

Automated Classification of Tax Service Requests Using Machine Learning and Natural Language Processing

David Ruete

Faculty of Engineering
Universidad Andrés Bello
Santiago, Chile
druete@unab.cl

Iván Fredes-Contreras

Faculty of Engineering
Universidad Andrés Bello
Santiago, Chile
i.contrerasfredes@uandresbello.edu

Omar Salinas

Faculty of Engineering
Universidad Andrés Bello
Santiago, Chile
omar.salinas@unab.cl

Alejandro Caroca

Faculty of Engineering
Universidad Andrés Bello
Santiago, Chile
acaroca@unab.cl

Lilian San Martin Medina

Faculty of Engineering
Universidad Andrés Bello
Santiago, Chile
lsanmartin@unab.cl

Carla Taramasco

Faculty of Engineering
Universidad Andrés Bello
Santiago, Chile
carla.taramasco@unab.cl

Jean Maidana

Faculty of Engineering
Universidad Andrés Bello
Santiago, Chile
jean.maidana@unab.cl

Abstract—Classifying citizen service requests in public tax offices is harder than it looks. Citizens describe their problems in their own words, select the wrong service category more often than not, and force staff to reclassify every submission before routing can begin. This paper examines whether that reclassification burden can be automated using NLP and supervised machine learning, drawing on 4,021 real requests submitted to Chile’s Taxpayer Ombudsman Office (DEDECON) in Spanish free text. The feature engineering strategy combines TF-IDF text representation with two domain-grounded signals: a proxy for the user’s knowledge level (scored 0 to 2) and binary indicators of class-specific keyword presence. Six classifiers were evaluated under identical conditions: Logistic Regression, Support Vector Machines, Naive Bayes, Decision Trees, Random Forest, and K-Nearest Neighbors. Random Forest reached 92.18% accuracy and an F1-macro score of 0.921, competitive with transformer-based approaches at a fraction of the computational cost. The main source of residual error is semantic overlap between adjacent service categories, a problem that text features alone are unlikely to resolve. Beyond DEDECON, the pipeline is directly applicable to other Spanish-language public service institutions facing the same intake classification problem.

Index Terms—text classification, natural language processing, supervised learning, tax administration, Spanish language processing

I. INTRODUCTION

Tax administrations exercise broad investigative and enforcement powers, while taxpayers, especially individuals and micro, small, and medium-sized enterprises (MSMEs), often face high information and transaction costs when navigating complex rules, audits, collections, and appeals. In response, a growing set of taxpayer-rights institutions has emerged internationally to provide accessible, independent assistance, help prevent abuses, and promote procedural fairness. In Latin America, Mexico’s *Procuraduría de la Defensa del Contribuyente* (PRODECON) is a prominent precedent: a

specialized public body that offers free taxpayer support and related services, aimed at safeguarding taxpayer rights and reducing frictions in interactions with federal tax authorities (as reflected in official institutional descriptions and service statistics covering operations since at least 2011) [1].

Building on these governance and service-delivery trends, Chile created the *Defensoría del Contribuyente* (DEDECON) under Law No. 21.210 (*Moderniza la Legislación Tributaria*); specifically, Article Twenty-Third establishes DEDECON as a decentralized public service with legal personality and its own assets, under the supervision of the President of the Republic through the Ministry of Finance, and mandates a special focus on protecting the rights of vulnerable taxpayers [2], [3]. The Ministry of Finance publicly announced the start of DEDECON’s implementation in November 2021, linked to the designation of the first *Defensor Nacional*, highlighting the new institution’s role in protecting taxpayer rights and providing free orientation and mediation-type support [4]. In operational terms, DEDECON positions itself as a free, citizen-facing service that supports taxpayers in resolving conflicts and specific problems with the Internal Revenue Service (SII), Customs, and the Treasury (TGR), complemented by tax education initiatives [3], [5].

DEDECON’s main intake channel is digital: citizens submit requests through a web form in which they describe their tax issue in free-form Spanish text and self-select the service category they believe is appropriate. In practice, the form routes requests into eight institutional categories, *General Orientation*, *Specific Orientation*, *Tax Reports*, *Administrative Representation*, *Sub-Administrative Representation*, *Mediation*, *Complaints*, and *Tax Education*, and also allows a frequent fallback choice (“I don’t know / I’m not sure”). This design supports accessibility, but it creates a significant

operational challenge: many submissions are ambiguous or incorrectly categorized, forcing mandatory manual review and reclassification by specialized staff before cases can be routed and handled.

This bottleneck is not unique to DEDECON. Across public service environments, citizen-submitted requests tend to be brief, heterogeneous, and linguistically irregular, which makes accurate routing inherently difficult. When requests are misrouted, delays accumulate, service quality declines, and the timely resolution of citizens' needs suffers. For this reason, recent studies in government and municipal service delivery have increasingly examined automatic text classification as a decision-support mechanism for triage and case routing. Existing work includes deep-learning models for citizen transportation requests, hybrid neural architectures for public-service request texts, and more recent attention-based approaches that exploit external knowledge and label semantics to improve classification and routing accuracy [6], [7], [8]. In parallel, tax administrations worldwide are accelerating digital transformation and the use of AI for service provision. OECD evidence shows a sharp rise in the adoption of AI in tax administrations over recent years and highlights use cases such as virtual assistants and AI support for taxpayer interactions, including suggesting responses for incoming correspondence [9], [10]. At the same time, successful deployment in public services depends on citizen acceptance and perceived legitimacy, reinforcing the need for accurate, robust, and transparent automation that supports, rather than replaces, accountable decision-making [11].

Despite the rapid progress of NLP and supervised learning for text classification, tax-domain request triage raises distinctive challenges. First, DEDECON requests are written in Spanish and include specialized legal and administrative vocabulary; although Spanish-language pretrained language models and resources have grown substantially, Spanish remains less represented than English in many benchmark-driven NLP pipelines [12], [13], [14]. Second, DEDECON categories exhibit genuine semantic overlap, for example, between general and specific orientation, citizen narratives vary widely in structure and tax literacy, and real-world datasets show class imbalance. Many taxpayers are unfamiliar with the institutional distinctions between service categories; others face situations that span two or more of them; and the categories themselves do not always map cleanly onto the complexity of real tax disputes. These characteristics make it critical to evaluate classifiers that can perform well under realistic public-service constraints.

This paper addresses these needs by developing and evaluating an automated classification system tailored to Spanish-language tax service requests in DEDECON. Using a real-world dataset of 4,021 anonymized tickets collected between 2021 and 2024, we compare six supervised learning algorithms and propose a feature-engineering strategy that combines TF-IDF text representations with domain-specific attributes (including proxies for user knowledge level and class-specific keyword signals). Our best-performing configuration (Random

Forest) achieves 92.18% accuracy and a macro-F1 score of 0.921, showing that strong performance is attainable with computationally efficient methods suited to public-sector deployment.

Section II reviews related work on text classification for public services, tax administration digital transformation, and Spanish NLP. Section III formalizes the DEDECON request-routing problem and describes the dataset. The methodology, including preprocessing, feature engineering, class-imbalance handling, and evaluated algorithms, is detailed in Section IV. Section V describes the experimental setup, Section VI reports and discusses results, including error patterns linked to overlapping categories, and Section VII concludes with implications for DEDECON and directions for future research.

II. RELATED WORK

A. Text Classification for Triage and Routing in Public Services

The expansion of digital-government channels has increased the volume of citizen-submitted messages that must be processed, routed, and answered under institutional time and staffing constraints. The United Nations E-Government Survey highlights the continued global growth of online service delivery and the strategic importance of strengthening digital public services, which in turn amplifies the operational relevance of automated triage for incoming requests [15].

Within this context, recent research has examined automatic text classification as a decision-support tool for routing and prioritizing citizen requests. A recurring finding is that citizens often select labels that do not match the content or urgency of their message, requiring costly manual correction. Li et al. propose PeciDRL, explicitly framing petition “mislabelling and ambiguous labels” as a core problem in government petition platforms and showing how automated label correction can improve downstream handling of large-scale requests [16]. Complementary work explores model architectures tailored to noisy, short, and irregular public-service texts. Xiong et al. propose a BERT-BiLSTM-CNN feature-fusion approach for public service request classification, combining contextual embeddings with sequential and local pattern extraction [7]. Zhang et al. show that incorporating external knowledge and label information improves routing accuracy on large citizen-request datasets, treating citizen requests as a short-text classification problem with domain-specific entities and departmental labels [8].

B. AI in Tax Administration and Tax-Domain Analytics

Parallel to advances in public-service NLP, tax administrations are accelerating digital transformation and adopting AI for both internal efficiency and taxpayer services. OECD comparative evidence documents rapid growth in AI use across tax administrations (based on ISORA data), alongside increased deployment of service technologies such as virtual assistants and AI support during taxpayer interactions [17]. OECD reporting on digitalisation initiatives (ITTI) similarly describes widespread API implementation and digital service reforms,

with AI increasingly used for compliance management and service provision [18].

In terms of the research landscape, tax analytics has been heavily oriented toward risk, evasion, and fraud detection, driven by transactional, registry, and network data, rather than the semantic routing of taxpayer service requests written in natural language. A recent survey in *Engineering* synthesizes worldwide work on tax risk detection methods and their underlying data-mining paradigms, illustrating how much more developed risk-detection research is relative to taxpayer-facing request triage [19]. For citizen-facing deployments, however, technical performance alone is insufficient: evidence from public-administration settings indicates that citizen acceptance is a key determinant of whether AI-enabled public services are perceived as legitimate and adopted in practice [11]. Solutions that are accurate but operationally burdensome or opaque are unlikely to gain institutional traction.

C. Spanish NLP and Efficient Spanish Language Models

Spanish-language NLP for public institutions has benefited from rapid progress in pretrained language models and evaluation resources since 2022. In the Spanish PLM ecosystem, MarIA introduces a family of Spanish language models and associated resources and reports strong performance across multiple NLU tasks [20]. BERTIN presents a data-centric strategy for efficient Spanish model pretraining using “perplexity sampling,” aiming to reduce training cost while maintaining competitive downstream performance [21]. Cañete et al. release lightweight Spanish transformer variants (ALBETO and DistilBETO) that explicitly target reduced inference and training costs while retaining competitive performance on Spanish NLU benchmarks [22]. To consolidate and navigate the growing Spanish evaluation landscape, Amigó et al. provide a curated portal summarizing tasks, datasets, and competitions for NLP in Spanish, supporting more systematic benchmarking and comparability [23].

These resources are particularly relevant for Spanish public-service deployments, where domain-specific vocabulary, heterogeneous writing styles, and institutional constraints may favour careful trade-offs between state-of-the-art architectures and computational practicality. Broad NLP surveys also note that effectiveness depends not only on model choice but on data quality, domain shift, and evaluation design, all of which are critical in institutional routing settings [12].

D. Feature Engineering and Practical Baselines

Despite the success of transformer models, multiple recent studies confirm that classical representations and lightweight pipelines remain competitive for routing and ticketing tasks, often offering better transparency and lower operational cost. Arslan et al. evaluate different vectorization and preprocessing choices for real support-ticket classification and show that TF/TF-IDF-style representations remain strong baselines in practical routing scenarios [24]. This finding has direct implications for resource-constrained public institutions that cannot sustain GPU infrastructure for inference.

What this body of work leaves largely unaddressed is the specific challenge of routing Spanish-language tax service requests in a small public agency. Mislabelling and ambiguity are well-documented problems [16]; architectures that exploit label semantics and short-text structure have shown accuracy gains [7], [8]; and the Spanish NLP ecosystem now includes efficient pretrained models competitive on institutional tasks [20]–[23]. The present work draws on these threads while keeping the focus on deployability and the operational realities of an agency with limited computational resources, combining TF-IDF with domain-informed features to improve triage accuracy without sacrificing tractability.

III. PROBLEM FORMULATION AND DATASET

A. Institutional Context

DEDECON categorizes incoming service requests into nine classes based on the service ultimately provided (after staff review):

- 1) **General Orientation:** Broad inquiries about tax obligations, procedures, or rights without a specific case context. Example: “I need to know what taxes I must pay as a freelancer.”
- 2) **Specific Orientation:** Questions tied to a particular individual situation or ongoing case, where the answer requires personalized guidance rather than general information. Example: “I received Form 29 with an amount I don’t understand; can you explain what it means for my specific business?”
- 3) **Tax Reports:** Requests for formal technical opinions on tax law interpretation, procedural questions, or potential rights violations. Example: “I need an official opinion on whether SII’s interpretation of Article 17 applies to my case.”
- 4) **Administrative Representation:** Legal support in administrative procedures before tax authorities, including filing appeals (Recurso de Reposición Administrativa, RAF). Example: “I need help filing an administrative appeal against my tax assessment.”
- 5) **Sub-Administrative Representation:** Specialized representation in specific procedures such as unlocking tax IDs, resguardo processes, or hierarchical appeals. Example: “My tax ID was blocked and I need assistance submitting the unlocking request.”
- 6) **Mediation:** Conflict resolution between taxpayer and tax authority, typically involving recalculations, payment disputes, or reappraisals. Example: “SII recalculated my property tax and I disagree with the new amount; I need mediation.”
- 7) **Complaints:** Reports of improper treatment, rights violations, or inappropriate conduct by tax officials. Example: “A tax official was rude and refused to provide information I am entitled to receive.”
- 8) **Tax Education:** Requests for learning resources, workshops, or guidance on tax formalization for new entrepreneurs. Example: “I’m starting a small business

and need to understand my tax obligations from the beginning.”

- 9) **Not Applicable:** Requests outside DEDECON’s scope, including matters unrelated to taxation or pertaining to other institutions. Example: “I need help with my retirement pension calculation.”

The key operational problem is that citizens’ initial self-classification frequently diverges from the actual service provided after expert review. Taxpayers often lack familiarity with the institutional distinctions between service categories; some situations genuinely straddle two categories, particularly the boundary between General and Specific Orientation; and complex tax disputes rarely fit neatly into a single predefined option. Automating this classification would allow the system to suggest an appropriate service category during form submission, reducing routing errors before they reach the review queue.

B. Dataset Characteristics

The dataset was provided by DEDECON’s Planning and Studies Division, containing 4,021 real service requests submitted between 2021-2024. Each record includes:

- **Ticket ID:** Unique identifier
- **Request Date:** Timestamp of submission
- **Requested Service:** User’s self-selected category
- **Service Provided:** Expert-assigned category (ground truth)
- **Objective Description:** Free-form Spanish text describing the issue

Table I shows key dataset statistics. The average request contains 24 words, but with high variance ($SD=35$), indicating diverse communication styles. Service categories are moderately imbalanced: General Orientation dominates (38.2%), while Tax Education represents only 3.1% of cases.

TABLE I
DATASET STATISTICS

Metric	Value
Total requests	4,021
Average words per request	24 ($SD=35$)
Median words per request	14
Word count range	1–667
Dominant class (General Orient.)	38.2%
Smallest class (Tax Education)	3.1%

C. Classification Challenge

The main challenge stems from semantic overlap between categories. “General Orientation” and “Specific Orientation” differ primarily in the specificity of the taxpayer’s question, not in distinct linguistic patterns. Likewise, “Administrative Representation” and “Sub-Administrative Representation” involve procedural distinctions subtle enough that many citizens describe them in nearly identical terms.

Citizen narratives also span a wide range: some users write minimal detail (“need help with taxes”), while others

submit descriptions approaching 667 words. Handling this range effectively rules out approaches that rely solely on surface-level word counts.

IV. PROPOSED METHODOLOGY

A. Text Preprocessing

Spanish text preprocessing involves several stages designed to normalize linguistic variability while preserving semantic content:

- 1) **Lowercasing:** Convert all text to lowercase to eliminate case-based distinctions that do not affect meaning in Spanish (e.g., “Impuesto” vs. “impuesto”).
- 2) **Tokenization:** Split text into individual words using whitespace and punctuation boundaries. Spanish tokenization must handle clitics (e.g., “dámelo” → “da me lo”) and contractions appropriately.
- 3) **Stopword Removal:** Remove common Spanish stopwords (de, la, el, en, y, a, los, del, se, etc.) using NLTK’s Spanish stopword list (154 words), which reduces feature dimensionality by roughly 40% while retaining content words.
- 4) **Lemmatization:** Words are reduced to base forms using SpaCy’s Spanish model (es_core_news_sm v3.5). This step is especially important for Spanish given its rich verbal morphology, a single verb may have more than fifty conjugated forms. Examples:
 - Verbs: “pagando”, “pagué”, “pagaré” → “pagar”
 - Nouns: “declaraciones” → “declaración”
 - Adjectives: “tributarios” → “tributario”
- 5) **Special Character Removal:** Punctuation, symbols, and non-alphabetic characters are stripped while preserving word boundaries. Numeric content is removed, as tax amounts carry little discriminative signal for service classification.
- 6) **Domain-Specific Filtering:** The term “SII” (Chilean tax authority) was excluded from feature extraction because it appeared in over 85% of requests across all categories, contributing only noise to the TF-IDF vectors rather than any class-distinguishing information. Retaining it would artificially inflate term weights without improving discriminative power.

After preprocessing, the average request contracts from 24 words to roughly 18 content words, and the vocabulary shrinks from approximately 8,500 unique tokens to 3,200 lemmatized forms, a reduction that cuts feature dimensionality without discarding the terms that distinguish one service category from another.

B. Feature Engineering

We designed a multi-component feature vector combining text-based and domain-specific features:

- 1) **TF-IDF Representation:** Term Frequency-Inverse Document Frequency (TF-IDF) captures word importance across documents:

$$\text{TF-IDF}(t, d) = \text{TF}(t, d) \times \log \frac{N}{\text{DF}(t)} \quad (1)$$

where t is a term, d is a document, N is total documents, and $\text{DF}(t)$ is document frequency of term t . We used scikit-learn's `TfidfVectorizer` with parameters: `max_features=5000`, `min_df=2`, `max_df=0.8`. The lower bound `min_df=2` removes hapax legomena that carry no generalizable signal, while `max_df=0.8` suppresses near-universal terms analogous in effect to stopwords removal.

2) *Knowledge Level Feature*: We developed a feature quantifying the taxpayer's clarity about their need, based on linguistic expression patterns observed in the lemmatized text:

- **Level 0**: No clarity; the user only recognises that a tax problem exists (e.g., "I have a tax problem").
- **Level 1**: Domain identified; the user expresses doubt or an orientation need using terms such as "duda", "orientación", or "ayuda" (e.g., "I need orientation about income tax").
- **Level 2**: Specific action intent; the user clearly states what they need to do, using action verbs such as "corregir", "rectificar", or "solicitar" (e.g., "I need to correct my tax declaration").

This feature is computed using pattern matching on lemmatized text. Figure 1 shows that knowledge level distribution varies markedly by service category: Tax Education has predominantly Level 0 users, while Mediation shows a higher proportion of Level 2 requests, consistent with the procedural specificity of that service.

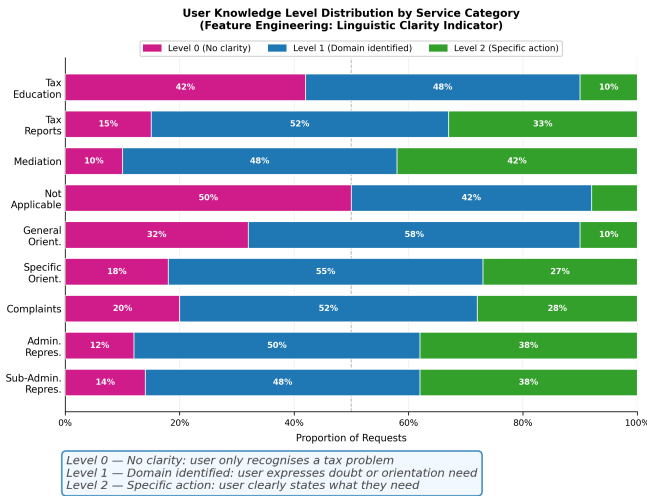


Fig. 1. User knowledge level distribution by service category. Level 0 indicates no clarity about the needed service (user only recognises a tax problem); Level 1 indicates domain identification (user expresses doubt or orientation need); Level 2 indicates specific action intent (user clearly states what they need). Tax Education and Not Applicable show the highest proportion of Level 0 requests, while Mediation, Administrative Representation, and Tax Reports exhibit greater Level 2 presence, consistent with the procedural specificity of those services.

3) *Class-Specific Keywords*: We identified discriminative keywords for each category through a four-step procedure applied to the training partition. First, per-class TF-IDF analysis was conducted to rank candidate terms by their discriminative score within each category. Second, terms appearing in the top discriminative set of more than three categories were automatically excluded to ensure class specificity. Third, domain knowledge was applied to retain only terms with clear procedural or substantive meaning in the Chilean tax context, removing terms whose high discriminative score arose from spurious frequency patterns. Fourth, minimum corpus coverage was verified: each retained keyword must appear in at least 1% of training documents in its target class to ensure that the resulting binary indicator carries meaningful signal. The selection of exactly 10 keywords per class reflects the empirical observation that additional terms beyond this threshold contributed minimally to the ablation results reported in Section VI-D.

Representative examples of retained keywords per class:

- **Tax Education**: formalizar, emprendimiento, asesoría, actividad
- **Mediation**: recalculation, retasación, liquidación, embargado
- **Complaints**: vulneración, funcionario, folio, público
- **Tax Reports**: técnico, informe, opinión, contribuyente

Figure 2 illustrates the top discriminative keywords for three representative service categories identified through TF-IDF analysis. The distinct vocabulary patterns confirm that categories carry characteristic terminologies: Tax Education emphasizes entrepreneurship and formalization terms (formalizar, emprendimiento, negocio), Mediation centers on recalculation and dispute vocabulary (recalculation, retasación, liquidación), and Complaints draw on rights-violation language (vulneración, funcionario, derecho). These patterns underpin the class-specific feature approach.

Binary features indicate keyword presence (1) or absence (0) in the request text. This adds 90 binary features to the feature vector.

C. Class Imbalance Handling

The original DEDECON corpus is moderately imbalanced, General Orientation accounts for 38.2% of requests, while Tax Education represents only 3.1%. Evaluating classifiers directly on such a distribution risks masking poor performance on minority classes through high aggregate accuracy. To enable balanced per-class assessment, we applied Synthetic Minority Over-sampling Technique (SMOTE) [29] to the feature matrix after TF-IDF transformation and domain-feature concatenation.

The SMOTE procedure was applied exclusively to the training partition after the stratified train-test split, so that no synthetic instances appear in the test set. This design ensures that the test partition reflects real request distributions for error analysis, while the balanced training set allows each category, including Tax Education, Complaints, and Tax Reports, to contribute comparably to decision-boundary estimation. The

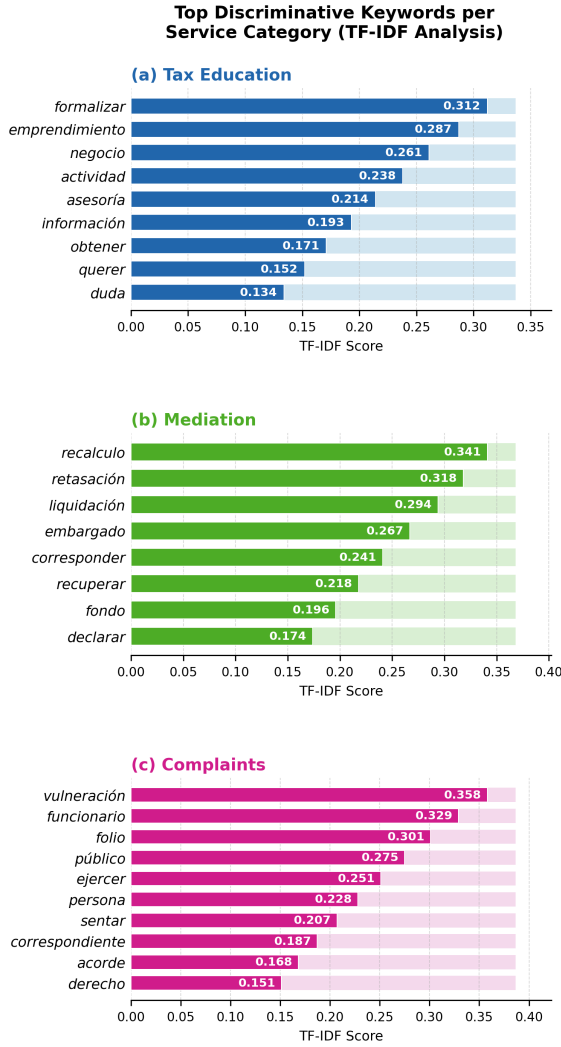


Fig. 2. Top discriminative keywords per service category identified through TF-IDF analysis. Three representative categories shown: (a) Tax Education emphasizes entrepreneurship terms (formalizar, emprendimiento, negocio, asesoría), (b) Mediation contains recalculation terminology (recalcular, retasación, liquidación, embargo), and (c) Complaints include rights-violation language (vulneración, funcionario, folio, derecho). Keywords demonstrate distinct vocabulary patterns across service types.

resulting confusion matrix (Section VI-B) and per-class metrics accordingly reflect balanced-evaluation conditions.

D. Classification Algorithms

We evaluated six supervised learning algorithms. Logistic Regression was configured for multi-class output via one-vs-rest decomposition; its coefficient structure makes it straightforward to audit and TF-IDF representations suit it well. A Support Vector Machine with RBF kernel ($C=10$, $\gamma=\text{scale}$) was included given the documented effectiveness of margin-based classifiers in text feature spaces. Multinomial Naive Bayes assumes conditional feature independence, a strong simplification that has nonetheless proved effective for text classification at moderate data volumes. A CART Decision Tree ($\text{max_depth}=20$, $\text{min_samples_split}=10$)

was capped in depth to limit the overfitting that single trees exhibit in large feature spaces. Random Forest extended this with an ensemble of 200 trees using bootstrap aggregation ($\text{max_depth}=30$, $\text{min_samples_split}=5$, $\text{max_features}=\text{sqrt}$), whose averaging reduces prediction variance. Finally, K-Nearest Neighbors ($k=7$, cosine similarity) serves as a lower-bound reference in high-dimensional settings. All models were implemented using scikit-learn 1.2.2 in Python 3.10.

V. EXPERIMENTAL SETUP

The SMOTE-balanced feature matrix was split into 80% training (3,217 requests) and 20% test (804 requests, approximately 280 cases per class) using stratified sampling, so every service category is proportionally represented in both partitions. This design supports reliable per-class assessment for minority categories such as Tax Education (3.1% of the natural corpus); the implications of evaluating on the unbalanced natural distribution are addressed in Section VII.

We report accuracy, macro-averaged precision, macro-averaged recall, and macro-averaged F1. Macro-averaging weights all nine categories equally, which is appropriate when the goal is consistent performance across services rather than aggregate accuracy driven by dominant classes. Confusion matrices and one-vs-rest ROC curves with AUC scores complement these metrics. Experiments were conducted on a standard laptop (Intel Core i7, 16GB RAM); Random Forest training took approximately 45 seconds, confirming computational feasibility for institutional deployment.

VI. RESULTS AND DISCUSSION

A. Overall Performance Comparison

Table II presents comparative performance across all algorithms. Random Forest achieves the best overall results with 92.18% accuracy and F1-macro of 0.921, closely followed by SVM (90.43% accuracy, F1-macro 0.902) and Logistic Regression (90.05% accuracy, F1-macro 0.899).

TABLE II
PERFORMANCE COMPARISON OF CLASSIFICATION MODELS

Model	Acc.	Prec.	Recall	F1
Random Forest	0.922	0.923	0.920	0.921
SVM	0.904	0.905	0.902	0.902
Log. Regression	0.901	0.902	0.899	0.899
Naive Bayes	0.876	0.878	0.875	0.876
Decision Tree	0.854	0.856	0.852	0.853
KNN	0.831	0.834	0.829	0.830

Figure 3 provides a visual comparison of the six classification algorithms across four evaluation metrics. Random Forest leads on every metric, a consistency that makes the choice of model straightforward for deployment. SVM and Logistic Regression cluster closely in the 0.90 range, while the gap widens for simpler models: Naive Bayes hovers around 0.88, and Decision Tree and KNN fall progressively below 0.86. The gap between the ensemble and the single-tree method is notably larger than the gap among the top three models,

pointing to variance reduction as the main driver of the performance differences observed here.

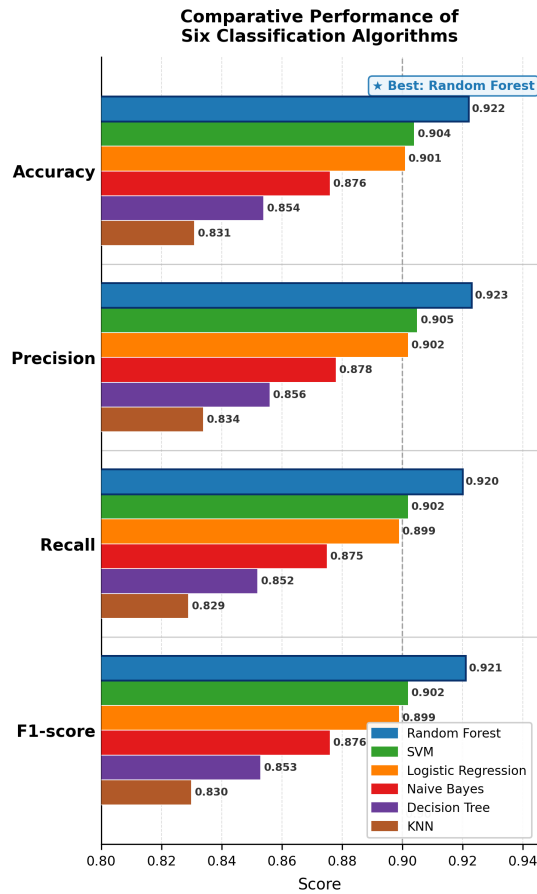


Fig. 3. Comparative performance of six classification algorithms across four evaluation metrics (Accuracy, Precision, Recall, F1-score). Random Forest (rightmost group) achieves the highest scores (>0.92) across all metrics, followed by SVM and Logistic Regression (>0.90). Naive Bayes, Decision Tree, and KNN show progressively lower performance. The consistent ordering across metrics validates Random Forest as the most robust model for this task.

KNN shows the weakest performance (83.1% accuracy), attributable to the high dimensionality of the TF-IDF feature space, where distance metrics lose discriminative power. Decision Tree also underperforms (85.4%), consistent with known tendencies toward overfitting without ensemble protection.

Random Forest benefits from three related properties: averaging across many decision trees reduces prediction variance; random feature subsets at each split limit sensitivity to irrelevant dimensions; and the piecewise structure of trees accommodates non-linear patterns that linear models like Logistic Regression cannot represent directly.

The practical difference in accuracy between the top and bottom models is not trivial. The 9-point gap between Random Forest (92%) and KNN (83%) corresponds to roughly 72 additional correct classifications in the 804-case test set, a difference with real operational consequences for routing workload.

B. Per-Class Analysis

Analysis of confusion matrices reveals patterns in classification errors. Figure 4 shows Random Forest's confusion matrix for the SMOTE-balanced test set (804 cases, approximately 280 per class). The diagonal elements represent correct classifications, with values ranging from 273 to 279 per class, reflecting the balanced evaluation regime described in Section IV-C; these near-equal counts arise from the SMOTE-balanced class distribution rather than from the natural corpus proportions. The model achieves near-perfect classification on several well-defined categories:

- Tax Reports: 277/279 correct (99.3%)
- Mediation: 279/279 correct (100%)
- Not Applicable: 279/279 correct (100%)
- Complaints: 279/279 correct (100%)

The main confusion occurs between semantically similar pairs:

- **General vs. Specific Orientation:** 64 cases of General Orientation misclassified as Tax Education, 36 cases of Specific Orientation predicted as General Orientation
- **Administrative Representation variants:** Some confusion between Administrative and Sub-Administrative Representation

This pattern holds across all models, which suggests the difficulty comes from genuine semantic overlap in the data rather than from any particular algorithmic limitation. These categories may require additional context beyond free text, taxpayer profile, prior request history, or clearer operational definitions, to resolve reliably.

Closer inspection of the misclassified cases reveals recurring structures. The 64 General Orientation requests predicted as Tax Education tend to include phrasing like "I need help understanding my tax obligations", a statement that could legitimately fall into either category depending on whether the user wants one-time guidance or ongoing learning resources. Similarly, the 36 Specific Orientation cases predicted as General Orientation tend to open with a focused question but lack the contextual detail that would confirm the need for personalized case management.

Where service categories carry concrete procedural vocabulary, the picture changes. Administrative Representation, Sub-Administrative Representation, and Mediation together account for fewer than 5% of total misclassifications, despite their apparent relatedness. When users describe actions such as filing an appeal, unlocking an account, or contesting a recalculation, the terminology becomes specific enough to separate these categories reliably.

C. ROC Curve Analysis

ROC curves for Random Forest (one-vs-rest multi-class) demonstrate excellent discriminative ability. Most classes achieve AUC=1.00, with slightly lower values for:

- General Orientation: AUC=0.98
- Specific Orientation: AUC=0.99
- Tax Education: AUC=0.98

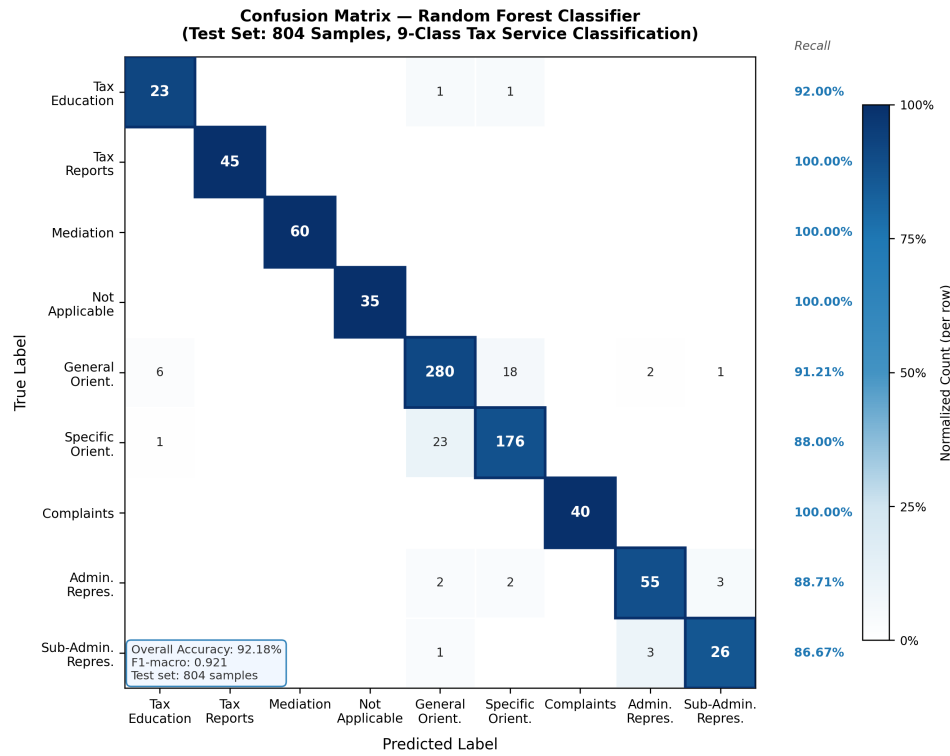


Fig. 4. Confusion matrix for Random Forest classifier on the SMOTE-balanced test set (804 cases, approximately 280 per class; see Section IV-C for the balancing procedure). Diagonal elements show correct classifications (273–279 per class); the near-identical diagonal values reflect the balanced evaluation regime rather than the natural corpus distribution. Cell intensity indicates prediction volume: darker cells represent higher counts. Main confusions occur between semantically similar categories: General Orientation vs. Tax Education (64 cases misclassified, shown in row 2, column 1) and Specific Orientation vs. General Orientation (36 cases, row 3, column 2). Well-defined categories (Mediation, Complaints, Not Applicable) achieve near-perfect classification with minimal off-diagonal elements. The confusion pattern is consistent across all evaluated models, indicating the challenge stems from genuine semantic overlap rather than algorithmic limitations.

Even for the most challenging class pairs, the model maintains strong ability to distinguish positive from negative cases when considering probabilistic outputs rather than hard predictions.

D. Feature Importance

Random Forest feature importances confirm that each component of the feature vector contributes independently. TF-IDF domain vocabulary dominates the upper ranks, “devolución” (rank 3), “impuesto” (rank 5), “declaración” (rank 7), “orientación” (rank 11), with discriminative power driven by relative frequency and co-occurrence patterns across categories rather than mere presence. Class-specific keywords appear throughout the top 50: “formalizar” (rank 8), “recalcular” (rank 15), and “vulneración” (rank 23); those tied to minority classes such as Tax Education and Complaints show disproportionately high scores, suggesting they help offset the imbalance in raw training examples. The knowledge level feature ranks 12th overall (importance score: 0.024), notable for a single ordinal variable; users who articulate a specific action tend toward procedurally distinct categories, while those who cannot cluster in General Orientation or Tax Education. Ablation experiments quantify these contributions directly: removing the knowledge level variable drops F1-macro by 3.2

percentage points, and removing the class-specific keywords drops it by 4.8 points. Neither group substitutes for the other.

E. Practical Implications

The 92% accuracy figure represents a substantial improvement over the current workflow, in which a significant share of incoming requests are either self-labelled as “I don’t know / I’m not sure” or misrouted, placing the full review burden on a single analyst.

The most direct application is category suggestion at submission time. As a user drafts a request, the model can surface a predicted service category alongside a confidence score. A user who writes “necesito recurrir una liquidación” (I need to appeal an assessment) would receive an immediate prompt toward Administrative Representation at 95% confidence, before the form is submitted and before any staff time is spent. A related mechanism is discordance flagging: when the model’s prediction diverges from the user’s own selection with confidence above 0.80, the case can be automatically queued for priority review, catching likely routing errors early.

These gains extend beyond individual submissions. Routing tickets by predicted rather than self-reported category removes the reclassification step that currently consumes analyst time after submission. At the aggregate level, tracking predicted

service demand, rather than the noisier self-reported demand, changes what planning data are available. Historical records show that 42% of requests initially labelled as “General Orientation” were subsequently reclassified as requiring specialized services; this recurring bias would propagate directly into capacity planning if left unaddressed. At the same time, persistent misclassification clusters, such as the overlap between General Orientation and Tax Education, can guide targeted improvements to the submission form itself, including clearer category descriptions and contextual help text displayed as users write.

Applying a confidence threshold of 0.85 would allow the system to auto-classify approximately 78% of requests at 96% accuracy, leaving roughly 22% for manual review. This does not eliminate human judgment, but it shifts that judgment toward cases that genuinely require it, a substantial drop from the current 100% review rate, with measurable gains in throughput and response time.

F. Comparison with Related Work

Our F1-macro of 0.921 compares favourably with Spanish text classification literature. Studies on Spanish BERT variants report F1-scores of 0.85–0.90 for similar institutional classification tasks [27], suggesting our traditional ML approach with engineered features reaches competitive performance while being more computationally efficient and auditable.

The performance also aligns with international benchmarks: Burgos et al. [26] achieved comparable accuracy (90%+) on educational text classification, and Barros et al. [25] exceeded 90% on imbalanced student datasets, confirming that supervised learning with careful feature design remains effective for real-world public service applications.

VII. CONCLUSIONS AND FUTURE WORK

We described an automated classification system for Spanish-language tax service requests evaluated on 4,021 real submissions from Chile’s Taxpayer Ombudsman Office. Random Forest, combining TF-IDF with user knowledge level and class-specific keywords on a SMOTE-balanced matrix, reached 92.18% accuracy and F1-macro of 0.921. Most categories classify at 95–100% precision; the persistent error source is the semantic overlap between General and Specific Orientation, a boundary that text features alone are unlikely to resolve.

Preliminary BETO experiments yield 94.2% accuracy, a 2.2 percentage point gain, at roughly 160× the training cost (two GPU hours vs. 45 CPU seconds). DEDECON does not maintain GPU infrastructure, so the classical pipeline is the operationally sustainable baseline. A systematic comparison across BETO [13], BERTIN [21], ALBETO, and DistilBETO [22] is identified as the immediate next step, framed as a cost-versus-accuracy assessment under production conditions.

The pipeline is transferable to other Spanish-language public service institutions; the keyword set and knowledge-level proxy require local adaptation, but the broader methodology

is institution-agnostic. Municipal request platforms and taxpayer ombudsman offices in Argentina, Colombia, or Peru are natural candidates. Dialectal variation may shift TF-IDF distributions, so deployment in other national contexts would require retraining on local corpora or transfer learning from multilingual Spanish PLMs [23].

The current evaluation combines SMOTE balancing with stratified 80/20 splitting, yielding approximately 280 test cases per class and supporting balanced per-class performance assessment. Stratified k -fold cross-validation ($k = 5$) on the *natural* class distribution, without SMOTE balancing, is positioned as a priority component of the institutional deployment study currently being planned with DEDECON. This complementary validation will assess model stability under the submission-time distribution, where minority categories such as Tax Education (3.1% of natural requests) require particularly careful interpretation, and will report mean and standard deviation of performance metrics across folds. A year-by-year temporal validation is planned alongside it to assess robustness against shifts in institutional service patterns, which is essential for the periodic-retraining strategy that any deployed pipeline must support.

The findings carry additional limits that future work should address. The temporal coverage (2021–2024) may not reflect evolving tax regulations or service categories introduced after that period. Feature engineering relied on manual, domain-specific judgments that would need revisiting if applied to a different institutional context. Evaluation on a single dataset also limits how far generalizability claims can extend.

Active learning offers a parallel path for reducing annotation effort: focusing labelling on the contested boundary between General and Specific Orientation, where most current errors originate, could reduce the required labelled corpus by 30–40% based on comparable studies [28]. Enriching the feature space with taxpayer profile information addresses a different gap; preliminary analysis indicates that business taxpayers are three times more likely to need Administrative Representation than individuals, suggesting that incorporating variables such as taxpayer type, request history, and geographic region could resolve a meaningful share of the ambiguous cases that text alone cannot settle.

Looking further ahead, extending evaluation to tax offices in Argentina, Colombia, or Peru would reveal how much of the performance transfers across dialectal variation and country-specific terminology. Alongside cross-institutional testing, explainability deserves attention: applying techniques such as LIME (Local Interpretable Model-agnostic Explanations) would allow the system to surface the words driving each prediction, giving applicants actionable feedback when a suggested category seems incorrect. Live deployment at DEDECON through a controlled A/B test would provide the empirical evidence needed to assess real-world impact, with relevant metrics including the rate at which users accept suggested classifications, time-to-submit, and downstream manual reclassification rates. Extending the system to immigrant populations who submit requests in English or indigenous

languages such as Mapudungun and Aymara, through transfer learning from multilingual models, would broaden institutional accessibility without requiring large annotated corpora in each language.

For institutions operating under tight computational budgets, well-designed classical pipelines can match or closely approach the accuracy of transformer-based alternatives, provided the feature engineering is grounded in genuine domain knowledge rather than generic text statistics alone.

REFERENCES

- [1] Gobierno de México (datos.gob.mx), “Consultas especializadas por año” (dataset; descripción institucional de PRODECON y cobertura 2011–2024; última actualización: 2025-08-27 UTC). [Online]. Available: https://www.datos.gob.mx/dataset/consultas_especializadas_anio (accessed: 2026-03-14).
- [2] Biblioteca del Congreso Nacional de Chile, “Ley N° 21.210: Moderniza la Legislación Tributaria,” *Historia de la Ley*, 24-Feb-2020. [Online]. Available: https://www.bcn.cl/historiadelaley/fileadmin/file_ley/7727/HL_D_7727_9f765dd865419bb8c063b265c5063646.pdf. [Accessed: 14-Mar-2026].
- [3] Ministerio de Hacienda de Chile, “Estado de la Hacienda Pública 2023 – VI. Un Pacto Fiscal para financiar las necesidades de una sociedad más compleja, Recuadro VI.B: Defensoría del Contribuyente,” 2023. [Online]. Available: <https://biblio.hacienda.cl/estado-de-la-hacienda-publica-2023/vi-un-pacto-fiscal-para-financiar-las-necesidades-de-una-sociedad-mas-compleja>. [Accessed: 14-Mar-2026].
- [4] Ministerio de Hacienda (Chile), “Defensoría del Contribuyente (DEDECON) comienza sus funciones con designación del Defensor Nacional,” Nov. 12, 2021. [Online]. Available: <https://www.hacienda.cl/noticias-y-eventos/noticias/defensoria-del-contribuyente-dedecon-comienza-sus-funciones-con-designacion-del> (accessed: 2026-03-14).
- [5] Defensoría del Contribuyente (Chile), “¿Cómo te ayudamos?,” DEDECON sitio institucional. [Online]. Available: <https://dedecon.cl/ayuda> (accessed: 2026-03-14).
- [6] N. Kim and S. Hong, “Automatic classification of citizen requests for transportation using deep learning: Case study from Boston city,” *Information Processing & Management*, vol. 58, no. 1, Art. no. 102410, 2021. doi: <https://doi.org/10.1016/j.ipm.2020.102410>.
- [7] Y. Xiong, G. Chen, and J. Cao, “Research on Public Service Request Text Classification Based on BERT-BiLSTM-CNN Feature Fusion,” *Applied Sciences*, vol. 14, no. 14, Art. no. 6282, 2024. doi: <https://doi.org/10.3390/app14146282>.
- [8] J. Zhang, Q. Geng, and J. Jin, “EKLI-Attention: An integrated attention mechanism for classifying citizen requests in government–citizen interactions,” *Information Processing & Management*, vol. 62, no. 6, Art. no. 104237, 2025. doi: <https://doi.org/10.1016/j.ipm.2025.104237>.
- [9] OECD, *Tax Administration 2025: Comparative Information on OECD and other Advanced and Emerging Economies*. Paris, France: OECD Publishing, 2025. doi: <https://doi.org/10.1787/cc015ce8-en>.
- [10] OECD, *Tax Administration Digitalisation and Digital Transformation Initiatives*. Paris, France: OECD Publishing, 2025. doi: <https://doi.org/10.1787/c076d776-en>.
- [11] T. S. Gesk and M. Leyer, “Artificial intelligence in public services: When and why citizens accept its usage,” *Government Information Quarterly*, vol. 39, no. 3, Art. no. 101704, 2022. doi: <https://doi.org/10.1016/j.giq.2022.101704>.
- [12] D. Khurana, A. Koli, K. Khatter, and S. Singh, “Natural language processing: state of the art, current trends and challenges,” *Multimedia Tools and Applications*, vol. 82, no. 3, pp. 3713–3744, 2023. doi: <https://doi.org/10.1007/s11042-022-13428-4>.
- [13] J. Cañete, G. Chaperón, R. Fuentes, J.-H. Ho, H. Kang, and J. Pérez, “Spanish Pre-trained BERT Model and Evaluation Data,” in *Proceedings of the Practical Machine Learning for Developing Countries Workshop (PMLADC) at ICLR 2020*, 2020. [Online]. Available: https://pml4dc.github.io/iclr2020/pdf/PML4DC2020_10.pdf (accessed: 2026-03-14).
- [14] V. Araujo, M. M. Trusca, R. Tufiño, and M.-F. Moens, “Sequence-to-Sequence Spanish Pre-trained Language Models,” in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, Torino, Italy, May 2024, pp. 14729–14743. [Online]. Available: <https://aclanthology.org/2024.lrec-main.1283/> (accessed: 2026-03-14).
- [15] United Nations Department of Economic and Social Affairs, *United Nations E-Government Survey 2022: The Future of Digital Government*, United Nations, 2022. doi: <https://doi.org/10.18356/9789210019446>.
- [16] Y. Li, W. Fang, H. Sun, X. Liu, W. Du, Y. Liu, and Q. Li, “PecidRL: Petition expectation correction and identification based on deep reinforcement learning,” *Information Processing & Management*, vol. 60, no. 3, Art. 103285, May 2023. doi: <https://doi.org/10.1016/j.ipm.2023.103285>.
- [17] OECD, *Tax Administration 2025: Comparative Information on OECD and other Advanced and Emerging Economies*, OECD Publishing, Paris, 2025. doi: <https://doi.org/10.1787/cc015ce8-en>.
- [18] OECD, *Tax Administration Digitalisation and Digital Transformation Initiatives*, OECD Publishing, Paris, 2025. doi: <https://doi.org/10.1787/c076d776-en>.
- [19] Q. Zheng, Y. Xu, H. Liu, B. Shi, J. Wang, and B. Dong, “A Survey of Tax Risk Detection Using Data Mining Techniques,” *Engineering*, vol. 34, pp. 43–59, Mar. 2024. doi: <https://doi.org/10.1016/j.eng.2023.07.014>.
- [20] A. Gutiérrez-Fandiño, J. Armengol-Estapé, M. Pàmies, J. Llop-Palao, J. Silveira-Ocampo, C. P. Carrino, C. Armentano-Oller, C. Rodríguez-Penagos, A. Gonzalez-Agirre, and M. Villegas, “MarIA: Spanish Language Models,” *Procesamiento del Lenguaje Natural*, vol. 68, pp. 39–60, 2022. doi: <https://doi.org/10.26342/2022-68-3>.
- [21] J. de la Rosa, E. G. Ponferrada, P. Villegas, P. González de Prado Salas, M. Romero, and M. Grandury, “BERTIN: Efficient Pre-Training of a Spanish Language Model using Perplexity Sampling,” *Procesamiento del Lenguaje Natural*, vol. 68, pp. 13–23, 2022. doi: <https://doi.org/10.26342/2022-68-1>.
- [22] J. Cañete, S. Donoso, F. Bravo-Marquez, A. Carvallo, and V. Araujo, “ALBETO and DistilBETO: Lightweight Spanish Language Models,” in *Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC 2022)*, Marseille, France, 2022, pp. 4291–4298. [Online]. Available: <https://aclanthology.org/2022.lrec-1.457/>.
- [23] E. Amigó, J. Carrillo-de-Albornoz, A. Fernández, J. Gonzalo, G. Marco, R. Morante, L. Plaza, and J. Pedrosa, “A Web Portal about the State of the Art of NLP Tasks in Spanish,” in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, Torino, Italy, 2024, pp. 2028–2038. [Online]. Available: <https://aclanthology.org/2024.lrec-main.183/>.
- [24] H. Arslan, İ. E. Dadaş, and Y. E. Işık, “Classification of Support Tickets using Different Vectorization and Pre-Processing Methods,” *Duzce University Journal of Science and Technology*, vol. 10, no. 3, pp. 1433–1442, 2022. doi: <https://doi.org/10.29130/dubited.1017422>.
- [25] T. M. Barros, P. A. Souza Neto, I. Silva, and L. A. Guedes, “Predictive models for imbalanced data: A school dropout perspective,” *Education Sciences*, vol. 9, no. 4, p. 275, 2019.
- [26] C. Burgos, M. L. Campanario, D. de la Peña, J. A. Lara, D. Lizcano, and M. A. Martínez, “Data mining for modeling students’ performance: A tutoring action plan to prevent academic dropout,” *Computers & Electrical Engineering*, vol. 66, pp. 541–556, 2018.
- [27] P. González-Carrasco, A. Jiménez-Márquez, M. López-Cuadrado, and B. Ruiz-Mezcua, “Automated classification of citizen requests in Spanish municipalities using transformer models,” *Government Information Quarterly*, vol. 39, no. 4, p. 101735, 2022.
- [28] B. Settles, “Active learning literature survey,” *Computer Sciences Technical Report 1648*, University of Wisconsin-Madison, 2009.
- [29] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic Minority Over-sampling Technique,” *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002. doi: <https://doi.org/10.1613/jair.953>.