Unpacking GDPR Enforcement: A Semantics Analysis and Topic Modeling Approach to Understanding Fines and Compliance Patterns

1. Orlando and M. Santoro

Istituto per le applicazioni del Calcolo “Mauro Picone” – I.A.C

Consiglio Nazionale delle Ricerche

a.orlando@iac.cnr.it; m.santoro@iac.cnr.it

The General Data Protection Regulation (GDPR) is a comprehensive data protection law that was passed by the European Parliament in 2016 and became enforceable on May 25, 2018. From this date onwards, all organizations that process personal data of individuals within the European Union (EU) were required to comply with its regulations. GDPR's primary goal is to harmonize and strengthen data protection laws across EU member states, providing individuals with greater control over their personal data. It replaces the 1995 European Data Protection Directive, which laid the foundation for data protection laws within the EU. However, the 1995 directive left room for national variations in interpretation and enforcement, leading to inconsistencies across member states. One of the core purposes of the GDPR is to address this fragmentation by creating a unified legal framework, simplifying compliance for businesses, and reducing administrative burdens for organizations operating across multiple EU countries.

The GDPR contains 99 articles, 42 of which have the potential to result in fines. Article 83 specifically outlines two levels of fines, which reflect the severity of different types of violations and the importance of certain GDPR principles. The first level of fines (Tier A) applies to breaches of Articles 8, 11, 25-39, 41, 42, and 43, and can result in a maximum penalty of either €10,000,000 or 2% of the company’s total global annual turnover, whichever is higher. The second level of fines (Tier B) is reserved for more severe breaches, including violations of Articles 5, 6, 7, 9, 12-22, 44-49, and 58, and carries a maximum fine of €20,000,000 or 4% of global annual turnover, again whichever is greater.

The GDPR fundamentally reshapes how organizations collect, process, and store personal data. While compliance may seem like a challenge, it also brings notable benefits, especially in terms of cybersecurity. By adhering to GDPR guidelines, organizations must adopt stricter security measures to protect personal data, which in turn helps mitigate cyber risks. This increased focus on data protection often leads companies to invest more in security technologies and practices, such as encryption and data minimization, which safeguard against data breaches. Additionally, some organizations opt to transfer residual risk by purchasing cyber insurance, which can help cover the costs of managing breaches, fines, and notifying affected individuals. The GDPR has, therefore, contributed to the growing demand for cyber insurance, as businesses seek to protect themselves against the financial consequences of non-compliance and data breaches.

Understanding how the GDPR is enforced across EU member states is of critical importance for organizations handling personal data, as it allows them to learn from past enforcement actions and improve their compliance practices. Cybersecurity experts also need to comprehend the link between GDPR compliance and cybersecurity to better protect organizations from potential data breaches. Moreover, researchers in fields such as information technology, law, and economics are keenly interested in the dynamics of GDPR enforcement, as it provides valuable insights into the regulation’s practical implications.

In this context, the goal of our research is to analyze GDPR fines by examining their motivations, amounts, the countries in which they were issued, the business sectors involved, and the severity of the violations. We aim to provide a comprehensive overview of GDPR fines from the regulation's implementation up to June 2023.

Despite the widespread enforcement of GDPR, there is no centralized database provided by the European Union that tracks enforcement actions by Data Protection Authorities (DPAs) across member states. This lack of a unified data source has led to the development of various online collections created by non-governmental organizations, companies, and other entities. For this study, we rely on the CMS Law-GDPR Enforcement Tracker (ET), which records fines imposed across all 27 EU member states, as well as in the United Kingdom and Norway. This dataset includes fines related to 10 different business sectors and is based on a publicly available archive containing numerous high-profile GDPR enforcement cases. This dataset helps address the challenges of annotating and labeling unstructured legal documents, providing a valuable resource for analyzing GDPR enforcement patterns.

For each fine in the ET dataset, several key details are provided:

- an “Enforcement Tracker ID” (ETid) to uniquely identify each case;

- information about the country where the decision was made, the date of the decision, the amount of the fine, the name of the controller or processor that was fined, and the relevant business sector;

- the specific GDPR articles that were breached;

- a generalized classification of the type of fine;

- a link to the data source;

- a summary text describing the case.

Our study employs advanced text analysis techniques, including semantic analysis and topic modeling, to explore patterns and relationships within GDPR enforcement data. By analyzing individual words, bigrams (two-word combinations), and trigrams (three-word combinations), we are able to gain deeper insights into the context and associations present in the enforcement decisions. We utilize a customized version of structural topic models (STM) to identify thematic structures within the text and enrich our understanding of the motivations behind fines.

To enhance our analysis, we use WordNet lexical graphs, which visualize the semantic relationships between words. This approach allows us to capture nuances in meaning and uncover latent patterns that might not be immediately apparent from the raw text. Specifically, we group fines based on their tier (Tier A, Tier B, or a combination of both) and categorize them according to the amount of the fine: less than €10,000, between €10,000 and €100,000, and more than €100,000.

The results of our analysis provide valuable insights into GDPR enforcement trends. For example, we examine how certain words or themes are more prevalent in specific categories of fines, helping to identify key drivers of enforcement actions. The WordNet lexical graphs further illuminate the relationships between terms within different categories, highlighting subtle differences in meaning that may affect how fines are assessed.

By using this framework, we can uncover latent thematic structures within GDPR enforcement data, offering valuable information for stakeholders such as policymakers, legal professionals, and researchers. Our findings contribute to a deeper understanding of GDPR compliance patterns, enabling organizations to better align their data protection practices with the regulation’s requirements and ultimately improve their overall compliance efforts.