

Use of Natural Language Processing Techniques for Forensic Analysis in Spanish [†]

Luis Alberto Martínez Hernández [‡], Ana Lucila Sandoval Orozco [‡] and Luis Javier García Villalba ^{*,‡}

Group of Analysis, Security and Systems (GASS), Department of Software Engineering and Artificial Intelligence (DISIA), Faculty of Computer Science and Engineering, Office 431, Universidad Complutense de Madrid (UCM), Calle Profesor José García Santesmases, 9, Ciudad Universitaria, 28040 Madrid, Spain; luisal11@ucm.es (L.A.M.H.); asandoval@fdi.ucm.es (A.L.S.O.)

* Correspondence: javiergv@fdi.ucm.es; Tel.: +34-91-394-7638

[†] Presented at the First Summer School on Artificial Intelligence in Cybersecurity, Cancun, Mexico, 3–7 November 2025.

[‡] These authors contributed equally to this work.

Abstract

In the digital forensics process, an essential step is the analysis of evidence contained in seized devices, a task that requires a significant investment of time to identify patterns and evidence that strengthen a judicial investigation. Advances in Natural Language Processing (NLP), particularly models based on Transformers, offer great potential for automating this analysis and facilitating the accurate detection of relevant information. However, there are still a few solutions in Spanish aimed at processing legal texts or identifying crimes. This work proposes an automated methodology for analysing digital evidence using NLP techniques trained for Spanish text. Its objective is to optimise the extraction of relevant information in the forensic context, reducing the analysis time and improving the accuracy of detecting data that is significant for investigations.

Keywords: artificial intelligence; digital information; forensic analysis; natural language; named entity recognition; relationship extraction; transformers

1. Introduction

The rapid advancement of technology has led to a significant rise in digital crimes, making the accurate identification of evidence and clues essential for the success of forensic investigations. Digital forensics has become a crucial discipline, not only for investigating cybercrimes but also for performing expert analyses on devices that may have been involved in other types of offences. One of its core challenges lies in analysing and classifying vast amounts of data stored on electronic devices to identify relevant evidence. However, this process remains complex due to the volume and heterogeneity of digital information. Although several tools enable partial automation, they often require case-specific adaptation and are typically limited to superficial analyses such as examining hashes, file sizes, or metadata without addressing the actual content of documents, which reduces their effectiveness in uncovering meaningful evidence. The digital evidence refers to any information with potential investigative value that is stored, transmitted, or received in digital form. Such evidence is obtained through the seizure and secure preservation of electronic devices and may include various data types, such as audiovisual material (audio, images, or videos), identity documents, travel records, credit cards, and other digital files.



Academic Editors: Héctor Manuel Pérez Meana and Gabriel Sánchez Pérez

Published: 4 February 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Advances in Machine Learning (ML) and Natural Language Processing (NLP) have transformed how large textual datasets are interpreted and analysed. These technologies now enable the extraction of meaningful insights and key ideas from complex documents, even in domains requiring specialised knowledge. This progress has been driven by the availability of large datasets and increasingly sophisticated algorithms. ML-based NLP models can understand the semantics, context, and intent within texts, supporting tasks such as document classification, entity recognition, sentiment analysis, and automatic summarisation with remarkable accuracy and efficiency.

This paper proposes a methodology for training NLP models to be integrated into digital forensic workflows, aiming to automate and enhance the extraction of information relevant to investigations. Special emphasis is placed on state-of-the-art NLP models and tools applicable to the forensic domain. The rest of this paper is organised as follows: Section 2 introduces Natural Language Processing; Section 3 discusses the digital forensic process and its main analytical challenges; Section 4 reviews the main pre-trained Transformer models for NLP; Section 5 details the experimental setup; and Section 6 concludes the study.

2. Natural Language Processing

Natural language processing is an area of computer science that is primarily concerned with transforming natural language into a formal language, such as programming language, so that a computer can process and understand it, mainly combining techniques from AI, linguistics, statistics and computing for the development of algorithms.

Semantic and Syntactic Analysis

It is a fundamental part of NLP as it focuses on understanding the meaning of words, phrases or sentences and how they are used in a specific context. This approach not only focuses on grammatical structure, as syntactic analysis would, but attempts to interpret the actual meaning of the text. This process is complemented by pragmatics, which attempts to understand how language is used in different situations and contexts. Although there have been great advances in semantic analysis, the challenge of understanding the basic meaning of language remains. Semantic analysis in NLP is crucial for tasks such as natural language production and understanding massive amounts of unstructured data in fields such as banking and medicine [1].

Syntactic analysis in NLP techniques is the process of understanding the structure of a sentence or phrase and extracting relevant information such as nouns, verbs, and adjectives, taking into account the grammatical structure of a sentence and understanding the syntactic relationships between words [2], in addition to identifying the roles that words play in the sentence. To identify the structure of the phrase, during the syntactic analysis process, the sentence is divided into its different parts: subject, verb, object, and modifiers. The hierarchical relationships between these parts are also identified, showing how they are grouped to form meaningful units. For example, in the sentence, “*The old cabin is in the forest*”, syntactic analysis would recognise “*The old cabin*” as a group, which acts as the subject, and “*is in the forest*” as another group that acts as the predicate. Among the most common approaches to syntactic analysis in NLP are rule-based analysis, statistical-based analysis, probabilistic syntactic analysis, dependency syntactic analysis, and neural networks.

3. Computer Forensics Process

The act of detecting, obtaining, preserving, and analysing electronic evidence that can be used in court is known as computer forensics [3]. This process involves using methodologies to search for evidence on digital devices such as computers, smartphones, servers, and networks.

In general terms, the digital forensics process consists of five main phases applied to the analysis and presentation of evidence from various storage sources, such as hard drives, USB sticks, SD cards, mobile phones, cloud services, IoT devices, and virtualised environments. The first phase, identification, consists of recognising possible sources of digital evidence. In the acquisition phase, data is extracted using forensic methodologies and tools that guarantee the integrity of the information. The analysis stage involves examining, interpreting, and classifying the data obtained to generate information relevant to the investigation. Subsequently, in the documentation phase, the corresponding reports and conclusions are prepared. Finally, the presentation phase allows the results to be used as valid evidence in a legal proceeding.

Two key steps stand out in the forensic process: the acquisition and analysis of evidence, which enable the resolution of a possible crime, the prevention of future crimes, and the protection of the integrity of the judicial process. However, this process faces several challenges, mainly the massive volume and complexity of digital data, which makes it difficult to identify relevant evidence. Additionally, the analyst's lack of knowledge about certain aspects of the case context can also have an impact, potentially hindering the identification of pertinent evidence.

Although there are technologies that enable the rapid extraction of information from a device, the analysis of all documents often requires manual processing. This is where NLP techniques can be beneficial in automating the process and extracting information that can support the investigation. This paper explores transformer models for automatic document analysis.

4. Pre-Trained Language Models

Thanks to advances in deep learning and the availability of large volumes of data, Natural Language Processing (NLP) has experienced significant growth in recent years. Among the most notable models are those based on Transformers, such as BERT, RoBERTa, and ALBERT, which have proven their effectiveness in multiple NLP tasks. BERT uses only the encoder part of the Transformer and is trained using two main tasks: the Masked Language Model (MLM), which predicts 15% of randomly masked tokens, and Next Sentence Prediction (NSP), which determines whether one sentence follows another, useful for question-answering tasks. During the fine-tuning phase, BERT is adjusted with specific data by reusing pre-trained weights, thus reducing the computational cost. RoBERTa optimises BERT by eliminating the NSP task and employing dynamic masking, where sentences are duplicated and masked tokens change at each epoch. It also increases the size of the mini-batch, the BPE and the training corpus, incorporating sets such as BOOKCORPUS and CC-NEWS. Finally, ALBERT reduces the high computational cost of BERT by factorising embeddings, sharing parameters between layers, and replacing NSP with the Sentence Order Prediction (SOP) task.

5. Experiments

The proposed solution focuses on the analysis phase of the digital forensic process, in which it is necessary to identify key entities within documents, images, or videos. Once the relevant devices have been identified, files that may contain evidence are extracted and converted into a structured format suitable for processing by an AI model that performs

entity recognition. The experiments were conducted in Spanish, using a corpus compiled from anonymised court rulings issued by the General Council of the Judiciary (CGPJ), which mainly dealt with crimes affecting the psychological and physical well-being of victims and referred to Spanish legislation. The dataset included ten labels (DRUGS, WEAPON, NAL, DI, PER, LOC, DATE, ORG, MISC, LAW) and underwent extensive cleaning, normalisation and lemmatisation to improve model performance.

After processing and balancing, six labels and their corresponding samples were retained for experimentation, ensuring a more robust and accurate evaluation of the PLN model, obtaining the following distribution, eliminating the Drugs, Weapons, Nationality, and Identity Document classes due to the limited data obtained: PER 1796, LOC 344, DATE 1663, MISC 668, ORG 1848, and LAW 3745.

The model used for the experiments was RoBERTa Base BNE, which has been pre-trained for Spanish text processing with a corpus of 570 GB of clean, duplicate-free texts from the annual collection made by the Spanish National Library of all “.es” domains between 2009 and 2019.

Evaluation

Due to the number of occurrences of each label and the fact that in certain cases they could be considered insufficiently present, it was decided to train a first model with the complete dataset (“Legal-GASS”) including the DRUGS, WEAPON, NAL and DI labels to check the metrics of the main labels. Subsequently, a second model was trained using only the main labels, called “Legal-GASS-lite”. The results of the trained model are shown in Table 1.

Table 1. Accuracy by class.

	PER	LOC	DATE	MISC	ORG	LAW
<i>Legal-GASS (Accuracy)</i>	86.05	46.94	74.95	46.11	62.30	58.98
<i>Legal-GASS-lite (Accuracy)</i>	80.61	54.69	80.61	40.68	62.37	56.93
<i>Legal-GASS (Recall)</i>	83.25	26.44	73.44	19.25	53.44	60.36
<i>Legal-GASS-lite (Recall)</i>	85.12	20.11	76.92	18	60.78	56.62
<i>Legal-GASS (F1-Score)</i>	84.63	33.82	74.19	27.16	57.53	59.66
<i>Legal-GASS-lite (F1-Score)</i>	82.80	29.41	78.73	24.96	61.56	56.78

As can be seen, depending on the class, the accuracy of the model increases. This is because texts are sometimes difficult to classify into a class. The most common problems are based on the ambiguity of phrases, for example, organisations with people’s names, the lack of labels or variety in examples of them, and single phrases in the corpus. Figure 1 shows an example of the output of entity recognition.



Figure 1. Graphical Representation of Recognised Entities.

6. Conclusions

This study showed that the use of NLP techniques for document analysis in a forensic process offers significant benefits. These techniques enable the rapid identification of patterns, themes, and relationships in unstructured information sources, facilitating the detection of evidence that may be relevant to an investigation and the possible reconstruction of events. Furthermore, the application of NLP in this context improves the efficiency of the process by automating tedious textual analysis tasks, allowing investigators to focus on deeper interpretations and conclusions. The capabilities of NLP can also help uncover hidden intentions or deception in documents, increasing the reliability and accuracy of forensic results.

Author Contributions: Conceptualization, L.A.M.H., A.L.S.O. and L.J.G.V.; methodology, L.A.M.H., A.L.S.O. and L.J.G.V.; validation, L.A.M.H., A.L.S.O. and L.J.G.V.; investigation, L.A.M.H., A.L.S.O. and L.J.G.V. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the European Commission under the Horizon 2020 research and innovation programme, as part of the project HEROES (<https://heroes-fct.eu>, Grant Agreement no. 101021801) and of the project ALUNA (<https://aluna-isf.eu/>, Grant Agreement no. 101084929). This work was also carried out with funding from the Recovery, Transformation and Resilience Plan, financed by the European Union (Next Generation EU), through the Chair “Cybersecurity for Innovation and Digital Protection” INCIBE-UCM. In addition, this work has been supported by Comunidad Autonoma de Madrid, CIRMA-CM Project (TEC-2024/COM-404). The content of this article does not reflect the official opinion of the European Union. Responsibility for the information and views expressed therein lies entirely with the authors.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in this study are included in the article.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Khurana, D.; Koli, A.; Khatter, K.; Singh, S. Natural language processing: State of the art, current trends and challenges. *Multimed. Tools Appl.* **2023**, *82*, 3713–3744. [[CrossRef](#)] [[PubMed](#)]
2. Li, J.; Liu, M.; Qin, B.; Liu, T. A survey of discourse parsing. *Front. Comput. Sci.* **2022**, *16*, 165329. [[CrossRef](#)]
3. Raghavan, S. Digital forensic research: Current state of the art. *CSI Trans. ICT* **2013**, *1*, 91–114. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.