CREFF: A Verifiable Human Baseline for Spanish Forensic Linguistics in the Age of Generative AI

CREFF: Una Base de Referencia Humana Verificable para la Lingüística Forense en Español en la Era de la IA Generativa

Desirée M. Carrión,[1] **Ronghao Pan,**[2] **Ángela Almela,**[1] **Rafael Valencia-García**[2]
[1]Department of English Philology, Faculty of Humanities, University of Murcia
[2]Department of Informatics and Systems, Faculty of Computer Science, University of Murcia
{desiree.martinezc, ronghao.pan, angelalm, valencia}@um.es

Abstract: Research in Spanish forensic linguistics and natural language processing (NLP) heavily relies on web-scraped, mono-genre datasets, facing challenges in distinguishing organic human writing from synthetic text generated by Large Language Models (LLMs). This paper introduces CREFF (Corpus de Respuestas en Español para Fines Forenses), a resource replicating Chaski's (2001) framework. The corpus comprises 393 verified human authors across 10 forensic scenarios compiled between 2019 and 2024, providing a solid pre-AI human baseline. A cross-domain authorship identification pilot study evaluates traditional machine learning models, Transformer architectures, and LLMs integrated with psycholinguistic features from UMUTextStats and LIWC-22. Results show that classic models combined with explicit stylistic cues significantly outperform deep neural approaches and LLMs in unseen domains. These findings underscore the need for explainable features and interpretability in judicial practice while paving the way for a novel forensic taxonomy.
Keywords: Forensic Linguistics, Generative AI, Authorship Identification, CREFF.

Resumen: La investigación en lingüística forense y procesamiento del lenguaje natural (PLN) en español depende principalmente de conjuntos de datos mono-género extraídos de la web, lo que dificulta discernir entre autoría humana y texto sintético generado por modelos de lenguaje (LLMs). Este artículo presenta CREFF (Corpus de Respuestas en Español para Fines Forenses), un recurso basado en el modelo de Chaski (2001) que compila textos de 393 autores verificados en 10 escenarios forenses entre 2019 y 2024, estableciendo una base de referencia pre-IA. Mediante un estudio piloto de identificación de autoría cross-domain, evaluamos modelos de aprendizaje automático clásicos, Transformers y LLMs, utilizando rasgos psicolingüísticos de UMUTextStats y LIWC-22. Los resultados demuestran que los sistemas clásicos combinados con rasgos lingüísticos explícitos superan a los enfoques neuronales en dominios no vistos. Estos hallazgos enfatizan la relevancia de la interpretabilidad, sentando las bases para una futura taxonomía forense.
Palabras clave: Lingüística Forense, IA Generativa, Identificación de Autoría, CREFF.

1 Introduction

The motivation behind this paper lies in the need to develop tailored linguistic resources for Natural Language Processing (NLP) in Spanish. The creation and subsequent publication of corpora or curated datasets are essential to enrich the NLP landscape, offering a foundation for the development, testing, and validation of novel computational methods. Following McEnery's (2022) words, corpora can serve as "either the raw fuel of NLP, and/or the testbed on which an NLP application is evaluated" (p. 494).

Nevertheless, human communication is inherently complex as it may encompass aspects that are multifaceted and highly context dependent. Thus, covering such variety within a single corpus becomes extremely dif-

ficult. While current Spanish NLP benchmarks comprehensively cover social media tasks such as hate speech or sentiment analysis,[1] they heavily rely on uncurated, web-scraped data.

This reliance poses a fundamental methodological challenge, as web-scraped data often results in unstructured inconsistencies (McEnery, 2022) that are further contaminated by the rise of Generative Artificial Intelligence (GenAI), threatening the authenticity of unverified online text (Moon et al., 2025).

To address these limitations, we introduce CREFF (Corpus de Respuestas en Español para Fines Forenses), a controlled and carefully curated corpus composed exclusively of authenticated human data. This corpus comprises several texts on diverse, relevant, and forensic topics, written primarily by native speakers of Castilian Spanish. Our aim is to provide the computational linguistics community with a balanced, representative linguistic resource that supports robust experimentation. With a strong linguistic foundation and verified ground-truth data, CREFF aspires to offer more reliable baselines for experimental NLP and forensic applications.

To demonstrate CREFF's potential, we present a cross-domain authorship identification pilot study, a classic forensic task that greatly benefits from NLP classification techniques. For that matter, we combine traditional classifiers (e.g., SVMs, Random Forest) and Transformer-based language models with psycholinguistic features extracted through the Spanish dictionary of LIWC-22 (Pennebaker et al., 2001; Ramírez-Esparza et al., 2007) and UMUTextStats (García-Díaz et al., 2022). Our preliminary findings demonstrate that explicit linguistic features combined with shallow machine learning models achieve higher accuracy, showing promising applications for explainable forensic analysis.

Thus, this paper is organised as follows: sections 2 and 3 review the principles on which CREFF is founded, as well as related work regarding the corpus' origins, its antecedents, and compilation. Next, section 4 presents the configuration of the study with their main results and pertinent discussion. After that, in sections 5 and 6 we provide concluding remarks, detailed limitations, and a roadmap of future research. Meanwhile, the paper ends with section 7 which discusses the ethical principles that guided the compilation of the corpus.

[1]In Amigó et al. (2024) a web portal to track Spanish NLP datasets is presented.

2 Background and Related Work

Authorship identification has traditionally been applied in several fields such as detecting plagiarism, phishing prevention, cybercrime investigation as well as in forensic linguistics –namely providing evidence for courtroom settings in anglophone countries or analysing anonymous text samples such as suicide notes (Juola, 2006; Chaski, 2005; Brennan et al., 2012). However, forensic practitioners face significant drawbacks regarding the nature of the disputed texts as they tend to be scarce, brief, and highly variable (Chaski, 2007). This issue is further exacerbated when dealing with digital evidence. As Li et al. (2006) remarked, there is no such thing as "digital fingerprints" since people hide or are not required to prove their identity on the Internet, so this hinders the task of tracing the author of a cybercrime.

Modern forensic linguistics is shifting from static idiolectal fingerprints towards a dynamic theory of linguistic individuality (Nini, 2023). This framework posits that capturing stylistic signatures requires observing how an author adapts across different communicative goals, as not every individual possesses the same configuration of grammar—rather, it depends on personal experiences and cultural backgrounds (Grant y MacLeod, 2020).

In the digital era, this task faces an unprecedented challenge due to the proliferation of LLMs. The daily reliance on GenAI, heavily optimised via Reinforcement Learning from Human Feedback (Ouyang et al., 2022; Bai et al., 2022), drives an algorithmic monoculture that inherently decreases lexical diversity (Moon et al., 2025; Padmakumar & He, 2024; Anderson et al., 2024). In fact, recent studies demonstrate that texts generated by advanced LLMs are nearly indistinguishable from human-written text on standard metrics (Huang et al., 2025). Consequently, unverified data scraped from the web cannot ensure genuine human authorship. Therefore, this highlights the need for more resources that cover ground-truth data featuring verified authors to develop resilient methodologies for authorship iden-

tification in the era of MGT (Machine-Generated Text).

3 CREFF

Designed to overcome the constraints and limitations of web-scraped data, we developed the Corpus de Respuestas en Español para Fines Forenses (CREFF). The structure and collection of CREFF are directly replicated from the rigorous framework established by Chaski (1997; 2001) and her Writing Sample Database (WSD), which collects multiple brief texts across diverse communicative settings from numerous individuals with similar sociolinguistic backgrounds. Following her validation and ethical guidelines, we selected subjects sharing key sociolinguistic attributes (dialect, age, and education level) engaged in regular academic writing in exchange for course incentives.

Note that these procedures mirror the protocols that were approved (Chaski, 1997, 2001) by the Institutional Review Board of the Institute for Linguistic Evidence, ensuring the requirements of strict data confidentiality and restricted access.

In the Spanish context, Almela et al. (2019) developed an antecedent for this type of controlled compilation based on Chaski's WSD. That corpus was compiled with contributions from college students alongside inmates that were convicted of gender violence. The authors based their approach on the use of LIWC (Pennebaker et al., 2001) to trace psycholinguistic features combined with ALIAS[2] TATTLER (Chaski, 2005) for the analysis of the collected texts. Their main contribution demonstrated the possibility to distinguish linguistic signatures of abusers from non-abusers, statistically. This emphasised the capacity of controlled multi-scenario corpora to provide valuable insights into the psycholinguistic and forensic aspects of language use.

The compilation of CREFF took place between 2019 and 2024. Second-year undergraduate students enrolled in an English Studies degree volunteered to write ten distinct texts on specific topics. This provided a highly homogeneous group: participants were students with similar educational level, were very close in age, and were native speakers of Castilian Spanish. To encourage genuine engagement, participants were incentivised with an extra credit mark in their final course grade. Nonetheless, they remained unaware of the final purpose of the study. Rather, they were "naïve writers" as Baayen et al. (2002) termed it, and their anonymity was strictly preserved to mitigate the Hawthorne effect.

Following Chaski's framework (2007, p. 140), the topics were selected to capture a wide range of intra-writer variation. Social context and communicative goals directly influence textual form as people do not expect to communicate identically when writing for personal consumption compared to business or professional settings. Thus, Chaski (1997; 2005) proposed a series of elicitations designed to "evoke both home and professional varieties" as well as "emotionally-charged and formal language across several text types" (Chaski, 2007, p. 140). **Table 1** outlines the list of topics validated by the referenced author and integrated in Spanish into CREFF.

CREFF adopts this multi-genre approach, not only to ensure topic diversity, but to effectively elicit intra-writer variation to isolate linguistic and emotional individuality from topic-dependent vocabulary. This multi-genre framework effectively elicits intra-writer variation, forcing authors to draw uniquely from their internal lexicogrammatical repertoires across distinct communicative goals (Nini, 2023).

Building upon these characteristics, CREFF currently comprises texts from 393 participants who met all the inclusion criteria and completed the assigned elicitations. As a result, the corpus represents a balanced dataset containing 3,930 excerpts, with each author contributing exactly 10 samples. To ensure sufficient linguistic density for the analysis of intra-writer variation, participants were required to write a minimum of 300 words per text. No maximum limit was imposed, thereby preserving natural variations in text length as an organic manifestation of the writer's linguistic individuality (Nini, 2023).

Moreover, its controlled compilation and collection timespan provide an invaluable experimental advantage, as there are data gath-

[2] Automated Linguistic Identification and Assessment System. Developed by Chaski (1997; 2001) as a programme to analyse texts from databases (e.g., lemmatising, n-graph sorting, Part-Of-Speech tagging, lexical frequency ranking...).

Nº	Topic
1	Describe a traumatic or terrifying event in your life and how you overcame it.
2	Describe someone or some people who have influenced you.
3	What are your career goals and why?
4	What makes you really angry?
5	A letter of apology to your best friend.
6	A letter to your sweetheart expressing your feelings.
7	A letter to your insurance company.
8	A letter of complaint about a product or service.
9	A threatening letter to someone you know who has hurt you.
10	A threatening letter to a public official (president, governor, senator, councilman, or celebrity).

Table 1: Topics validated by Chaski (1997; 2005) on her WSD.

ered between 2019 and 2022 which establishes guaranteed pre-AI human baseline, free from any generative tools like ChatGPT. Furthermore, data collected between 2023 and 2024 retains verified human provenance, as every sample is connected to an authenticated author who provided explicit demographic, personal data, and consent. This resource contrasts directly with anonymous web-scraped data. In fact, Huang et al. (2025) recommended to leverage this task to collect human data created before the popular use of chatbots or LLMs to guarantee genuine human authorship. Further, the high emotional loads and personal narratives elicited by CREFF's tasks act as an organic disincentive for AI reliance,[3] offering a highly reliable testbed for computational and forensic linguistics.

Consequently, CREFF stands as a competitive and versatile Spanish NLP resource suitable not only for foundational tasks such as sentiment analysis or emotion detection, but also for forensic and computational linguistic research. This structural versatility is of particular significance, since in the field of Spanish forensic linguistics, robust datasets are notably scarce and often restricted to single genre, as seen in Cremades (2016).

To further illustrate CREFF's unique positioning in the Spanish forensic landscape, **Table 2** provides an illustrative comparative overview against the main existing controlled resources previously mentioned in this section.

Additionally, the Spanish NLP community is on the trail of computational techniques and resources to tackle the challenges of MGT, as observed in shared tasks like AuTexTification (Sarvazyan et al., 2023) and, for Iberian languages, IberAuTexTification (Sarvazyan et al., 2024). In this context, the urgency and necessity of verified human baselines become paramount. To address this framework, the following section outlines an experimental pilot study focusing on a cross-domain authorship identification task.

4 Experimental Setup

To illustrate CREFF's capabilities, we conducted an author identification experiment using three main modelling approaches: TF-IDF classifiers, Transformer-based encoders (e.g., BERT-like architectures), and instruction-tuned LLMs.

While traditional NLP studies occasionally report results on raw data, forensic linguistics demonstrates that classification performance improves significantly when combining lexical features with explicit syntactic or psycholinguistic variables (Chaski, 2001; Stamatatos et al., 2001). For instance, a robust method for author identification is ALIAS SynAID (Chaski, 2001), which is linguistically sophisticated, as it analyses syntactic markedness applying Discriminant Function Analysis (DFA). Similarly, early statistical approaches established foundational computational baselines for stylistic differentiation (Baayen et al., 2002; Juola & Baayen, 2005).

Consequently, we deliberately withheld 17 unseen authors as a blind, gold-standard benchmark for a future shared task evaluation. Thus, for the purpose of this pilot study we focus on 376 verified authors, totalling 3,760 texts to address. To extract explicit

[3] As Huang et al. (2025) and Anderson et al. (2024) noted, the language generated by LLMs tends to be less emotional, more formal, and highly structured.

Dimension / Feature	Cremades (2016)	Almela et al. (2019)	CREFF (Ours)
Data Provenance	Social Media & Deep Web (Anonymous / Unverified)	Controlled Laboratory (Authenticated Humans)	Controlled Laboratory (Authenticated Humans)
Genre Variety	Mono-Genre (Suicidal & Deep Emotional Messages)	Multi-Genre (Chaski's WSD)	Multi-Genre (Chaski's WSD)
Sample Size	102 Texts (66 Deep Web, 36 Social Media)	106 Verified Authors (13 Inmates, 93 Students, 1,060 Texts)	393 Verified Authors (3,930 Texts)
Temporality	Synchronic (Static Point, 2016)	Synchronic (Static Point, 2019)	Diachronic (2019–2024)
Supported Tasks	Psychological Profiling (Deep Emotion & Suicidal Ideation)	Criminal Profiling (Abuser vs. Non-Abuser Differentiation)	Cross-Domain Authorship Identification, MGT Detection & Pragmatic-Emotional Forensic Analysis

Table 2: Illustrative comparison of CREFF against existing Spanish forensic linguistic resources.

psycholinguistic cues, all selected texts were processed using the Spanish dictionary of LIWC-22 (Pennebaker et al., 2001; Ramírez-Esparza et al., 2007) to capture lexical and emotional valence features, complemented by UMUTextStats (García-Díaz et al., 2022) to account for specialised Spanish morphosyntactic structures and complex verb tenses. However, these tools rely primarily on simple term-counting, without considering the context in which words occur. It may, therefore, overlook grammatical variation that is relevant to linguistic and forensic analyses.

Hence, the LIWC-22 and UMUTextStats variables used in this study are automatically extracted computational descriptors. They should not be interpreted as manually validated or gold-standard forensic annotations. The present experiment evaluates their usefulness as classification features.

The experiment is formulated as a *closed-set cross-domain authorship identification* task. Given a questioned text, the model must identify its author from a fixed set of 376 known candidate authors.

All candidate authors are represented in the training data. However, the two communicative scenarios used for final testing are excluded from both training and validation. The objective is therefore to determine whether the models can transfer author-related linguistic patterns across communicative contexts, rather than merely recognising scenario-specific topics or vocabulary.

4.1 Data and Splits

We used the complete balanced experimental subset of 376 authors. Each participant was assigned an Author ID. To set up a strict cross-domain evaluation scenario, we split the 10 available tasks by genre rather than splitting texts randomly.

For the training set, we selected a subset of 8 tasks including Influence Essay, Goal Essay, Anger Essay, Complaint Letter, Insurance Letter, Threat Letter to Known Person, Apology Letter, and Lover Letter (Tasks 2 to 9, totalling 3,008 excerpts).

For the evaluation set, we reserved two unseen text types (Trauma Narrative and Threat Letter to Public Unknown Person – texts from tasks 1 and 10, respectively) to test the ability of the models to generalise to new genres from the same authors (752 writing compositions). Additionally, we removed their author IDs to avoid potential bias, thereby forcing models to generalise in a fully blind evaluation based solely on linguistic information rather than on ID association.

The data were grouped by author and text type, and split using a Group Shuffle Split to ensure that no author's text types appeared in both training and validation sets. Specifically, 80% of data were used for training and 20% for validation.

This internal split prevents the same author–scenario instance from appearing in both training and validation. It does not remove authors from either stage, since the ex-

periment is a closed-set identification task in which the same candidate-author inventory must be learned and evaluated.

Table 3 summarises the experimental partitions. The final test partitions are balanced because each author contributes exactly one document to each unseen scenario.

Partition	Authors	Scenarios	Texts
Training	376	8 observed	2,406
Validation	376	8 observed	602
Test: Trauma Narrative	376	1 unseen	376
Test: Public Threat Letter	376	1 unseen	376

Table 3: Training, validation, and cross-domain test partitions. The final test documents were identical for all models.

4.2 Models

TF-IDF baselines. Traditional classifiers were trained on TF-IDF representation of words and character n-grams. We experimented with Linear SVM (Cortes & Vapnik, 1995), and Random Forest (RF; Breiman, 2001). Some variants also included additional linguistic features extracted by LIWC-22 (Pennebaker et al., 2001) and UMUTextStats (García-Díaz et al., 2022) concatenated to the TF-IDF vectors.

The word-level TF-IDF representation included unigrams and bigrams, with a minimum document frequency of 2, a maximum document frequency of 0.95, sublinear term-frequency scaling, and a maximum vocabulary size of 200,000 features. Word- and character-level representations were combined to capture lexical, orthographic, punctuation, and sublexical stylistic information.

The Linear SVM used balanced class weights and random seed 42. The Random Forest used 500 estimators, square-root feature sampling, balanced subsample class weighting, parallel execution, and random seed 42.

For the hybrid models, missing or non-finite linguistic values were replaced with zero. The linguistic features were aligned by author and text type and concatenated with the sparse TF-IDF vectors before classification.

Transformer-based encoder models. We fine-tuned different monolingual Transformer models pretrained on Spanish: BETO (Cañete et al., 2023), BERTIN (De la Rosa et al., 2022), ALBETO, and DistilBETO (Cañete et al., 2022), for sequence classification. In addition, a Longformer (Beltagy et al., 2020) variant was included to handle longer inputs. All models used cross-entropy loss with class weighting to compensate for author imbalance.

BETO, BERTIN, ALBETO, and DistilBETO used a maximum sequence length of 512 tokens. Longformer used a maximum sequence length of 4,096 tokens and assigned global attention to the initial classification token.

LLMs. We also trained different medium-sized LLMs for author prediction, including Gemma-3-1B (Gemma Team et al., 2025), LLaMA-3.2-1B, and LLaMA-3.2-3B (Grattafiori et al., 2024). Each author was represented by a special token USR_i, and the models were fine-tuned using supervised instruction-style prompts. Evaluation was performed by selecting the most probable USR_i token as the predicted author. The prompt structure followed an instruction–input–response format (see Annex A).

This setup allowed the models to leverage their instruction-following capabilities while remaining constrained to a discrete output space representing the set of known authors. During fine-tuning, all USR_i tokens were added to the tokeniser and model vocabulary to ensure direct mapping between predicted tokens and author identities.

During inference, the response label was omitted. The logits of the next token were then restricted to the 376 registered author tokens, and the token with the highest probability was selected as the predicted author.

Although these systems are generative language models, the evaluation is therefore more precisely described as constrained next-token classification over a closed inventory of author labels, rather than unrestricted text generation. This setup allows the models to produce an explicit author label and provides token-level scores over the complete set of candidate authors.

4.3 Training Details

For the traditional TF-IDF classifiers, models were trained on the corresponding training subset. Word- and character-level features were combined to capture lexical and stylistic information. The SVM and Random Forest configurations described in Section 4.2 were applied consistently across the baseline and

linguistic-feature variants.

For the Transformer-based encoders, BETO, BERTIN, ALBETO, and DistilBETO were trained for 20 epochs using a learning rate of 1×10^{-5}, a training batch size of 4, an evaluation batch size of 8, a weight decay of 0.01, and a warm-up ratio of 0.06. The maximum sequence length was 512 tokens.

Longformer was trained using a maximum sequence length of 4,096 tokens. Its definitive number of epochs, batch sizes, and learning rate are reported in Annex C according to the final run configuration.

Transformer models were evaluated after each epoch. The checkpoint obtaining the highest validation macro-F1 was retained. Dynamic padding was applied within each batch, and mixed-precision training was used when supported by the available hardware.

For the instruction-tuned LLMs, training was conducted using supervised fine-tuning. The models were trained for one epoch with a per-device training batch size of 2, an evaluation batch size of 2, two gradient-accumulation steps, a learning rate of 2×10^{-5}, a weight decay of 0.05, 100 warm-up steps, and a maximum sequence length of 4,092 tokens. Gradient checkpointing and bfloat16 precision were enabled (see Annex C which illustrates all the principal hyperparameters).

4.4 Evaluation Protocol and Metrics

Evaluation was performed exclusively on the Trauma Narrative and Threat Letter to Public Unknown Person domains, which were unseen during training. Their inclusion allows for a fully blind author identification setting, and testing how well models generalise to ethically and stylistically challenging material independently of topic or genre. We report top-1 accuracy and macro-averaged F_1 score for every architecture (see Annex B for formal definitions). Both test sets are perfectly balanced, containing exactly 376 instances matching the candidate author inventory.

A prediction was considered correct when the author assigned by the model matched the anonymised author identifier of the test document. Since each test domain contains one document from each of the 376 candidate authors, both test sets are balanced and contain 376 instances.

The same test documents and author-label mapping were used for all models, allowing a direct comparison between traditional classifiers, Transformer encoders, and instruction-tuned language models.

4.5 Results and Discussion

Table 4 presents the performance of all evaluated models on the CREFF test set, which includes the two unseen text types. The reported metric is accuracy (%), computed separately for each column.

4.5.1 Overall Performance

Results show a clear performance gap between traditional TF-IDF models and deep learning approaches when evaluated in a completely unseen domain.

The highest accuracy for Trauma Narrative was obtained by the TF-IDF + SVM model, with 15.16%. The hybrid TF-IDF + UMUTextStats + SVM model produced a similar result of 14.36%. In the Threat Letter to a Public Unknown Person domain, the best-performing system was TF-IDF + UMUTextStats + SVM, which achieved 27.39%, followed by the standard TF-IDF + SVM model with 23.40%.

Although these values are modest in absolute terms, they are significantly higher than those obtained by any Transformer-based encoder or LLM, so it shows that shallow lexical-stylistic representations retain limited but consistent authorial information even across domains.

Ultimately, these empirical limitations underscore that the modest baseline performance is not due to model limitations, but a reflection of the inherent complexity of testing stylistic generalisation across unseen forensic domains. Additionally, the diachronic nature of CREFF (compiled between 2019 and 2024 with a different student pool for each year) introduces a key valuable challenge for testing the robustness of NLP benchmarks against shifting linguistic trends and generational slang across different subject groups over time. This dynamic temporal layer establishes a more realistic and demanding testing ground than the static, synchronous setups characteristic of previous Spanish forensic resources.

4.6 Classic vs. Neural Encoders

The Transformer-based encoders (BETO, BERTIN, ALBETO, DistilBETO, and Long-

Model	Trauma Narrative Accuracy (%)	Threat Letter Accuracy (%)
TF-IDF SVM	**15.16**	23.40
TF-IDF RF	3.46	12.50
TF-IDF SVM + linguistic features (UMUTextStats)	14.36	**27.39**
TF-IDF RF + linguistic features (UMUTextStats)	1.60	2.66
TF-IDF SVM + linguistic features (LIWC-22)	11.97	23.14
TF-IDF RF + linguistic features (LIWC-22)	5.59	11.17
BETO	4.52	8.24
BERTIN	5.05	10.90
ALBETO	0.53	1.06
DistilBETO	2.39	5.58
Longformer	4.26	10.64
Gemma-3-1B	0.27	0.27
Llama-3.2-1B	0.27	0.27
Llama-3.2-3B	0.00	0.27

Table 4: Performance of all models on the CREFF test set using unseen text types (Trauma Narrative and Threat Letter to Public Unknown Person). The table reports accuracy (%) for each column, showing the difficulty of cross-domain author identification when the test texts come from sensitive and stylistically distinct domains.

former) showed low cross-domain generalisation, with accuracies between 0.5% and 10.9%. Their performance was considerably below that of the TF-IDF models, indicating that while contextual embeddings capture richer semantics, they are less robust to domain and topic shifts in author identification.

Among them, BERTIN and Longformer performed best, reaching 5.05% on Trauma Narrative and 10.90% on Threat Letter to Public Unknown Person. Longformer obtained similar results, with 4.26% and 10.64%, respectively. BETO reached 4.52% and 8.24%, whereas DistilBETO and ALBETO obtained lower accuracies. These results suggest that, under the present configuration, contextual representations were less robust than sparse lexical and stylistic representations when transferred to unseen communicative domains. Transformer encoders are primarily designed to capture contextual and semantic information, which may facilitate topic modelling and text classification but does not necessarily provide an optimal representation of stable author-specific stylistic variation.

The relatively small number of training documents per author may also have limited the neural models. Each encoder had to learn a 376-class classification problem from only a few examples for each author. As a result, the models may have learned semantic or scenario-related patterns that did not transfer reliably to Trauma Narratives and public Threat Letters.

Longformer was included to determine whether a longer context window could improve author identification by retaining more information from each document. However, its results were comparable to those of BERTIN. This indicates that sequence truncation alone does not explain the lower performance of the neural models.

4.7 Effect of Linguistic Features

The inclusion of both LIWC-22 and UMUTextStats (lexical and syntactic metrics such as word length, vocabulary richness, and sentence complexity) enhanced classification accuracy. Especially evident for the SVM classifier in the Threat Letter domain, where accuracy rose from 23.4% (TF-IDF alone) to 27.4% when augmented with UMUTextStats.

For Trauma Narrative, however, TF-IDF SVM achieved 15.16%, while the addition of UMUTextStats resulted in a slightly lower

accuracy of 14.36%. Therefore, the contribution of explicit linguistic features was not uniform across the two domains.

The TF-IDF SVM model combined with LIWC-22 reached 11.97% on Trauma Narrative and 23.14% on Threat Letter to Public Unknown Person. Although this configuration outperformed the neural encoders and LLMs, it did not surpass the standard TF-IDF SVM model or the UMUTextStats-enhanced configuration.

One possible explanation is that UMUTextStats includes a broader set of Spanish-specific grammatical, lexical, syntactic, and readability variables. These features may capture aspects of linguistic variation not represented explicitly by TF-IDF, particularly verb-tense usage, sentence structure, punctuation, and vocabulary richness.

By contrast, LIWC-22 is primarily based on dictionary-category frequencies. While these variables can represent emotional, cognitive, and thematic dimensions, their reliance on term counting may provide less detailed information about grammatical structure and language-specific stylistic variation.

However, the results do not demonstrate that UMUTextStats features are universally more effective than LIWC-22. Their advantage was mainly observed with the linear SVM in the public Threat Letter domain. In addition, the corresponding Random Forest models obtained considerably lower results. TF-IDF RF combined with UMUTextStats reached only 1.60% and 2.66%, whereas TF-IDF RF combined with LIWC-22 obtained 5.59% and 11.17%.

4.8 Large Language Models (LLMs)

The instruction-tuned language models obtained the lowest accuracies among all evaluated systems. Gemma-3-1B and Llama-3.2-1B achieved 0.27% in both test domains, while Llama-3.2-3B obtained 0.00% on Trauma Narrative and 0.27% on Threat Letter to Public Unknown Person. These values resemble random baseline for a closed candidate set of 376 authors, demonstrating that generative "black-box" architectures in this configuration failed to map stable stylistic signatures across domains, reinforcing the need for Explainable AI (XAI) for legal scenarios (Rudin, 2019; Billah, 2025).

As formalised in Annex B, this near-chance performance is primarily attributed to three interrelated factors within our constrained decoding framework: (1) severe data scarcity (mapping a massive 376-class problem from limited documents per candidate); (2) the absence of prior semantic features in the arbitrary special tokens (`<USR_i>`) used as author labels; and (3) the restrictive nature of next-token optimisation, which forces the models to select a single token from a fixed inventory rather than leveraging free-form instruction following or abstracting idiosyncratic stylistic cues.

To examine this performance collapse, an analysis of the predicted label distributions revealed severe mode collapse (see Annex D). Gemma-3-1B clustered its predictions into just 5 distinct labels for the Trauma Narrative, while LLaMA-3.2-1B collapsed entirely onto a single label (100% frequency), accounting for 376 of the predictions. Rather than distributing decisions across the complete author inventory, optimisation converged towards a limited subset of labels that received disproportionately high next-token scores across many unrelated inputs. This behaviour demonstrates that constrained token spaces cause generative models to resort to majority-class or arbitrary label collapses when exposed to highly emotional, out-of-domain forensic writing.

The analysis does not establish whether this collapse was caused by model size, label-token initialisation, the amount of supervision, optimisation dynamics, prompt design, or the constrained next-token formulation. Isolating these factors would require controlled ablation experiments.

5 Conclusions

This paper has presented CREFF, a versatile, high-quality linguistic resource designed to overcome the limitations of web-scraped and unverified text datasets in Spanish NLP. By providing a strictly controlled and balanced human-authored baseline, CREFF establishes a reliable foundation for forensic and computational linguistics.

To demonstrate its capabilities, we conducted a pilot study of cross-domain authorship identification. We combined classic machine learning, transformer-based encoders, and LLMs along with psycholinguistic features extracted from the Spanish version of LIWC and UMUTextStats. This hybrid ap-

proach shows that classic models combined with explicit linguistic features not only enhance classification accuracy in unseen genres but also offer the explainability required in judicial proceedings.

Following Open Science principles, the public portion of the CREFF corpus (376 authors, comprising approximately 95% of the total dataset) will be made available upon publication in open repositories such as Zenodo and OSF under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International license (CC BY-NC-SA 4.0). A subset of 17 authors (5%) has been deliberately withheld to serve as a restricted, blind gold-standard benchmark for a future shared task evaluation.[4]

6 Limitations and Future Work

A current limitation of this pilot study is that CREFF is largely unannotated at a granular level. Forensic research requires comprehensive tracking of specific language patterns across disputed authors. While the current baseline relies on automated tool extraction, the lack of deeply layered manual annotation may constrain the performance.

Additionally, the rigorous control applied to sociolinguistic variables may limit the generalisation of our results, as noted by Moon et al. (2025). This control ended in a homogeneous participant pool, which could limit immediate generalisation to broader, more diverse demographic groups. Finally, as Huang et al. (2024) warned, the scalability remains a great limitation in forensic practice. The vast number of disputed authors in real-world scenarios may significantly hinder the effectiveness of the closed-set classification method.

Despite these constraints, this pilot study lays the foundation for an ongoing project. Our immediate future work focuses on validating a methodology based on an integrative taxonomy that entails both the pragmatic dimensions of the Speech Act Theory (Austin, 1962; Searle, 1979) and emotional categories derived from constructivist psychology (Barrett, 2017). Based on Huang et al. (2025) suggestions, we intend to integrate a multidisciplinary benchmark methodology (i.e., uniting linguistics, computer science, forensic science, and psychology) that reflects real-world scenarios through a broad range of written human-authored data.

Currently, we are automating corpus segmentation by leveraging AI for initial parsing, followed by a rigorous human-in-the-loop revision. This is intended to capture the granularity of linguistic and lexical choices people use in different communicative contexts and how they construct their linguistic, cognitive, and emotional identities. We aim to validate our methodology by testing it on established datasets, such as the MEACorpus from the EmoSPeech shared task at IberLEF (Pan et al., 2024), shifting our features from authorship identification toward emotion detection and sentiment analysis tasks.

Additionally, taking advantage of the temporal compilation process of CREFF (2019-2024), we propose a diachronic comparative study to analyse potential shifts in human writing style between the pre-AI and post-AI collected texts. Specifically, we aim to adapt the empirical framework of Moon et al. (2025) regarding the potential homogenising effects of LLMs on writing diversity within our own human-authored corpus. While Moon et al. (2025) focused on general lexical and creative variance, our future work will utilise our custom pragmatic-emotional taxonomy to quantify style flattening. By tracking shifts in speech act density and emotional granularity across the pre- and post-AI approaches, we seek to investigate how widespread public exposure to generative writing tools affects individual human stylistic variance.

7 Ethical Considerations

The protocols guiding CREFF compilation strictly adhere to the ethical standards of Chaski's Writing Sample Database (Chaski, 2001), which received institutional approval from an independent interdisciplinary board. For the Spanish repository, compliance was fully verified and approved by the equivalent institutional ethics committee (Almela et al., 2019). All data collection prioritised absolute participant anonymity, confidentiality, legal data governance, and restricted access, mitigating any risks of privacy violation while preserving the structural integrity required for forensic research replication.

[4]The CREFF corpus splits used for this study for replicability are available at: `https://doi.org/10.5281/zenodo.20584621`

Acknowledgments

This work is part of the research project LaTe4PoliticES (PID2022-138099OB-I00) funded by MICIU/AEI/10.13039/501100011033 and the European Regional Development Fund (ERDF/EU – FEDER/UE)-a way of making Europe. Ms. Desirée Martínez Carrión is supported by University of Murcia through the predoctoral programme.

References

Almela, Á., G. Alcaraz-Mármol, Á. García-Pinar, y C. Pallejá. 2019. Developing and analyzing a spanish corpus for forensic purposes. *Linguistic Evidence in Security, Law and Intelligence*, 3(0):1–13.

Amigó, E., J. Carrillo-de Albornoz, A. Fernández, J. Gonzalo, G. Marco, R. Morante, J. Pedrosa, y L. Plaza. 2024. A web portal for the state of the art of nlp tasks in spanish. En *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, páginas 2028–2038.

Anderson, B. R., J. H. Shah, y M. Kreminski. 2024. Homogenization effects of large language models on human creative ideation. En *Creativity and Cognition*, páginas 413–425.

Austin, L. 1962. *How to Do Things with Words*. Oxford University Press.

Baayen, H., H. Halteren, A. Neijt, y F. Tweedie. 2002. An experiment in authorship attribution. En *Proceedings of JADT 2002*, páginas 29–37.

Bai, Y., A. Jones, K. Ndousse, A. Askell, A. Chen, N. DasSarma, D. Drain, S. Fort, D. Ganguli, T. Henighan, y others. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Barrett, L. F. 2017. *La Vida Secreta del Cerebro. Cómo se Construyen las Emociones*. Paidós.

Beltagy, I., M. E. Peters, y A. Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Billah, M. 2025. Explainable ai for digital forensics: Ensuring transparency in legal evidence analysis. *Journal of Forensic Science and Research*, 9(2):109–116.

Breiman, L. 2001. Random forests. *Machine Learning*, 45(1):5–32.

Brennan, M., S. Afroz, y R. Greenstadt. 2012. Adversarial stylometry: Circumventing authorship recognition to preserve privacy and anonymity. *ACM Transactions on Information and System Security (TISSEC)*, 15(3):1–22.

Cañete, J., G. Chaperon, R. Fuentes, J. H. Ho, H. Kang, y J. Pérez. 2023. Spanish pre-trained bert model and evaluation data.

Cañete, J., S. Donoso, F. Bravo-Marquez, A. Carvallo, y V. Araujo. 2022. Albeto and distilbeto: Lightweight spanish language models. En *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, páginas 4291–4298. European Language Resources Association.

Chaski, C. E. 1997. Who wrote it? steps toward a science of authorship identification. En *National Institute of Justice*, volumen 233, páginas 15–22.

Chaski, C. E. 2001. Empirical evaluations of language-based author identification techniques. *Forensic Linguistics*, 8(1):1–65.

Chaski, C. E. 2005. Who's at the keyboard? authorship attribution in digital evidence investigations. *International Journal of Digital Evidence*, 4(1):1–13.

Chaski, C. E. 2007. The keyboard dilemma and authorship identification. En *Advances in Digital Forensics III*, volumen 242, páginas 133–146. Springer.

Cortes, C. y V. Vapnik. 1995. Support-vector networks. *Machine Learning*, 20(3):273–297.

Cremades, S. Z. 2016. Corpus de mensajes suicidas y emociones profundas en las redes sociales y deep web. Master's thesis, Universidad de Alicante.

De la Rosa, J., E. G. Ponferrada, P. Villegas, P. G. d. P. Salas, M. Romero, y M. Grandury. 2022. Bertin: Efficient pretraining of a spanish language model using perplexity sampling.

García-Díaz, J. A., P. J. Vivancos-Vicente, Á. Almela, y R. Valencia-García. 2022. Umutextstats: A linguistic feature extraction tool for spanish. En *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*, páginas 6035–6044.

Gemma Team, A. Kamath, J. Ferret, S. Pathak, N. Vieillard, R. Merhej, S. Perrin, T. Matejovicova, A. Ramé, M. Rivière, L. Rouillard, T. Mesnard, G. Cideron, J. Grill, S. Ramos, E. Yvinec, M. Casbon, E. Pot, I. Penchev, y L. Hussenot. 2025. Gemma 3 technical report. Informe técnico, Google.

Grant, T. y N. MacLeod. 2020. *Language and Online Identities: The Undercover Policing of Internet Sexual Crime*. Cambridge University Press, 1.a edición.

Grattafiori, A., A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, A. Yang, A. Fan, A. Goyal, A. Hartshorn, A. Yang, A. Mitra, A. Sravankumar, A. Korenev, A. Hinsvark, y Z. Ma. 2024. The llama 3 herd of models.

Huang, B., C. Chen, y K. Shu. 2024. Can large language models identify authorship? En *Findings of the Association for Computational Linguistics: EMNLP 2024*, páginas 445–460.

Huang, B., C. Chen, y K. Shu. 2025. Authorship attribution in the era of llms: Problems, methodologies, and challenges. *ACM SKIGKDD Explorations Newsletter*, 26(2):21–43.

Juola, P. 2006. Authorship attribution. *Foundations and Trends in Information Retrieval*, 1(3):233–334.

Juola, P. y H. Baayen. 2005. A controlled-corpus experiment in authorship identification by cross-entropy. *Literary and Linguistic Computing*, 20(1):59–67.

Li, J., R. Zheng, y H. Chen. 2006. From fingerprint to writeprint. *Communications of the ACM*, 49(4):76–82.

McEnery, T. 2022. Corpora. En R. Mitkov, editor, *The Oxford Handbook of Computational Linguistics*. Oxford University Press, 2nd edición, páginas 494–507.

Moon, K., A. E. Green, y K. Kushlev. 2025. Homogenizing effect of large language models (llms) on creative diversity: An empirical comparison of human and chatgpt writing. *Computers in Human Behavior: Artificial Humans*, 6:1–10.

Nini, A. 2023. *A Theory of Linguistic Individuality for Authorship Analysis (Elements in Forensic Linguistics)*. Cambridge University Press.

Ouyang, L., J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. F. Christiano, J. Leike, y R. Lowe. 2022. Training language models to follow instructions with human feedback. En *Advances in Neural Information Processing Systems*, volumen 35, páginas 27730–27744.

Padmakumar, V. y H. He. 2024. Does writing with language models reduce content diversity? En *ICLR 2024*, páginas 1–28.

Pan, R., J. García-Díaz, M. A. Rondríguez-García, F. García-Sánchez, y R. Valencia-García. 2024. Overview of emospeech at iberlef 2024: Multimodal speech-text emotion recognition in spanish. *Procesamiento de Lenguaje Natural*, (73):359–368.

Pennebaker, J. W., M. E. Francis, y R. J. Booth. 2001. *Linguistic Inquiry and Word Count (LIWC): LIWC 2001*. Erlbaum Associates.

Ramírez-Esparza, N., J. W. Pennebaker, F. Andrea García, y R. Suriá. 2007. La psicología del uso de las palabras: Un programa de computadora que analiza textos en español. *Revista Mexicana de Psicología*, 24(1):85–99.

Rudin, C. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215.

Sarvazyan, A. M., J. Á. González, F. Rangel, P. Rosso, y M. Franco-Salvador. 2024. Overview of iberautextification at iberlef 2024: Detection and attribution of machine-generated text on languages of

the iberian peninsula. *Procesamiento del Lenguaje Natural*, 73:421–434.

Sarvazyan, A. M., F. Rangel, B. Chulvi, y P. Rosso. 2023. Overview of autextification at iberlef 2023: Detection and attribution of machine-generated text in multiple domains. *arXiv preprint arXiv:2309.11285*.

Searle, J. R. 1979. *Expression and Meaning: Studies in the Theory of Speech Acts*. Cambridge University Press.

Stamatatos, E., N. Fakotakis, y G. Kokkinakis. 2001. Computer-based authorship attribution without lexical measures. *Computers and the Humanities*, 35(2):193–214.

A LLM Prompt Structure

The instruction-tuned models (Gemma and LLaMA) were evaluated using a standardised supervised instruction-style template. The prompt structure followed a strict instruction–input–response format as detailed below:

```
### Question:
Identify the author of the following text.
Answer ONLY with their tag.

### Input:
Reference text:
[DOCUMENT]

### Response:
<USR_i>
```

B Evaluation Metrics Formalisation

To ensure mathematical replicability, the top-1 classification performance for the closed-set cross-domain authorship identification task is formalised using standard Accuracy, defined as:

$$\text{Accuracy} = \frac{\text{N of correctly identified texts}}{\text{total N of test texts}}. \quad (1)$$

C Model Hyperparameters Configuration

To ensure exact training replicability, Table 5 details the primary hyperparameters utilised during the fine-tuning of both the Transformer encoders and the instruction-tuned Large Language Models.

Model	Epochs	Batch Size	LR	Max. Len
BETO	20	4 (Train) / 8 (Eval)	1×10^{-5}	512
BERTIN	20	4 (Train) / 8 (Eval)	1×10^{-5}	512
ALBETO	20	4 (Train) / 8 (Eval)	1×10^{-5}	512
DistilBETO	20	4 (Train) / 8 (Eval)	1×10^{-5}	512
Longformer	30	4 (Train) / 8 (Eval)	1×10^{-5}	4,096
Gemma-3-1B	1	2 (Train) / 2 (Eval)	2×10^{-5}	4,092
Llama-3.2-1B	1	2 (Train) / 2 (Eval)	2×10^{-5}	4,092
Llama-3.2-3B	1	2 (Train) / 2 (Eval)	2×10^{-5}	4,092

Table 5: Main hyperparameters used to fine-tune the Transformer encoders and instruction-tuned language models. Train BS denotes the per-device training batch size, Eval BS the per-device evaluation batch size, LR the learning rate, and Max. length the maximum input sequence length. The instruction-tuned models used two gradient-accumulation steps, resulting in an effective training batch size of 4.

D Distributional Error Analysis

This table provides the empirical distribution of the author labels predicted by the generative models, illustrating the label concentration effect discussed in Section 4.8.

Model	Test Domain	Dist.	Freq.	Freq. (%)
Gemma-3-1B	Trauma Narrative	5	367	97.61
Gemma-3-1B	Public Threat Letter	6	366	97.34
Llama-3.2-1B	Trauma Narrative	1	376	100.00
Llama-3.2-1B	Public Threat Letter	4	370	98.40
Llama-3.2-3B	Trauma Narrative	7	202	53.72
Llama-3.2-3B	Public Threat Letter	8	272	72.34

Table 6: Distributional analysis of predicted author labels across the 376 test documents.