*CORSO DI DOTTORATO DI RICERCA IN BIG DATA MANAGEMENT*

*XXXV-ciclo*

*COORDINATORE: PROF: VALERIO ANTONELLI*

# Abstract Tesi di Dottorato

*Candidato: Domenico Alfano*

*Tutor: Prof. Domenico Parente*

Digital transformation has triggered a monumental shift in how data are generated and collected, ushering in an era of unprecedented information. This phenomenon, often described as a genuine "data explosion," has opened the doors to a universe rich in information useful across numerous fields of knowledge. However, it has also introduced significant challenges in processing and analyzing this immense volume of data. One of the most pertinent challenges is the ability to detect non-standard patterns or deviations, a field that has established itself as a crucial area of research in data analysis.

In this context, Anomaly Detection plays a central role. This approach is vital for distinguishing relevant data within a sea of information.

This work aims to examine and delve into the importance of Machine Learning in tackling the complexities associated with anomaly detection. The focus is on three main aspects: the analysis of statistical irregularities, the investigation into the semantic depth of data, and the identification of errors.

Indeed, the second chapter of the thesis outlines the theoretical foundations of Anomaly Detection, offering an in-depth view of definitions, challenges, and key methodologies in this field. From the importance of understanding the nature of irregularities in data to the application of advanced Machine Learning and Natural Language Processing algorithms.

This chapter provides the conceptual basis necessary for understanding the scope of our research.

The core of the research develops in the following three chapters, each dedicated to a specific perspective of Anomaly Detection, offering practical applications of methodologies through specific case studies. Through the analysis of results obtained in real-world contexts, we aim to demonstrate the effectiveness of the methodological proposals in concrete application situations, providing a critical framework of traditional approaches and the latest innovations in Anomaly Detection. Through this analysis, we aim to position the work within the current context, identifying gaps in the existing literature and outlining the originality of the contributions.

The third chapter addresses the analysis of statistical irregularities, examining how Machine Learning and Natural Language Processing techniques can identify anomalous patterns in data.

Using models and supervised learning techniques, the limitations of traditional methods are overcome, improving the precision and reliability of anomaly detection in the field of cryptocurrencies. Cryptocurrencies, with their growing popularity as an investment vehicle, have led to a significant increase in opportunities and, unfortunately, scams.

Among the most widespread frauds is the Pump and Dump (P&D), a tactic used by unscrupulous speculators to artificially manipulate the value of a cryptocurrency.

Pump and Dump is a well-orchestrated fraud that involves market manipulation through the spread of false

information to artificially inflate the price of a cryptocurrency. In this scheme, fraudulent operators gradually accumulate a significant number of tokens of a specific cryptocurrency. Then, they actively promote this cryptocurrency, spreading false or exaggerated news about it through online channels such as social media and group chats. This phase of promotion is called "pumping." Once the price of the cryptocurrency has been artificially inflated due to market manipulation, the scammers proceed to quickly sell off the accumulated tokens to unsuspecting investors, causing a rapid drop in price. This phase of selling is known as "dumping." Investors who purchased the cryptocurrency based on false information during the pumping phase suffer significant losses when the price collapses.

Detecting Pump and Dump schemes is a complex challenge, but in recent years, innovative approaches have been developed, often based on advanced natural language analysis and social media.

One of the most promising methods for identifying Pump and Dump schemes is the analysis of social media and online chats, where fraudulent operators often coordinate their activities. Platforms like Telegram have become favored places for planning and executing P&D. Natural Language Processing (NLP) emerges as a crucial tool in combating Pump and Dump. This discipline allows for advanced analysis of natural language used in online messages, enabling the identification of patterns and trends that could indicate fraudulent activity.

Through NLP, it is possible to extract opinions and sentiments from social media and chat users. Identifying linguistic signals that suggest intentional manipulations, such as the use of hyperbolic expressions or the spread of contradictory information, can contribute to the early detection of P&D.

Scammers often create specific public groups to spread false information and coordinate Pump and Dump. NLP can be employed to monitor these groups, identifying suspicious behavioral patterns, and detecting the use of misleading language. The constant progress in technology and the development of advanced algorithms enable the refinement of detection techniques. Machine learning algorithms can be trained on large volumes of historical data, identifying distinctive traits of Pump and Dump schemes, and continuously adapting to new strategies adopted by scammers.

In conclusion, this study demonstrates the possibility of identifying a Pump and Dump scheme as soon as it begins, highlighting that financial exchanges have superior detection capabilities thanks to access to more detailed data. This data includes in-depth information on transactions, their volumes, and the actors involved in Pump and Dump. The study also underscored the importance of information from social media in detecting P&D in cryptocurrencies.

Thanks to the use of features derived from social media, previous approaches have been surpassed, achieving results that position themselves at the forefront of the state of the art.

The fourth chapter focuses on the assessment of semantic richness in data, introducing the role of semantic understanding and how it can enrich the anomaly identification process by considering contexts and relationships that go beyond simple statistical irregularities. Through this perspective, the aim is to grasp the complexity and depth of the information contained in the data, further improving anomaly detection capabilities. The focus in this case is on legal text, where anomalies are considered a semantic richness to be leveraged for correctly segmenting the text.

In the complex world of legal documents, text segmentation plays a fundamentally important role. Legal documents typically consist of a series of segments that, while being semantically coherent and linked by a common logical thread, can vary greatly in terms of length and themes addressed. This structural feature makes them particularly complex and often difficult to navigate, especially for those who are not experts in the legal field.

Text segmentation comes into play precisely to tackle this challenge: it allows for the division of verbose and continuous textual documents into smaller units, defined as segments, which are easier to manage and analyze. In the legal domain, where time is often a critical factor and accuracy is imperative, effective text segmentation can represent a considerable advantage. For legal professionals, having the ability to identify the relevant sections of a document quickly and accurately can mean a better understanding of the case and more efficient strategic planning. Moreover, text segmentation proves to be a crucial step within legal Natural Language Processing pipelines, especially in the preliminary phases of processing a document, when it is necessary to prepare the text for more complex operations such as Named Entity Recognition (NER) or text summarization.

However, the segmentation of legal documents is not without challenges. Legal terminology, by its nature, is complex and full of specificities that can vary greatly from one context to another. Furthermore, the length of legal documents can be extensive, with texts often extending for hundreds or even thousands of pages. These factors, combined with the potential noise introduced by handwritten, scanned, or somehow altered texts, make segmentation a particularly daunting task. In response to these challenges, various neural approaches to text segmentation have emerged, some of which have been specifically designed to adapt to the legal context. These methods employ advanced artificial intelligence technologies to analyze documents and identify the various segments efficiently and accurately.

Despite these advancements, many of these methods do not incorporate a crucial functionality: the classification of segments within the same text segmentation activity. In other words, while they can divide the text into segments, they do not categorize these segments based on their content or function within the document. Furthermore, many approaches also present another difficulty: they analyze the entire document and, to avoid the token limit problem, use techniques such as chunking, which reduce the model's accuracy.

To overcome these limitations, this thesis introduces a new approach based on the detection of "break-points." This innovative method not only identifies segments within a legal document but also classifies the lines of text that signal the start of a new segment.

The basis of this method is a transformer-based model, an advanced form of neural network, which has been pre-trained on pre-processed texts from the Italian National Jurisprudence Archive. This preliminary training provides the model with a solid knowledge base on the structure and language of legal documents, allowing it to adapt and react more effectively to the various types of text it may encounter. The proposed approach goes beyond simple text segmentation, integrating both break-point detection and section classification.

This translates into a more complete and versatile solution for segmenting legal documents, which is robust even in the face of textual disturbances and noise. Indeed, the model has proven to be much more reliable than a rule-based engine specifically built on the textual structure of documents, showing remarkable adaptability and flexibility.

It is important to highlight that, although the presented model is focused on Italian legal documents, the methodology and techniques employed can be adapted and extended to tackle the challenges of text segmentation in legal documents of other languages. This opens the way to a vast field of applications, allowing the approach to be adapted to different linguistic and legal contexts.

In the fifth chapter, anomalies are considered as possible errors, carefully analyzing errors in the data, and highlighting them as true anomalies. This perspective aims to improve the accuracy of the detection process, providing a more robust methodology for identifying anomalous patterns related to incorrect information. In this case, the focus is on Optical Character Recognition systems.

Optical Character Recognition systems, commonly known as OCR, represent one of the most significant innovations in the field of digital information processing. OCR, which literally means "optical character recognition," is a technology that allows converting various types of documents into digital textual data. This transformation makes the text machine-readable and thus susceptible to being edited, searched, and electronically stored, revolutionizing the way we manage and store information, making data easily accessible, editable, and searchable.

The concept of OCR is not new; its roots can be traced back to the early 20th century, with the development of the first character reading machines. These early devices were rudimentary and limited in recognizing only a restricted set of fonts and numbers. However, with the advent of the digital age and the development of computing, OCR has undergone rapid evolution.

Modern OCR systems use complex algorithms and machine learning techniques to identify characters in nearly every type of font and in various languages, with ever-increasing accuracy.

Despite the advancements in OCR technologies, the results are often not perfect and require an additional step called post-processing. This process aims to improve the quality of the recognized text by identifying and correcting errors that occurred during the initial scanning and recognition.

OCR post-processing is divided into two main parts: error detection and error correction. Both phases are crucial for ensuring the accuracy of the final document. Error detection involves identifying parts of the text that may have been incorrectly recognized. This can range from simple typographical errors to more complex issues like confusion between similar characters.

For error correction, once errors are identified, the system attempts to correct them. This can be done through various methods, such as comparing with a database of correct words or analyzing the context to determine the most probable word. In this case, the study aims to capture errors stemming from OCR, developing a specific model for detecting these anomalies.

It was necessary to build a dataset containing typical errors of an OCR system, using the technique of "Data Augmentation", commonly employed in machine learning and deep learning, to increase the quantity and diversity of training data. This practice is essential in the presence of a limited dataset, as it helps prevent overfitting and improve the model's generalization capability on new data. Textual data augmentation artificially increases the amount of textual data available, enhancing the performance of machine learning and artificial intelligence models, especially in scenarios with limited data. Common techniques include paraphrasing, back translation, introduction of intentional spelling or grammatical errors, synonym substitution, random word insertion, random word deletion, and sentence rearrangement.

Specifically, for this study, the perturbation technique was used to simulate real-world conditions in which legal professionals work with text documents digitized through OCR.

The resulting dataset was used to train a Transformer-based classifier, pretrained on texts from the Italian National Archive of Jurisprudence. ITALIAN-LEGAL-BERT was chosen for its ability to capture the nuances of legal language. This study is part of a broader effort to solve the problem of detecting OCR system errors and correcting them. This work, named REMOAC (A Retroactive Explainable Method for OCR Anomalies Correction in Legal Domain), employs cutting-edge methods to address the complex challenges presented by OCR errors in this specialized field. The primary goal of REMOAC is to significantly improve the performance of legal text classifiers by correcting errors introduced during the OCR scanning phase.

Finally, in chapter six, the conclusions and prospects summarize the results obtained, reflect on the importance of the findings in the broader context of data analysis, and suggest directions for future research. In this way, the thesis makes a significant contribution to the evolution of Anomaly Detection, integrating the potential of Machine Learning and Natural Language Processing in a coherent manner.