrated that combining transformer-based embeddings with clustering enables effective discovery of thematic structures and semantic groupings in social media datasets [11, 17].

To explore whether semantically meaningful clusters can reflect tampering-related patterns, this study applies K-Means clustering to the embedding representations of posts and videos. Among partition-based clustering algorithms, K-Means remains one of the most widely used due to its simplicity, scalability, and empirical robustness [21]. By partitioning the high-dimensional embedding space into distinct clusters, the algorithm reveals how similar content naturally groups according to its latent semantics. In the context of this research, setting the number of clusters to two allows an empirical examination of whether tampering and non-tampering content form separable regions in vector space. Hence, clustering serves not only as an unsupervised exploratory tool but also as a diagnostic mechanism to evaluate how well embedding models capture the semantic distinctions relevant to automotive tampering behavior.

2.5 Large Language Models (LLMs)

Large Language Models (LLMs) are transformer-based neural architectures trained on massive text corpora to predict the next token in a sequence [13], have proven valuable for understanding and manipulating natural language and demonstrated exceptional zero-shot and few-shot reasoning capabilities across a wide range of classification and moderation tasks [22, 18, 19]. For example, Fayyazi et al. [5] explored using LLMs, such as GPT-3.5 and Bard, along with supervised training-based BaseLLMs, to classify cybersecurity descriptions into ATT&CK tactics.

In this thesis work we employ two categories of LLMs with distinct purposes:

1. **Open-source instruction-tuned LLMs for direct tampering classification.** We rely on three lightweight and widely adopted models:

    - **Llama 3**[23] - Meta-Llama-3-8B-Instruct.
    - **Mistral 7B**[19] - Mistral-7B-Instruct-v0.2
    - **Qwen**[24] - Qwen2.5-7B-Instruct

    These models are used to perform the **tamper/non_tamper/uncertain** classification on Reddit posts and Youtube videos. Each item is fed to the model through a structured instruction prompt designed specially for automotive tampering detection.

    We select these LLMs because these models belong to instruction-tuned families that are known to deliver strong zero-shot and few-shot performance, making them suitable for classification tasks without domain-specific fine-tuning. Moreover, their moderate parameter size (7–8B) strikes a practical compromise between accuracy and computational cost, allowing us to process the full Reddit and YouTube datasets efficiently while still retaining enough representational capacity to reason about tampering behavior.

2. **GPT-4o-mini for external evaluation.** To ensure that the open-source model outputs were consistent, we employ OpenAI's GPT-4o-mini as an evaluator performing two validation tasks: (1) **7-dimensional semantic scoring**, assigns tampering-related semantic scores (0–1) based on the multi-dimensional

framework adapted from Huq and Suleiman (2025)[25]. (2) **Independent label assignment**, produces an external **tamper / non_tamper / uncertain** label used to construct dataset for validation and verified by the manually annotated gold dataset.

Prompt Engineering Prompt engineering refers to the systematic design of textual instructions that guide LLMs toward desired behaviors and outputs. Since LLMs are trained through next-token prediction over massive corpora, their performance in downstream tasks depends heavily on how queries are phrased and contextualized [26]. Early studies treated prompts as static templates; however, more recent work has shown that subtle variations in lexical framing, ordering, and contextual grounding can significantly affect accuracy and reasoning quality [27, 28]. In this thesis work, we design platform-specific structured prompts for Reddit and YouTube. Each prompt explicitly:

- defines the three allowed labels (tamper, non_tamper, uncertain),
- describes positive and negative decision rules,
- constrains the output to a strict JSON.

Few-shot prompting Few-shot prompting includes examples in the prompt, giving the model additional context which aids in boosting its performance by guiding the model in generating outputs that mirror the patterns in the examples [29]. From a theoretical standpoint, few-shot prompting bridges the gap between fully supervised fine-tuning and unsupervised inference [30]. By providing minimal yet representative examples, it allows the model to internalize labeling conventions and decision boundaries while preserving flexibility and generalization [26]. In this context, few-shot prompting is used to guide open-source LLMs toward the classification of tamper, non-tamper or uncertain content.

Chapter 3

Methodology

The development of our framework involves multiple steps. This section will discuss our methodology in data curation, data analysis and results evaluation.

3.1 Data Collection

The dataset used in this work is derived from and extends the corpus introduced by Scarano *et al.* in their study [3]. Their SOCMATI framework demonstrated that social media platforms such as **YouTube** and **Reddit** constitute rich reservoirs of open-source automotive cybersecurity intelligence (OSCTI). Following the same methodological principles, we focus on YouTube and Reddit as publicly accessible and ethically manageable sources, where posts and video metadata can be systematically collected through official APIs and web-scraping interfaces using keywords such as "def delete", "def removal", "dpf delete" and "dpf removal". The selected YouTube videos and Reddit posts were manually inspected to ensure thematic relevance. Following the ethical guidelines discussed by Scarano *et al.*, only publicly available content was accessed, and no personally identifiable information was collected. All user IDs were anonymized, and private or deleted posts were excluded. The resulting dataset therefore complies with the GDPR[31] principles of data minimization and purpose limitation.

Specifically, the dataset collected from YouTube and Reddit comprise a total of **5,551 YouTube videos** and **1,496 reddit posts**. Each platform contributes different content types and metadata:

- **YouTube**: *id, title, description, tags, transcript.*
- **Reddit**: *id, title, description, comments.*

Figure 3.1 and Figure 3.2 show the real examples from the YouTube and Reddit datasets respectively. Among the 5,551 YouTube videos collected, 2,412 entries

YouTube Data Example

```
{
"id":  "GD1vx5poql0",
"title":  "How to remove DEF INJECTOR on Ram Trucks #shorts",
"description":  "How to remove DEF INJECTOR on ram Cummins truck.  The injector is
located on the exhaust near the rear of the truck.  To remove you will need a 10mm
socket and ratchet.  Remove the diesel exhaust fluid supply line by pushing in the
line an squeezing the white clip together then pull off to remove.  The next thing
will be to use a pick to remove the red safety lock on the electrical connector.
After it is unlocked then depress the black tab to remove connector.  Proceed will
removing two 10mm bolts that hold injector to exhaust.  Clean build up with hot water.
Clean port in exhaust.  Install new gasket if needed.  Repeat the removal process in
reverse to reinstall.  Do not tighten holding bolts very tight they could break.  Use
anti seize on bolts.  Full length video on the channel.  Look for how good is peak def
for your ram truck.  @4MRanch #shorts #repairs #def",
"tags":  ["Ram trucks DEF diesel exhaust fluid injector removal how to cleaning
corroded injector 10mm bolts truck service peak blue def fluid shorts Maintenance"],
"transcript":  "last time I had it awesome mine just squeezes and Pops off like so and
then this you push in and squeeze together just like that and that comes right off
there so this clip squeeze it push in pull off so now that that’s out of the way that
actually looks good because sometimes it’ll be in the past when I looked at this it
was kind of crystallized around the supply line and if that’s happening it could be
that it’s not sealing well and you may have to replace the supply line and I’ll leave
a link in the description below to where you can pick one of those up so now we’re
going to take our 10 mil ratchet I’m going to take these bolts out right here"
}
```

Figure 3.1: A real YouTube data item from the collected dataset.

Reddit Data Example

```
{
"id":  "EDC17C56_tuning_1fmk3j3",
"title":  "EDC17C56 tuning",
"description":  "I have read a lot about diesel tuning and want to build my first own
stage for my F11 N57D30O1.  If anyone can give me directions for egr off and dpf off
(or any information), I‘d be glad!",
"comments":  [ "Download winOLS demo, try to find a definition or a similar, well
defined file and use that to define yours." ]
}
```

Figure 3.2: A real Reddit data item from the collected dataset.

included native transcripts provided by the platform. If the transcripts were unavailable, our pipeline used yt-dlp[32], a python library, to download video. Subsequently, OpenAI's Whisper-Large-v3[33], was used to generate text from those videos. After generating the transcripts, a total of 2,771 videos in the YouTube dataset contained usable transcripts data. These textual transcripts play a crucial role in downstream embedding-based analysis and LLM direct classification.

We systematically evaluated whether they could provide richer semantic context compared to titles and descriptions alone.

In addition to the structural characteristics of the dataset, the collected content spans a wide range of behaviours, intentions and technical discussions. To support the downstream analysis and classification tasks, we label each item into one of three categories: *tamper, non_tamper*, and *uncertain*. The following sections describe these three categories in detail, outlining the types of content included in each group and the rationale behind their definitions. This categorisation provides the foundation for evaluating both unsupervised clustering and LLM-based classification methods throughout this thesis.

3.1.1 Tampering Content

Items labelled as tamper include posts and videos that explicitly describe, request, or promote practices aimed at modifying or disabling a vehicle's emissions-control or safety-related functions. In both YouTube and Reddit subsets, these items frequently involve technical discussions on DPF/DEF/EGR removal, catalyst deletion, and straight-pipe conversions, as well as ECU remapping, tuning, and flashing using tools such as KESS, K-TAG, MPPS, PCMFlash, Autotuner, or WinOLS. Many entries provide procedural guidance, troubleshooting steps for failed flashes, or advice on bypassing immobilizers and security mechanisms. Others involve sharing or requesting binary files, damos, or stage maps intended to alter torque limiters, fuel/air ratios, or injection parameters. These characteristics make tamper items clear examples of illicit or regulation-avoidance modifications.

3.1.2 Non-Tampering Content

The non-tamper class contains legitimate and lawful automotive content. Across Reddit and YouTube, these items typically describe routine maintenance procedures—such as DPF cleaning (not removal), EGR valve cleaning, replacement of sensors, regeneration, wiring repairs, or common mechanical troubleshooting. Many posts focus on interpreting fault codes (e.g., P0401, P2002), identifying symptoms like misfires, limp-mode activation or smoke, or requesting advice on drivability issues. Some videos include ECU readings or diagnostics purely for inspection or educational purposes, without modifying calibration files. This category is essential for preventing false positives, as it demonstrates that emissions-related terminology does not always imply illegal modification activity.

3.1.3 Uncertain Content

Items are assigned the uncertain label when the available information is insufficient to reliably determine whether tampering is intended or performed. These cases frequently arise in Reddit discussions where users mention tuning tools or ECU access but provide no clear indication of purpose. For example, asking whether a tool is compatible, whether an ECU "needs a flash," or inquiring about general tuning concepts without specifying intent. On YouTube, uncertainty often results from incomplete descriptions, missing or noisy transcripts, or videos that refer to improved performance without clarifying whether modifications were mechanical, software-based, or purely diagnostic. This category reflects the real ambiguity present in social-media OSINT data and plays a crucial role in assessing model robustness against borderline or ambiguous cases.

3.2 Preprocessing

This section describes the text preprocessing pipeline applied to YouTube and Reddit samples prior to embedding, clustering, and LLM-based classification. Before the analysis, a data cleaning step is necessary to remove unreliable data. In our pipeline, the preprocessing stage, show at Figure 3.3, was designed with three main objectives:

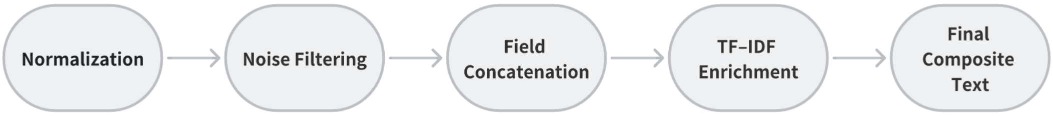


Figure 3.3: Preprocessing Step

1. **Normalization**

   Normalize heterogeneous fields (titles, descriptions, tags, transcripts, comments) into a consistent textual format.

2. **Noise Reduction**

   Several steps of data cleaning were performed here:

   - Convert all text to lowercase
   - Remove punctuation and special characters
   - Replace non-alphanumeric symbols with whitespace
   - Collapse repeated spaces into single tokens

- Remove URLs and platform artifacts (e.g., “http”, “com”, “amzn”)
- Filter unrelated lexical noise while preserving automotive terminology

All cleaned text fields corresponding to a video and post are concatenated into a single composite text entry, representing one sample.

3. **Semantic Enrichment**

   Integrates additional enrichment strategies inspired by the multi-feature preprocessing paradigm proposed by Tarekegn *et al.* [11], who demonstrated that leveraging LLMs and multi-level keyword extraction can substantially improve the representational richness and subsequent clustering robustness of large-scale textual datasets. The specific procedure was as follows: After completing basic data cleaning, we first constructed a TF-IDF model across the entire corpus to capture the most discriminative terms between documents. We then extracted the highest-weighted keywords for each text segment and appended them back to their corresponding synthesized texts. This process increased the density of contextually relevant vocabulary in the corpus, thereby enhancing the effectiveness of subsequent semantic representation and clustering analysis.

The outcome of this stage is a high-quality dataset of normalized and semantically enhanced text entries, forming the foundation for embedding-based and LLM-based detection of tampering content in the subsequent experimental phases.

3.3 Embedding and Clustering Approach

The central idea of this stage is to examine whether the semantic structure of social media content can reveal tampering-related behavior without supervision. In particular, we employed the *E5* model [16], which optimizes contrastive learning objectives for general-purpose text similarity; *Llama3* [18] and *Mistral* [19], two open-source instruction-tuned LLMs capable of capturing context-rich semantics across long documents. Following recent findings that LLM embeddings outperform shallow representations such as TF–IDF and GloVe in clustering tasks [17, 34], each document was encoded into a normalized vector of fixed dimension d using mean-pooled hidden states of the final transformer layer. The resulting embeddings were ℓ_2-normalized to ensure cosine-based similarity comparability across models.

By representing each post or video as a point in a high-dimensional vector space, conceptually similar samples are expected to occupy neighboring regions. Under this assumption, posts or videos that discuss tampering activities—such as illegal vehicle modifications—are expected to occupy a similar region of the embedding space, forming a cluster distinct from non-tampering discussions. Given the overall goal of distinguishing *tampering* from *non-tampering* content, an unsupervised clustering approach was applied, with the **number of clusters** fixed at **2** to reflect the assumed division between tampering and non-tampering content. This formulation allows the task of tampering detection to be reframed as a clustering problem: if the embedding model truly captures the semantics of tampering behavior, then the cluster assignments should implicitly recover the tamper/non-tamper categories without explicit labels. To characterise the resulting cluster structures, we compute internal clustering metrics such as the Silhouette Score, Davies–Bouldin Index, and Calinski–Harabasz Score. These metrics provide an easy to understand way to assess the capacity of embedding models to capture semantic distinctions relevant to tampering behavior without relying on labeled data.

3.4 LLM Direct Classification Approach

In addition to the embedding–clustering pipeline, a complementary approach based on LLMs was employed to perform direct classification of social-media content into the categories *tamper*, *non-tamper* and *uncertain*. Unlike the unsupervised method that infers latent clusters from embeddings, the LLM-based approach directly leverages the model's internal semantic understanding to evaluate intent, risk, and legality embedded in textual descriptions.

3.4.1 Prompt Design.

A unified three-class instruction template was developed for Reddit and YouTube samples. Each prompt contains the following elements:

- **Task definition:** a description of the classification objective with precise definitions of the three output classes (*tamper*, *non_tamper*, *uncertain*) and their decision rules.
- **Strict output constraint:** the model is required to output a single-line JSON object of the form showed in Figure 3.4.
- **Input block:** the available textual fields for each sample, namely:
  - YouTube: *title, description, tags, transcript*;
  - Reddit: *title, description, comments.*
- **Few-shot examples:** one representative annotated example for each category (tamper, non_tamper, uncertain) is appended to the instruction to stabilise the model's behaviour.
- **System prompt:** a short, high-priority directive enforcing automotive–cybersecurity expertise and JSON-only output, shows at Figure 3.5.

Figure 3.6, 3.7 shows the specific promot for reddit and YouTube data, this structured prompt ensures consistent interpretation of the task across models and reduces variability across platforms.

Example: Expected JSON Output Format

```
{"label": "tamper", "confidence": 0.95, "rationale": "The content discusses
the removal of DPF and DEF systems for power, indicating tampering with emissions
systems."}
```

Figure 3.4: Structured JSON response format required from all LLM classifiers.

System Prompt

```
You are an expert in automotive cybersecurity and emission-system compliance.
Always respond with a SINGLE LINE of valid JSON only (no markdown/prose).  If
uncertain, choose "uncertain" with lower confidence.
```

Figure 3.5: System-level instruction shared by all LLM classifiers.

3.4.2 Inference Procedure.

During inference, all models were loaded and run in 4-bit quantization. Compared to full-precision models, this setup significantly reduces memory usage, enabling more efficient processing of large-scale datasets within the same hardware environment without perceptible degradation in semantic reasoning capability. This compromise makes LLM inference feasible in resource-constrained research settings.

Model outputs uniformly adopt a structured JSON format containing three key elements: predicted label, confidence score, and a concise rationale. This standardized structure facilitates automated parsing while enhancing the transparency and auditability of decision-making processes. To ensure data quality, the system automatically filters responses that violate JSON formatting, contain missing fields, or present incomplete content, guaranteeing that all results entering the analysis phase are reliable and interpretable.

Overall, this inference pipeline achieves an optimal balance between automated efficiency and interpretability requirements. On one hand, it enables efficient batch processing of large-scale social media data; on the other, the structured output format provides clear foundations for subsequent statistical analysis.

Reddit Classification Prompt

```
Classify the Reddit post into exactly one of three classes:
- tamper:  content that demonstrates, promotes, or requests information about illegal
modification, deletion, or disabling of vehicle emission or safety systems.
Examples:  DPF/DEF/EGR delete, ECU tuning/remapping, performance modification, or the use
of enabling tools/software.
- non_tamper:  legitimate and lawful content for maintenance/cleaning/repair/ diagnostics
that does not alter or bypass emission control logic.
Examples:  DPF regeneration, EGR valve cleaning, sensor troubleshooting, ECU dump reading
for diagnostic verification only.
- uncertain:  ambiguous or incomplete cases where intent or legality cannot be confidently
determined.

Rules:
1) Any explicit intent/action/enabling info for illegal modification -> "tamper".
2) Pure maintenance/diagnostics/verification without modification intent -> "non_tamper".
3) Insufficient context or unclear intent -> "uncertain".

Return STRICT single-line JSON only:
{"label":"tamper|non_tamper|uncertain","confidence":0.xx,"reasons":"..."}

FEW-SHOT (1 tamper + 1 non_tamper + 1 uncertain)
Title:  EGR delete tuning options Description:  i have an autel mx808 j5234 tuning software
to delete EGR? Comments:  how to disable EGR via tuning tools (Autel MX808, J5234) and
remove EGR codes
EXPECTED_JSON {"label":"tamper","confidence":0.91, "reasons":"EGR delete intent"}

Title:  New to ECU repairs Description:  plan to do crash data / odometer calibration using
Iprog+ V77; looking for tool recommendations Comments:  electronics repair & diagnostics;
no performance tuning or emission bypass
EXPECTED_JSON {"label":"non_tamper","confidence":0.85,"reasons":"no modification intent"}

Title:  Mercedes-Benz CAN ID codes Description:  paid coding services to unlock features;
want correct codes Comments:  mentions DTS Monaco/Vediamo; unclear whether emission/safety
systems targeted
EXPECTED_JSON {"label":"uncertain","confidence":0.68,"reasons":"no clear tamper"}
```

Figure 3.6: Instruction and few-shot structure used for Reddit post classification.

3.5 Evaluation

The evaluation phase of this thesis work aims to assess whether embedding-based clustering approach or LLMs can reliably identify tampering-related content in social media posts and videos, and to determine which approach is best suited for this highly domain-specific task. To this end, we examine two distinct evaluation strategies: one is a multi-dimensional scoring framework 3.5.1 based on prior work in harmful content assessment, and the other is a direct three-class classification method 3.5.2 utilizing OpenAI's GPT-4o-mini model. These complementary perspectives enable us not only to investigate the capability of embedding-based approach and LLMs in detecting illicit modification behaviors but also to explore

YouTube Classification Prompt

```
Classify the YouTube video into exactly one of three classes:
- tamper:  content that explicitly shows/promotes/describes illegal modification, tuning,
or disabling of vehicle emission or safety systems.
Includes DPF/DEF/EGR delete, ECU remapping/tuning, enabling tools (KESS, K-TAG, FoxFlash,
PCMFlash, CMDFlash, WinOLS, etc.)  or modes (OBD, bench, boot, BDM, Tricore, VR file,
checksum).
- non_tamper:  legitimate maintenance/diagnostic/educational content for
cleaning/repair/verification without changing emission-control logic.
Examples:  DPF regeneration, EGR cleaning, DEF pressure troubleshooting, ECU dump reading
for fault inspection.
- uncertain:  ambiguous/incomplete cases where intent or legality cannot be confidently
determined.

Rules:
1) Any modification intent or enabling procedure -> "tamper".
2) Only maintenance/cleaning/diagnostics -> "non_tamper".
3) Intent/legality unclear -> "uncertain".

Return STRICT single-line JSON only:
{"label":"tamper|non_tamper|uncertain","confidence":0.xx,"reasons":"..."}

FEW-SHOT (1 tamper + 1 non_tamper + 1 uncertain)
Title:  2020 Cummins EGR Delete Description:  removed EGR cooler and installed block-off
plates; power increase Transcript:  describes removing EGR cooler and installing plates
EXPECTED_JSON {"label":"tamper","confidence":0.93,"reasons":"explicit EGR delete"}

Title:  Jeep 3.0L ECODIESEL - What I've Learned Description:  DEF & DPF usage; soot
levels & regeneration behavior Transcript:  monitoring DEF level and DPF regeneration;
legal maintenance EXPECTED_JSON {"label":"non_tamper","confidence":0.87,"reasons":"no
modification"}

Title:  ORRP Part 1 - An Intro to Opioid Rapid Response Program Description:
unrelated podcast; 'DPF' appears as another acronym EXPECTED_JSON
{"label":"uncertain","confidence":0.55,"reasons":"keyword mismatch"}
```

Figure 3.7: Instruction and few-shot structure used for YouTube video classification.

the limitations of structured, rule-based scoring methods when applied to complex automotive cybersecurity discussions.

3.5.1 Multi-Dimensional Scoring Framework

We initially adopted a dimensional scoring architecture inspired by the framework proposed by Huq and Suleiman [25]. Their framework introduced an LLM-based content evaluation pipeline for detecting and scoring harmful content on YouTube. This approach relies on defining semantic dimensions to guide the LLM in evaluating each dimension separately, followed by rule-based aggregation to combine scores and derive final content labels. Therefore, building upon this idea, rather than relying solely on discrete labels such as tamper or non-tamper, the framework decomposes each post or video into several interpretable dimensions, each reflecting a specific aspect of tampering intent, technical depth, or behavioural risk.

The following subsections describe the implementation of this evaluation framework.

Dimension Design

The primary goal of designing dimensions for our thesis work is to understand the intensity of various metrics of a particular content from post or video. Eventually, we designed seven dimensions:

- **Tampering Guidance**

  Evidence that the content provides actionable instructions, step-by-step procedures, tool recommendations, or parameter settings that directly enable emissions-system deletion, ECU remapping, or the bypassing of regulatory or safety controls. This dimension captures the operational aspect of tampering.

- **Tampering Claim**

  Indicates whether the user explicitly states they have performed, are performing, or intend to perform tampering activities. This includes self-reports of DPF/DEF/EGR deletion, tuning results, or statements indicating prior success or failure. It reflects self-disclosed engagement in illegal modification.

- **Illegal Modification**

  Assesses the extent to which the content involves or encourages activities that circumvent emissions regulations, safety requirements, or vehicle control logic. This dimension includes references to disabling emission-control systems, using prohibited tools, defeating diagnostic protections, or bypassing compliance mechanisms.

- **Manipulative Intent**

  Captures whether the content actively encourages others to carry out tampering, promotes unsafe shortcuts, or adopts a tone that trivialises or normalises illegal modifications. This reflects a social influence dimension—whether the user is pushing others toward harmful or unlawful actions.

- **Safety Risk**

  Evaluates the degree to which the actions described—if followed—could jeopardise vehicle integrity, operational safety, or road-user wellbeing. Examples include disabling safety-critical subsystems, altering engine behaviour unpredictably, or recommending operations that compromise braking, stability, or emissions control. This captures the physical hazard associated with tampering.

- **Misinformation**

  Identifies the presence of technically inaccurate, misleading, or oversimplified claims that could cause users to behave unsafely or engage in illegal modifications under false assumptions. This reflects the risk that misleading information directly contributes to tampering behaviour.

- **Social Harm**

  Measures the potential for the described behaviour to influence broader adoption of unsafe or unlawful modifications. This includes content that normalises emissions-system deletion, encourages bypassing inspections, or demonstrates tampering techniques in a way that could propagate harmful practices within the community.

Each dimension is designed to be independent, enabling the evaluator to analyse different behavioural signals without forcing them into a single categorical judgment.

Mapping Rules

To convert the seven dimension scores derived from GPT-4o-mini into a final binary decision (tamper vs non-tamper), the framework applies a set of deterministic mapping rules. These rules combine strong, weak, and aggregated evidence to determine whether a piece of content exhibits tampering-related behaviour.

1. **Hard Triggers**

   Certain dimensions serve as direct indicators of tampering when their score exceeds a fixed threshold. In these cases, strong evidence in a single dimension is sufficient to classify the content as tampering-related. The thresholds are:

   - Tampering Guidance ≥ 0.66
   - Tampering Claim ≥ 0.66
   - Illegal Modification ≥ 0.66
   - Safety Risk ≥ 0.80

   If any of these criteria are met, the system immediately flags the content as tamper without requiring further evidence. These dimensions represent behaviours that are inherently high-risk or strongly indicative of illegal modification intent.

2. **Soft Evidence Pairs** In some cases, weaker signals in multiple dimensions jointly indicate tampering even if none of the individual scores pass a hard trigger threshold. To capture this, the framework defines several soft-evidence pairs, where moderate evidence in both dimensions suggests tampering. The pairs are:

   - Illegal Modification ≥ 0.33 and Safety Risk ≥ 0.33
   - Tampering Claim ≥ 0.33 and Manipulative Language ≥ 0.33
   - Misinformation ≥ 0.50 and Safety Risk ≥ 0.33

   If any pair meets its respective thresholds, the content is classified as tampering. This rule models the idea that certain behaviours, when combined, express stronger tampering intent than when considered independently.

3. **Global Mean Threshold** When neither hard triggers nor soft pairs apply, the decision falls back on a global aggregation rule. A post or video is classified as tamper if:

   - the mean score across all seven dimensions is ≥ 0.45,
   - and at least three dimensions have non-zero evidence.

   This rule captures diffuse or mild tampering signals distributed across multiple dimensions, ensuring that borderline cases are still meaningfully captured.

Prompt-Based Scoring Procedure

To assign dimension scores at scale, the framework uses a structured prompt applied to GPT-4o-mini. The prompt instructs the model to:

1. Read all available fields of the content (title, description, tags, transcript, comments).

2. Provide a short rationale explaining the presence or absence of behaviours related to the seven dimensions.

3. Output a strict JSON object with numerical scores for each dimension based on the mapping rules.

Figure 3.8 is a representative example of the scoring prompt.

YouTube Multi-Dimensional Scoring Prompt

```
You are a security analyst for automotive cybersecurity.
Task:  Evaluate the following user-generated content
(title/description/tags/transcript/comments) and score each dimension from 0.00 to 1.00.
Definitions (tamper context):
- tamper_guidance:  captures content with step-by-step "how to do it" instructions.
- tamper_claim:  content claims tampering was performed or is being demonstrated.
- illegal_mod:  indicates evasion of inspections/emissions/registration/safety compliance.
- manipulative_language:  persuasive tone or intent encouraging others to tamper (not
necessarily steps).
- safety_risk:  non-trivial mechanical or operational risk if the content is followed.
- misinfo:  technically wrong yet plausible guidance likely to cause dangerous attempts.
- social_harm:  potential for broader harm (coordination, scale, influence on inexperienced
users).

Output STRICT JSON with this schema and nothing else:
{
"scores":  {
"tamper_guidance":  0.00,
"tamper_claim":  0.00,
"illegal_mod":  0.00,
"manipulative_language":  0.00,
"safety_risk":  0.00,
"misinfo":  0.00,
"social_harm":  0.00
},
"rationale":  "one concise sentence"
}

Content:
```

Figure 3.8: Instruction structure used for multi-dimensional scoring of YouTube content.

3.5.2 OpenAI direct classification

The purpose of this part is to generate a consistent, deterministic set of labels for all items in the dataset, enabling a uniform reference dataset for downstream validation work. The overall structure of the task is conceptually similar to the LLM direct classification procedure described earlier, but implemented with OpenAI's GPT-4o-mini through the OpenAI Chat Completion API. These output labels are later used within the evaluation pipeline to analyse classifier behaviour and assess alignment between the open-source models and the reference classifier.

3.5.3 Golden Dataset

To demonstrate the reliability of the two evaluation methods (3.5.1, 3.5.2), we constructed a verified *golden dataset* manually by sampling 100 posts and 100 videos from both Reddit and YouTube sources. All selected entries explicitly reference

Diesel Particulate Filter (DPF), Diesel Exhaust Fluid (DEF), or Exhaust Gas Recirculation (EGR) systems, covering a diverse range of maintenance, diagnostic, and modification scenarios. Each item was carefully reviewed and annotated according to a three-class taxonomy:

- **tamper**: content that demonstrates, promotes, or discusses illegal modifications such as DPF/DEF/EGR deletion, ECUs remapping, performance tuning, or emission-control bypass;
- **non-tamper**: legitimate operations related to system maintenance, cleaning, regeneration, troubleshooting, or diagnostic data analysis (e.g., ECUs dump inspection) that do not alter the emission-control logic;
- **uncertain**: ambiguous or incomplete cases in which the intention or legality of the activity cannot be confidently determined.

This curated dataset serves as the ground-truth reference for validating the accuracy and robustness of the two evaluation methods. For both the multi-dimensional scoring framework and the OpenAI direct-classification model, we first evaluate their predictions on the golden dataset by measuring three key thresholds:

- overall accuracy (0.8)
- tamper precision (0.8)
- and tamper recall (0.8)

Only methods that satisfy these criteria are then applied to the full-scale Reddit and YouTube datasets.

Chapter 4

Results

This section presents the experimental findings obtained from the embedding-based clustering analysis, multi-dimensional scoring framework, and the direct LLM classification experiments.

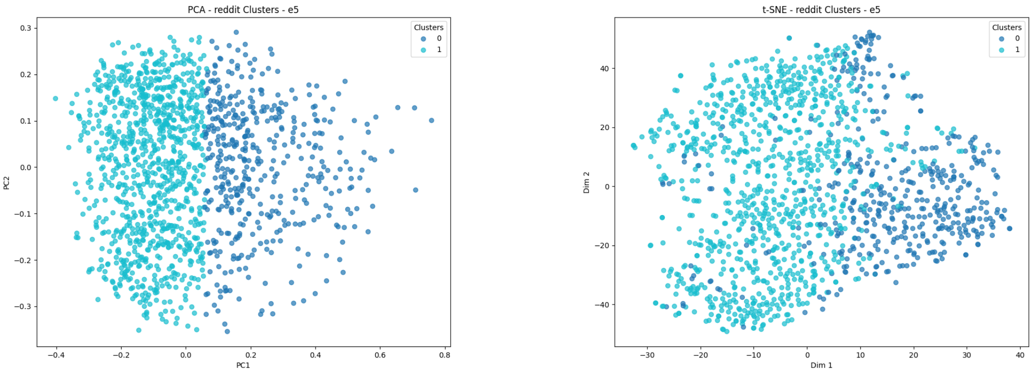
4.1 Embedding-Based Clustering

4.1.1 Reddit

Table 4.1 reports quantitative clustering metrics for the Reddit dataset using three embedding models: Meta-Llama-3-8B-Instruct, Mistral-7B-Instruct-v0.2, and Qwen2.5-7B-Instruct. All models yield low silhouette scores (0.10–0.11), indicating weak intra-cluster cohesion and substantial overlap between the two inferred clusters. Similarly, the Davies-Bouldin index (approximately 3.0) reveals high within-cluster variance and poor inter-cluster separation, while Calinski-Harabasz scores (129–133) further confirm the absence of well-defined structure in the embedding space.

Under our initial hypothesis, tampering-related and non-tampering content should form two reasonably separable regions in the embedding space due to their semantic differences in definition. However, the consistently poor clustering metrics indicate that such separation does not emerge in practice. This suggests that unsupervised partitioning based solely on embeddings is not well-suited for distinguishing between tampering and non-tampering content in this context.

The following sections present the PCA and t-SNE visualisations generated from the three embedding models, providing a qualitative view that complements the quantitative results and further illustrates the lack of separable structure in the embedding space.



(a) PCA projection (E5)

(b) t-SNE projection (E5)

Figure 4.1: Reddit clusters visualized using PCA and t-SNE for the E5 embedding model.

E5 Embeddings

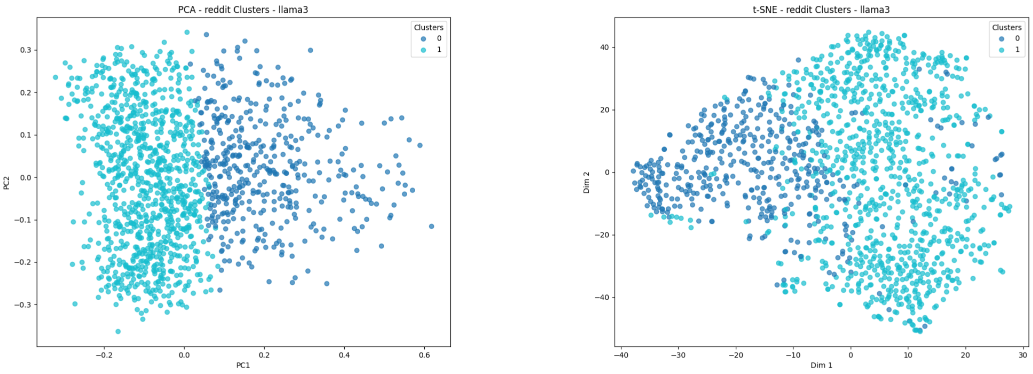
The PCA and t-SNE projections of E5 embeddings (Figures 4.1a and 4.1b) display an approximately continuous point cloud with only minor local density variations. Neither PCA nor t-SNE projections revealed separable regions. This observation is consistent with the low Silhouette Score (0.102) and indicates that the E5 embedding space does not exhibit discriminative ability for separating tampering and non-tampering content on Reddit.

Llama3 Embeddings

Llama3 projections (Figures 4.2a and 4.2b) show that both clusters still occupy overlapping areas without clear geometric separation. Although Llama3 achieves the lowest Davies–Bouldin Index (3.015), the visual patterns confirm that this marginal difference does not translate to practical separability.

Mistral Embeddings

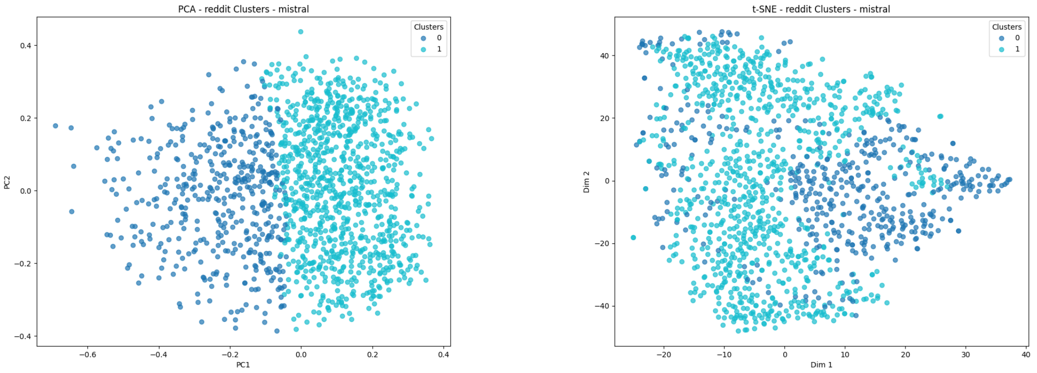
Mistral projections (Figures 4.3a and 4.3b) show distributions similar to the previous models. Although Mistral yields the highest Silhouette Score (0.109), the visualizations do not reveal any identifiable cluster boundaries.



(a) PCA projection (Llama3)

(b) t-SNE projection (Llama3)

Figure 4.2: Reddit clusters visualized using PCA and t-SNE for the Llama3 embedding model.



(a) PCA projection (Mistral)

(b) t-SNE projection (Mistral)

Figure 4.3: Reddit clusters visualized using PCA and t-SNE (Mistral embedding).

4.1.2 Youtube

Tables 4.2 and 4.3 summarize the clustering metrics for YouTube videos under two input configurations: using only title, description and tags, and using the same fields augmented with automatically extracted transcripts. In both settings, silhouette scores remain modest (0.09–0.17), while the Davies–Bouldin indices (2.2–3.1) indicate that the two clusters are far from well separated. The Calinski–Harabasz values (approximately 600–930 without transcripts and 280–440 with transcripts) suggest the presence of only weak global structure, with none of the

Table 4.1: Clustering metrics for Reddit embeddings using E5, Llama3, and Mistral.

Model	Silhouette	Davies–Bouldin	Calinski–Harabasz
E5-7b	0.1022	3.0862	129.62
Llama3-8B	0.1051	3.0154	133.41
Mistral-7B	0.1094	3.1055	130.21

Table 4.2: Clustering metrics on YouTube (without transcripts).

Model	Silhouette	Davies–Bouldin	Calinski–Harabasz
E5-7b	0.1212	2.5053	812.64
Llama3-8B	0.1452	2.1609	929.41
Mistral-7B	0.0947	2.9105	605.14

embedding models yielding a clearly superior partition of tamper versus non-tamper content.

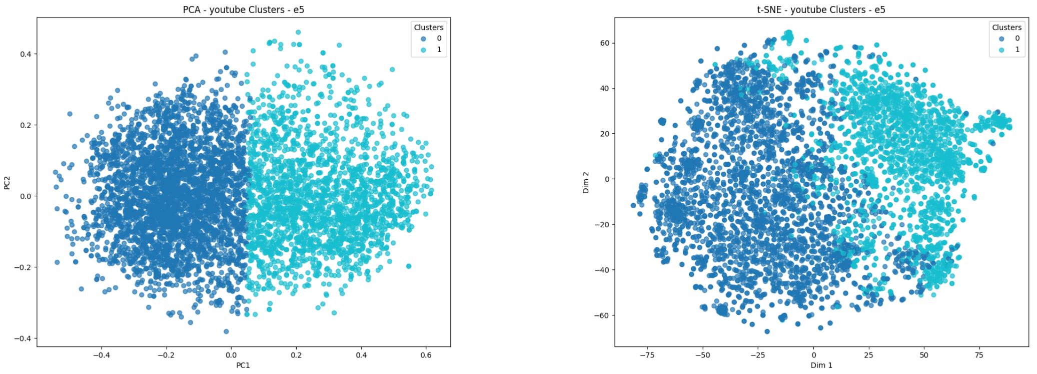
The PCA and t-SNE projections in Figures 4.7, 4.8 and 4.9 (shown for the configuration with transcripts) visually corroborate these quantitative findings. For all three models, the two clusters appear as highly overlapping clouds in the two-dimensional projections, with only a mild shift along the first principal component or t-SNE axis. Adding transcript information slightly sharpens this separation for LLama3 and Mistral, which is consistent with the slight increase in silhouette scores. However, substantial intermixing between the clusters persists. In the configuration without transcripts, the PCA and t-SNE plots shown in Figures 4.4, 4.5 and 4.6 exhibit the same qualitative pattern: all three embedding models generate broad, overlapping point clouds with no indication of two naturally separable regions corresponding to tamper and non-tamper content. The absence of transcripts does not materially alter the geometry of the embedding space, and the clusters remain intermixed across all models.

4.1.3 Summary

Overall, the results above show that unsupervised embedding-based clustering is not working on distinguishing tampering and non-tampering content in this context. Across both Reddit and YouTube, the embedding spaces produced by all tested models form largely continuous and overlapping point distributions, with no evidence of naturally separable semantic regions. Even when we enriched the YouTube data by incorporating video transcripts, the resulting clusters remained

Table 4.3: Clustering metrics on YouTube (with transcripts).

Model	Silhouette	Davies–Bouldin	Calinski–Harabasz
E5-7b	0.1189	3.0490	285.81
Llama3-8B	0.1503	2.4451	439.99
Mistral-7B	0.1663	2.6224	374.80



(a) PCA projection (E5)

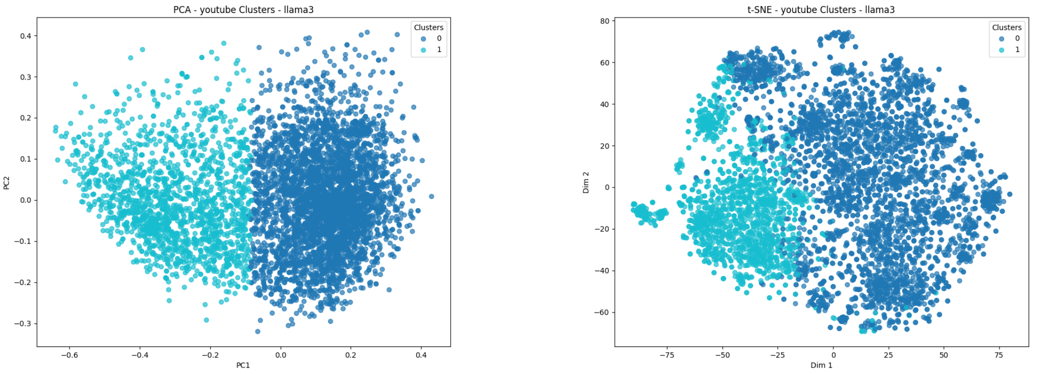
(b) t-SNE projection (E5)

Figure 4.4: YouTube clusters visualized using PCA and t-SNE for the E5 embedding model (without transcripts).

highly intermixed, failing to form the two distinct groups we anticipated. Overall, these findings indicate that the combination of general-purpose sentence embeddings and unsupervised clustering cannot effectively capture the nuanced patterns required for reliable tampering detection.

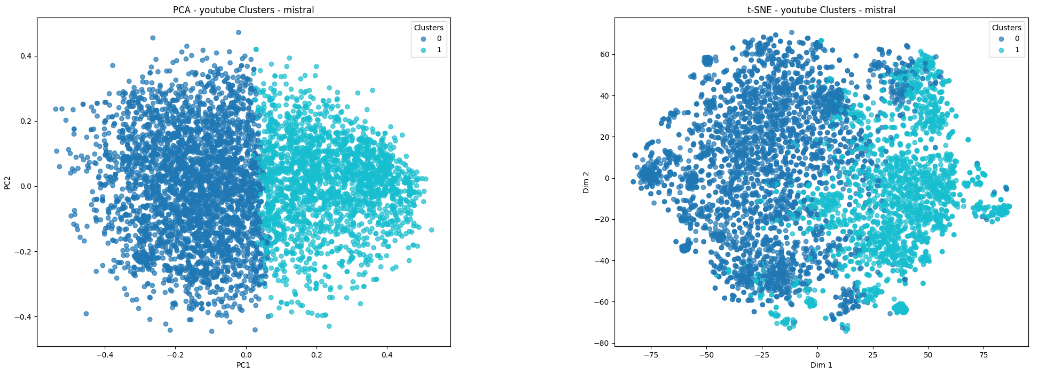
4.2 Evaluation of Validation Mechanisms

Before validating the results of open-source LLM direct classification, we first evaluated the reliability of our verification mechanisms. Both approaches were first evaluated on the manually curated golden dataset. Only if a method satisfied the predefined thresholds for overall accuracy, tamper precision, and tamper recall would it be adopted for validating open-source LLM outputs at scale.



(a) PCA projection (Llama3) **(b)** t-SNE projection (Llama3)

Figure 4.5: YouTube clusters visualized using PCA and t-SNE for the Llama3 embedding model (without transcripts).



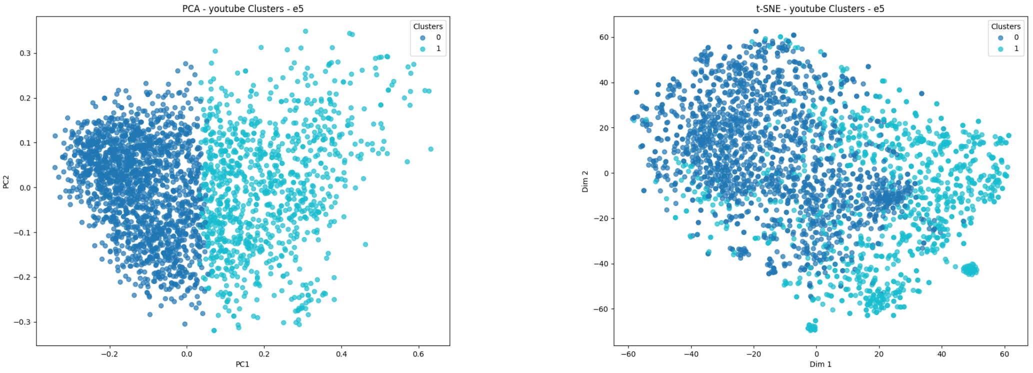
(a) PCA projection (Mistral) **(b)** t-SNE projection (Mistral)

Figure 4.6: YouTube clusters visualized using PCA and t-SNE for the Mistral embedding model (without transcripts).

4.2.1 Multi-Dimensional Scoring Framework

Table 4.4 reports the performance of the multi-dimensional scoring framework on the golden dataset. The results show that the framework failed to meet acceptable thresholds across all three metrics for both Reddit and YouTube data.

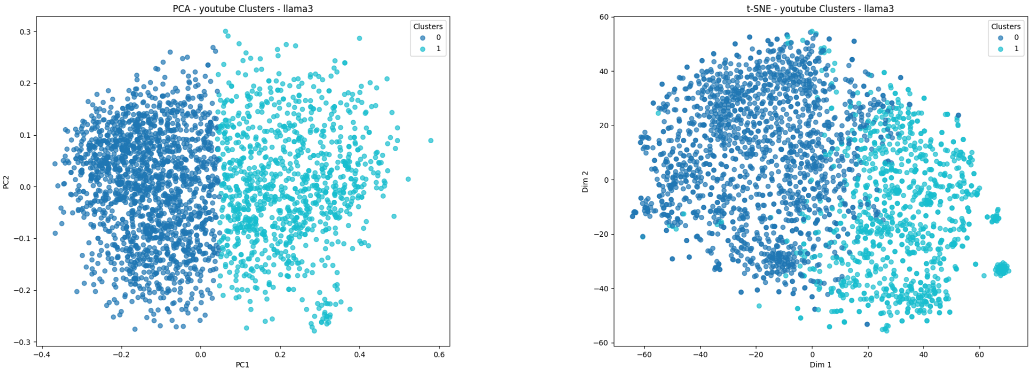
Upon closer examination of the framework's outputs, we observed that some items manually annotated as tamper—including explicit inquiries about DPF or EGR deletion procedures—received uniform scores of 0.0 across several critical



(a) PCA projection (E5)

(b) t-SNE projection (E5)

Figure 4.7: YouTube clusters visualized using PCA and t-SNE for the E5 embedding model (with transcripts).



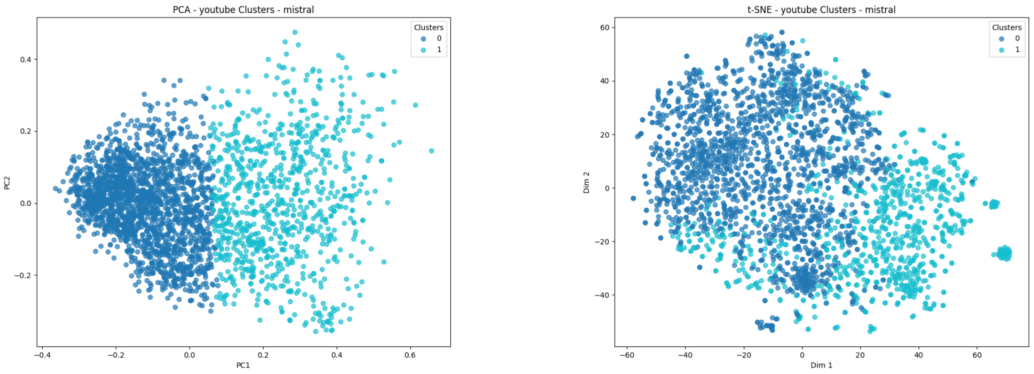
(a) PCA projection (Llama3)

(b) t-SNE projection (Llama3)

Figure 4.8: YouTube clusters visualized using PCA and t-SNE for the Llama3 embedding model (with transcripts).

semantic dimensions such as tamper_guidance, tamper_claim, and illegal_mod. In these cases, despite clear illegal modification intent, the model failed to elevate scores in any dimension.

Consequently, this method was neither applied to the full dataset nor used to evaluate the direct classification results from LLMs, as its outputs proved too unreliable for validation purposes.



(a) PCA projection (Mistral)

(b) t-SNE projection (Mistral)

Figure 4.9: YouTube clusters visualized using PCA and t-SNE for the Mistral embedding model (with transcripts).

Table 4.4: Performance of the multi-dimensional scoring framework on golden dataset.

Data	Accuracy	Tamper Precision	Tamper Recall
Reddit	0.77	0.68	0.69
YouTube	0.72	0.48	0.53

4.2.2 OpenAI Direct Classification

We next evaluated whether OpenAI's GPT-4o-mini model could serve as a more robust automatic validator. To ensure methodological consistency, the model was queried using the same instruction template employed in the open-source LLM direct-classification experiments, producing a predicted label for each item in the golden dataset.

As shown in Table 4.5, although GPT-4o-mini demonstrated stronger performance than the multi-dimensional scoring framework, it still failed to meet the predefined thresholds required for dependable validation. In particular, its precision–recall balance remained insufficient, preventing the model from serving as a stable substitute for human supervision.

4.3 Direct LLM Classification

Since neither the multi-dimensional scoring framework nor the OpenAI-based classifier met the reliability thresholds established in Section 4.2, no automated

Table 4.5: OpenAI GPT-4o-mini performance on the golden dataset.

Dataset	**Accuracy**	**Tamper Precision**	**Tamper Recall**
Reddit	0.53	0.58	0.98
YouTube	0.72	0.65	0.52

Table 4.6: Distribution of Predicted Labels by Open-Source LLMs across Full Reddit and YouTube Datasets.

Dataset	Setting	Model	non_tamper	tamper	uncertain
Reddit	–	Llama3	696	363	437
		Qwen	1084	248	164
		Mistral	552	183	761
YouTube	No transcript	Llama3	3559	1341	650
		Qwen	4044	1104	402
		Mistral	3195	1011	1344
	With transcript	Llama3	1936	618	217
		Qwen	2157	518	96
		Mistral	1929	581	261

Average confidence: Reddit (Llama3: 0.88, Qwen: 0.85, Mistral: 0.77), YouTube–No transcript (Llama3: 0.897, Qwen: 0.885, Mistral: 0.88), YouTube–With transcript (Llama3: 0.905, Qwen: 0.908, Mistral: 0.93).

method could be adopted as a validation mechanism for full-dataset evaluation. Consequently, in this section we focus on analysing the behaviour of open-source LLMs directly on the golden dataset. The goal is not to validate them automatically, but rather to understand how these models respond to tampering-related content and to identify any distinctive strengths or limitations they exhibit in detecting illicit modification behaviours.

Table 4.6 provides an overview of the prediction results, and the following subsections report each open-source LLM's prediction performance on the manually annotated golden dataset.

Reddit Table 4.7 summarises the performance of the three open-source LLMs on the Reddit portion of the golden dataset. Overall accuracy is very low (0.37–0.49), confirming that none of the models can be used as reliable stand-alone classifiers.

Specifically, both Llama3 and Qwen exhibit relatively strong sensitivity to tampering content. Llama3 achieves a high tamper precision (0.781) and a reasonably

Table 4.7: Performance of open-source LLMs on the Reddit golden dataset.

Model	**Accuracy**	**Tamper Precision**	**Tamper Recall**
Llama3-8B-Instruct	0.49	0.781	0.694
Mistral-7B-Instruct	0.46	0.739	0.472
Qwen2.5-7B-Instruct	0.37	0.722	0.722

Table 4.8: Performance of open-source LLMs on the YouTube (no transcript) golden dataset.

Model	**Accuracy**	**Tamper Precision**	**Tamper Recall**
Llama3-8B-Instruct	0.679	0.750	0.600
Qwen2.5-7B-Instruct	0.670	0.682	0.500
Mistral-7B-Instruct	0.547	0.591	0.433

high tamper recall (0.694), meaning that it identifies most tampering cases while keeping false alarms at a moderate level. Qwen produces an even higher tamper recall (0.722), which indicates that it misses fewer true tampering posts, although its overall accuracy is lower. In contrast, Mistral is the most conservative model, with the lowest tamper recall (0.472), often labelling tampering posts as non-tamper or uncertain.

YouTube Table 4.8 reports the performance of the three open-source LLMs on the YouTube golden dataset without transcripts and Table 4.9 reports the results with transcripts. Compared to the no-transcript setting, the models show positive changes in both accuracy and tampering-related performance, reflecting the influence of additional semantic context provided by transcripts. These results from YouTube platform show noticeable ability to detect tampering-related content.

Table 4.9: Performance of open-source LLMs on the YouTube (with transcript) golden dataset.

Model	**Accuracy**	**Tamper Precision**	**Tamper Recall**
Llama3-8B-Instruct	0.618	0.917	0.458
Qwen2.5-7B-Instruct	0.636	1.000	0.458
Mistral-7B-Instruct	0.600	0.800	0.500

Chapter 5

Discussion

5.1 Comment on the results

The embedding-based clustering results using E5, Llama3, Mistral show that tampering and non-tampering contents do not form separable clusters. This applies to both Reddit and YouTube. Analysis of PCA/t-SNE visualizations combined with clustering results reveals that discussions about illegal modifications and those concerning legitimate maintenance or troubleshooting are heavily interwoven in the semantic space. The keyword analysis derived from clustering results indicates that this semantic overlap occurs because both categories frequently employ identical technical terminology—such as "DPF," "EGR," "regeneration cycles," "sensor readings," or "ECU data. However, this kind of distinction between legal and illegal depends on contextual nuances and intent, which unsupervised embedding models cannot adequately capture. Consequently, even with additional textual inputs like video transcripts, unsupervised clustering methods remain unsuitable for this task.

In contrast to the limitations observed with unsupervised clustering, the direct LLM classification experiments demonstrate that instruction-tuned models are capable of distinguishing, to some extent, between tampering, non-tampering, and uncertain content at the semantic level. Unlike embedding-based methods, which rely solely on distributional similarity, LLMs incorporate the surrounding intent, action type, and narrative context when interpreting a post or video. The prediction distributions reveal clear differences in how each model interprets and classifies the content. As shown in Tables 4.6, the prediction distributions reveal clear differences in how each model interprets and classifies the content. Llama3 exhibits a comparatively balanced prediction pattern, Qwen makes more decisive assignments with fewer uncertain outputs, and Mistral produces the largest share of uncertain labels, reflecting a more cautious attitude. Despite these differences, all models consistently show greater sensitivity to tampering-related signals than to non-tampering cases, with tamper recall generally outperforming recall in the other categories. This behaviour is also reflected in the prediction distributions across the full datasets for Reddit and YouTube. Our analysis further found that YouTube transcripts provided additional contextual information to LLMs. This enhancement led to improved a notable reduction in "uncertain" predictions. In contrast, the same transcripts did not benefit embedding-based clustering approaches. The inherent noise in automatically generated captions actually introduced additional instability to the embedding space, further compromising clustering performance.

5.2 Limitations

A fundamental limitation of the current framework lies in its strong reliance on textual information, neglecting multimodal signals. Manual inspection of YouTube samples reveals that many tampering-related videos convey critical information primarily through visual content—such as screen recordings, diagnostic tool interfaces, ECU editing software, or specific procedural demonstrations—rather than verbal explanations. In numerous videos, background music, poor audio quality, or complete absence of speech make it difficult for transcription systems like Whisper to extract useful text. Consequently, text-only analysis pipelines inevitably miss crucial visual evidence, including diagnostic tool operations, usage of DPF deletion or ECU editing software. These visual cues often determine whether content constitutes tampering, yet remain entirely invisible to text-only models.

Even when transcripts are available, their quality is often inconsistent. Automatically generated subtitles frequently contain noise, timing misalignments, and semantic drift—issues particularly pronounced in technical contexts involving complex engine components and diagnostic terminology. Background music, echo, multiple speakers, and varied accents further degrade transcription quality, and these noise artifacts propagate to both embedding extraction and LLM classification, undermining overall performance.

Finally, the evaluation framework itself has methodological limitations. While the open-source models in this study demonstrate emergent capability in identifying tampering-related content, they remain insufficient for reliable automated classification. Subsequent work could explore targeted prompt engineering or task-specific prompt optimization to enhance classification performance without requiring full model retraining, thereby improving detection reliability for automotive cybersecurity applications.

Chapter 6

Conclusion

In this thesis, we investigate the feasibility of using open-source embedding models and instruction-tuned LLMs to identify automotive tampering-related content in social media environments. As OSINT gains increasing importance in automotive cybersecurity, we focus our analysis on Reddit and YouTube platforms where both legitimate maintenance discussions and illicit modification practices are openly shared.

We evaluated two methodological approaches: unsupervised embedding-based clustering and direct LLM classification. Additionally, we explored the use of a multi-dimensional scoring framework and OpenAI models as potential automated validation mechanisms to assess whether these methods could provide reliable performance indicators before large-scale inference. In order to verify reliability of these approaches, we constructed a manually validated gold-standard dataset centered around key components such as EGR, DPF, and DEF systems.

Our evaluation incorporated clustering metrics, dimensionality reduction visualizations, and classification results across multiple datasets and configurations. The findings reveal that tampering and non-tampering content do not form semantically separable structures in embedding space, rendering clustering methods unsuitable for this task. In contrast, direct LLM classification, while not yet reliable enough for automated validation, demonstrated meaningful sensitivity to tampering-related semantics and produced more interpretable outputs than unsupervised approaches. Although the validation mechanisms tested in this study are not yet ready for deployment, experimental trends suggest that refining prompt design and incorporating multimodal inputs based on LLM semantic reasoning remain promising directions for future OSINT applications in automotive cybersecurity.

Bibliography

[1] Charlie Miller and Chris Valasek. «Adventures in Automotive Networks and Control Units». In: *Proceedings of DEF CON 21.* Available at: `https://illmatics.com/car_hacking.pdf`. Las Vegas, NV, USA, 2013, pp. 260–264 (cit. on pp. 1, 2, 5).

[2] Backlinko. *YouTube Stats: How Many People Use YouTube in 2024?* [Online; accessed 3-Nov-2025]. Mar. 2023. URL: `https://backlinko.com/youtube-users` (cit. on p. 2).

[3] Nicola Scarano, Luca Mannella, Alessandro Savino, and Stefano Di Carlo. «Can Social Media Shape the Security of Next-Generation Connected Vehicles?» In: *2024 IEEE International Conference on Dependable Systems and Networks (DSN).* Turin, Italy: IEEE, 2024, pp. 1–9 (cit. on pp. 2, 5, 6, 10).

[4] Stephen Checkoway et al. «Comprehensive Experimental Analyses of Automotive Attack Surfaces». In: *Proceedings of the 20th USENIX Conference on Security.* SEC'11. San Francisco, CA: USENIX Association, 2011, p. 6 (cit. on p. 5).

[5] R. Fayyazi and S. J. Yang. «On the Uses of Large Language Models to Interpret Ambiguous Cyberattack Descriptions». In: *arXiv preprint arXiv:2306.14062* (2023). [Online; accessed 3-Nov-2025]. URL: `http://arxiv.org/abs/2306.14062` (cit. on pp. 5, 8).

[6] ISO/SAE. *ISO/SAE 21434: Road Vehicles – Cybersecurity Engineering.* International Standard. [Online]. Available: `https://www.iso.org/standard/70918.html`. Geneva, Switzerland and Warrendale, PA, USA: International Organization for Standardization and SAE International, 2021 (cit. on p. 5).

[7] Karl Koscher et al. «Experimental Security Analysis of a Modern Automobile». In: *2010 IEEE Symposium on Security and Privacy.* IEEE. 2010, pp. 447–462. DOI: `10.1109/SP.2010.34` (cit. on p. 5).

[8] Yong-Woon Hwang, Im-Yeong Lee, Hwankuk Kim, Hyejung Lee, and Donghyun Kim. «Current Status and Security Trend of OSINT». In: *Security and Communication Networks* 2022 (2022), pp. 1–15. DOI: `10.1155/2022/1290129` (cit. on p. 6).

[9] Peng Gao, Xiaoyuan Liu, Edward Choi, Sibo Ma, Xinyu Yang, and Dawn Song. «ThreatKG: An AI-Powered System for Automated Open-Source Cyber Threat Intelligence Gathering and Management». In: *arXiv preprint arXiv:2212.10388* (2022). URL: `https://arxiv.org/abs/2212.10388` (cit. on p. 6).

[10] Anton O. Bryushinin, Alexandr V. Dushkin, and Maxim A. Melshiyan. «Automation of the Information Collection Process by Osint Methods for Penetration Testing During Information Security Audit». In: *2022 Conference of Russian Young Researchers in Electrical and Electronic Engineering (ElConRus)*. 2022, pp. 242–246. DOI: `10.1109/ElConRus54750.2022.9755812` (cit. on p. 6).

[11] Adane Nega Tarekegn. «Large Language Model Enhanced Clustering for News Event Detection». In: *IEEE Access* (2024). Available online. URL: `https://arxiv.org/abs/2406.10552` (cit. on pp. 6, 7, 14).

[12] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. «Efficient Estimation of Word Representations in Vector Space». In: *arXiv preprint arXiv:1301.3781* (2013). [Online; accessed 6-Nov-2025]. URL: `https://arxiv.org/abs/1301.3781` (cit. on p. 7).

[13] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. «Attention Is All You Need». In: *Proceedings of the 31st Conference on Neural Information Processing Systems (NeurIPS 2017)*. Long Beach, CA, USA: Curran Associates, Inc., 2017, pp. 6000–6010 (cit. on pp. 7, 8).

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. «BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding». In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*. Minneapolis, MN, USA: Association for Computational Linguistics, 2019, pp. 4171–4186. URL: `https://aclanthology.org/N19-1423` (cit. on p. 7).

[15] Nils Reimers and Iryna Gurevych. «Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks». In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Hong Kong, China: Association for Computational Linguistics, 2019, pp. 3982–3992. URL: `https://aclanthology.org/D19-1410` (cit. on p. 7).

[16] Yang Wang et al. «Text Embeddings by Weakly-Supervised Contrastive Pretraining». In: *arXiv preprint arXiv:2212.03533*. 2022. URL: `https://arxiv.org/abs/2212.03533` (cit. on pp. 7, 14).

[17] Nuno Fachada Alina Petukhova João P. Matos-Carvalho. «Text Clustering with LLM Embeddings». In: *arXiv preprint arXiv:2403.15112* (2024). URL: `https://arxiv.org/abs/2403.15112` (cit. on pp. 7, 14).

[18] Hugo Touvron et al. «LLaMA 2: Open Foundation and Fine-Tuned Chat Models». In: *Meta AI Technical Report* (2023). URL: `https://arxiv.org/abs/2307.09288` (cit. on pp. 7, 8, 14).

[19] Albert Q. Jiang et al. «Mistral 7B». In: *arXiv preprint arXiv:2310.06825* (2023). URL: `https://arxiv.org/abs/2310.06825` (cit. on pp. 7, 8, 14).

[20] Rui Xu and Donald Wunsch. «Survey of Clustering Algorithms». In: *IEEE Transactions on Neural Networks* 16.3 (2005), pp. 645–678. DOI: `10.1109/TNN.2005.845141` (cit. on p. 7).

[21] Anil K. Jain. «Data Clustering: 50 Years Beyond K-Means». In: *Pattern Recognition Letters* 31 (2010), pp. 651–666 (cit. on p. 7).

[22] OpenAI et al. «GPT-4 Technical Report». In: *arXiv preprint arXiv:2303.08774* (2023). URL: `https://arxiv.org/abs/2303.08774` (cit. on p. 8).

[23] Meta AI. *Meta Llama 3*. `https://ai.meta.com/blog/meta-llama-3/`. Accessed: 2025-02-06. 2024 (cit. on p. 8).

[24] B. Bai et al. «Qwen Technical Report». In: *arXiv preprint arXiv:2309.16609* (2023) (cit. on p. 8).

[25] Syed Mahbubul Huq and Basem Suleiman. «Content Filtering on YouTube: An LLM Approach for Detecting and Scoring Harmful Content». In: *Companion Proceedings of the ACM on Web Conference 2025 (WWW '25)*. Sydney, NSW, Australia: Association for Computing Machinery, 2025, pp. 1988–1992. ISBN: 979-8-4007-1331-6. DOI: `10.1145/3701716.3718388`. URL: `https://doi.org/10.1145/3701716.3718388` (cit. on pp. 9, 18).

[26] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. «Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing». In: *ACM Computing Surveys* 55.9 (2023), 195:1–195:35. DOI: `10.1145/3560815` (cit. on p. 9).

[27] Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. «Finetuned Language Models are Zero-Shot Learners». In: *arXiv preprint arXiv:2109.01652* (2022). URL: `https://arxiv.org/abs/2109.01652` (cit. on p. 9).

[28] Yujia Zhou, Nathanael Schärli, Le Hou, Quoc V. Le, Jason Wei, Slav Petrov, Philippe Gabry, Adams Wei Yu Liu, and et al. «Large Language Models Are Human-Level Prompt Engineers». In: *arXiv preprint arXiv:2211.01910* (2022). URL: `https://arxiv.org/abs/2211.01910` (cit. on p. 9).

[29] Hugging Face. *LLM Prompting Guide*. [Online; accessed 3-Nov-2025]. 2024. URL: `https://huggingface.co/docs/transformers/main/tasks/prompting` (cit. on p. 9).

[30] Tianyu Gao, Adam Fisch, and Danqi Chen. «Making Pretrained Language Models Better Few-Shot Learners». In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL)*. Online: Association for Computational Linguistics, 2021, pp. 3816–3830. DOI: `10.18653/v1/2021.acl-long.295`. URL: `https://aclanthology.org/2021.acl-long.295` (cit. on p. 9).

[31] *Regulation (EU) 2016/679 of the European Parliament and of the Council. General Data Protection Regulation (GDPR)*. `https://eur-lex.europa.eu/eli/reg/2016/679/oj`. Official Journal of the European Union. Accessed: 2025-02-06. 2016 (cit. on p. 10).

[32] yt-dlp contributors. *yt-dlp*. `https://github.com/yt-dlp/yt-dlp`. GitHub repository. Accessed: 2025-09-06. 2024 (cit. on p. 11).

[33] Alec Radford et al. *Whisper Large-v3*. `https://huggingface.co/openai/whisper-large-v3`. Accessed: 2025-9-06. 2024 (cit. on p. 11).

[34] Imed Keraghel, Stanislas Morbieu, and Mohamed Nadif. «Beyond Words: A Comparative Analysis of LLM Embeddings for Effective Clustering». In: *Advancing Intelligent Data Analysis (XXII)*. Vol. 14641. Intelligent Data Analysis. Kista, Sweden: Springer, 2024. ISBN: 978-3-031-58547-0. DOI: `10.1007/978-3-031-58547-0_17` (cit. on p. 14).