CHAPTER 1

INTRODUCTION

In the current digital era, digital forensics is crucial for supporting the analysis of digital evidence or academic research. This thesis aims to deepen understanding of digital forensics, focusing on social media that processes large data sets when used as digital evidence. The increasing volume of data generated by various social media applications necessitates analytical solutions that are both rapid and efficient and reliable for decision-making by investigators or academics. These challenges underscore the urgent need for innovation and continuous improvement in digital forensics technologies, which are central to this research. Consequently, this study aims not only to assist investigators in enhancing the efficiency of digital data retrieval but also to make a significant contribution to decision-making efficiency, as investigators and academics can disregard non-essential data. This chapter discusses these challenges and the possibility of handling them in detail.

1.1 Rationale

The rapid growth of social media usage in contemporary times has drawn the attention of practitioners and researchers in digital forensics. Recent data indicates that as of early 2025, there were approximately 5.24 billion active social media users worldwide, representing 63.9% of the total global population. Annual growth reached 4.1%, equivalent to an addition of roughly 206 million new users in the past year. This trend is further highlighted by the fact that the average social media user interacts with 6.8 different platforms each month and spends an average of 2 hours and 21 minutes daily engaging on these platforms.[1] This high level of multi-platform adoption and duration of daily usage directly contributes to creating large volumes and a wide variety of digital data for each individual. Consequently, users' digital footprints are no longer confined to a single platform but are spread across various services, each with distinct data structures, formats, and APIs. This results in a complex and fragmented data landscape.

Alongside technological advancements, social media has become a breeding ground for cybercrimes. An illustrative case involving social media investigation is the 2016 murder of Rito Llamas-Juarez ('Llamas') by Larry Jo Thomas ('Thomas'), which involved investigating the 'Offer Up' social media platform, as well as a Facebook account linked to this platform[2, 3]. This increase in social media usage, and its role in facilitating cybercrimes, emphasizes the urgent need for more advanced and efficient digital forensic methods to effectively address and mitigate the prevalence of criminal activities on social media platforms.

In addition to increasing crime on social networks, the increasing volume and unstruc-

tured data present several open issues, as mentioned in [4]. One prominent issue in social media forensic is the complexity problem, which poses challenges in the analysis and correlation phases. This problem arises from the increasing volume of social media forensic data (quantity problem), and the heterogeneity of evidential data and diverse architectures of social media platforms (diversity problem). These problems impede the efficiency of investigations.

In this context, data generated from user activity on social media—such as text posts, images, videos, private messages, location and time metadata, connection logs, friend or follower lists, and profile information, often serves as highly valuable and irreplaceable digital evidence[5]. The public or semi-public nature of much social media content, combined with the platforms' inherent chronological features and their ability to map social networks among users, makes these data a rich source of information for reconstructing event timelines, identifying motives, uncovering relationships between suspects and victims, and validating alibis. However, this richness of information comes with its complexities. Factors such as the dynamic of platforms, the ease with which data can be manipulated or deleted, varying privacy policies, and challenges in authentication and maintaining the chain of custody make forensic analysis of social media significantly more complex compared to the analysis of digital evidence from more static sources, such as computer hard drives.

A review of the current state of the art shows that various approaches have been proposed to address the challenges in social media forensics. Previous research has focused on developing tools for individual platforms or common data integration frameworks, such as the CISMO[6] ontology oriented towards criminal intelligence from online social networks (OSN) or the layered semantic framework proposed by Arshad et al. [7] for data integration. While these approaches have made significant contributions, a major challenge that remains and constitutes a research gap is the lack of a unified conceptual model specifically designed for generalizing forensic data across fundamentally different platform categories (e.g., between short text-based social media like X, video-based ones like TikTok, and forum-based ones like Reddit). This gap in terms of standardization of data representation that is able to accommodate intercategory heterogeneity is the primary focus and justification of this research.

To address this gap, this study proposes an ontology-based approach. This choice is based on the fundamental advantage of ontology in formally representing domain knowledge, especially when dealing with dynamic domains such as social media. While platform features continue to evolve, the core concepts underlying user interactions—such as the existence of '*Users*', '*Posts*', and various forms of '*Interactions*'—tend to be more stable. With their ability to explicitly define semantic concepts and relationships, ontologies are well-suited to modeling these core concepts. Compared to other approaches such as Knowledge Graphs (KG), which focuses more on large-scale data connections, ontologies provide a rich formal and semantic schema, ensuring consistency and enabling automated reason-

ing capabilities. This capability is crucial for the forensic domain which demands high accuracy. Furthermore, their flexible and extensible nature makes them more adaptable to future platform changes than rigid custom database schemas.

Therefore, this research will focus on developing and applying an ontology-based approach to generalize forensic data originating from various categories of social media platforms. The aim is to significantly enhance the effectiveness and efficiency of SMF investigations considering the challenges posed by the SMF in the era of big data.

1.2 Problem Formulation

In the complex digital era, SMF faces significant challenges in managing and analyzing massive amounts of unstructured data. Data from various social media platforms is highly diverse in format and volume, which can impede the investigation process. Many tools operate in silos, only capable of processing data from one or a few specific platforms. This forces investigators to examine each data source separately and then attempt to correlate findings manually—a process that is not only extremely time-consuming but also highly prone to human error and oversight[8].

Although the potential of ontologies for knowledge representation, data integration, and standardization has been recognized in forensic and AI literature[5], their specific and comprehensive application for forensic data generalization across various heterogeneous social media platforms remains very limited. It represents an area that remains largely unexplored[9]. There is not yet an established and comprehensive domain ontology specifically designed to facilitate this generalization process, which can capture common concepts while accommodating the specific characteristics of diverse platform categories. The lack of universal ontology support is recognized as one of the primary challenges faced in the SMF domain today.

To address these issues and solve the existing problems, this research focuses on the following key research questions:

1. What type of social media forensic data is considered significant by digital forensic examiners?

2. How can social media forensic data be generalized?

These questions are the primary guide in the development and evaluation of the proposed approaches in this research.

1.3 Objective and Hypothesis

To achieve the main objective, this study is systematically designed to answer a series of research questions through specific and measurable objectives, as follows:

1. Identify and validate forensically significant types of data, features, and digital artifacts from various categories of social media platforms, through the process of formulating and validating Competency Questions (CQs) with domain experts.

2. Designing and building a generalized ontology model capable of representing and connecting data from heterogeneous social media sources, through a feature extraction, generalization, and ontology mapping & merging process methodology.

The hypotheses in this study are as follows:

Premise 1: Facing the complexity and scale of Social Media Forensic (SMF) data, data generalization strategies have proven crucial to simplify, categorize, and manage the vast diversity of data, thereby improving the efficiency of forensic analysis. [10] [11]

Premise 2: Ontology-based approaches have great potential to address SMF challenges by providing a common vocabulary, semantic interoperability, standardization of data representation, and support for automated reasoning that can uncover hidden relationships and inconsistent data. [12] [13]

Thus, the hypothesis in this study is: By applying a generalized global ontology model (GENOSIS), forensic data analysis can be performed more effectively and efficiently across heterogeneous social media platform categories.

1.4 Scope and Delimitation

To conduct this research, it is necessary to have a scope of work so that the objectives can be fulfilled within a limited time. Other than that, several limitations need to be considered in this research. These include:

1. This model has only been tested on 4-four categories, each with two platforms as a research scope with the version determined in Table 1.1.

Table 1.1: Social Media Categories

Social Media Categories	***Social Media Platforms***
Online Social Network Sites (OSN)	Facebook, Instagram
Microblogging Platforms (MP)	X (Twitter), Tumblr
Media-content Sharing Sites (MCSS)	YouTube, TikTok
Online Forums/Blogs (OFB)	Reddit, Quora

2. This research focuses on client-side media, in this case, mobile phones, and is limited to the device versions and technologies available when the research was conducted.

The categorization of social media platforms used in this study, as presented in Table 1.1, is compiled by referring to existing classifications from a study conducted by Basumatary et al. [4], which maps various platforms based on their functions. Based on this mapping, this study then makes modifications and selections to narrow down the categories that are most relevant to the purpose of generalizing data for digital forensics purposes. This study focuses on four primary categories, namely Online Social Network Sites (OSN), Microblogging Platforms (MP), Media-content Sharing Sites (MCSS), and Online Forums/Blogs (OFB). The selection of categories and social media platforms was conducted to maintain a manageable and in-depth scope of the study by establishing research limitations while ensuring that the selected platforms possess characteristics that are representative enough to be analyzed in the context of digital forensics.

1.5 Significance of Study

This research has important theoretical and practical significance in digital forensics. Theoretically, this research contributes to the development of literature, especially in generalizing of SMF data through an ontological approach. The resulting GENeralized Ontology for Social Media Investigation Support (GENOSIS) is expected to be a reference for other researchers interested in exploring similar phenomena or developing theories about handling heterogeneous data from various social media platforms. Practically, the results of this study provide substantial benefits for digital forensic investigators and examiners. The developed GENOSIS ontology model can help them understand the complex and diverse data structures of various social media platforms in an integrated manner. This facilitates a faster and more efficient investigation process, as it allows investigators to identify, correlate, and analyze cross-platform data more effectively, and focus attention on the most relevant digital artefacts, thereby reducing the time required for examination. Furthermore, the GENOSIS model has the potential to become a standardization framework in the development of triage and data analysis systems to support digital forensics and Open-Source Intelligence (OSINT) processes in the social media domain, enabling a more structured and comprehensive analysis of public data.