Procesamiento del Lenguaje Natural, Revista nº 73, septiembre de 2024, pp. 139-149 recibido 02-04-2024 revisado 12-05-2024 aceptado 15-05-2024

Extraction and Structuring of Financial Terminology

Extracción y estructuración de terminología financiera

Jordi Porta-Zamorano, Blanca Carbajo-Coronado, Antonio Moreno-Sandoval
Laboratorio de Lingüística Informática
Universidad Autónoma de Madrid
{jordi.porta, blanca.carbajo, antonio.msandoval}@uam.es

Abstract: This study focuses on automatic term extraction to detect domain-specific terms from Spanish financial reports using BERT and RoBERTa monolingual and multilingual language models. We have evaluated the performance of the models, paying attention to their ability to identify terms that were not present during training. Additionally, we have conducted a thorough analysis of false positives, false negatives, and true positives. To further enhance our analysis, we have employed social network analysis techniques to systematically organize the extracted terms into relevant clusters. Our findings emphasize that transformer language models are a cost-effective choice for identifying such terms and show how clustering allows us to organize them into coherent and meaningful groups.
Keywords: Financial concepts, term extraction, community detection.

Resumen: Este estudio se centra en la extracción automática de términos específicos del dominio de informes financieros españoles utilizando los modelos de lenguaje BERT y RoBERTa, tanto monolingües como multilingües. Evaluamos el rendimiento de los modelos, enfocándonos en su habilidad para generalizar términos no mostrados durante el entrenamiento. Enriquecemos esta evaluación con un análisis exhaustivo de los falsos positivos, falsos negativos y verdaderos positivos. Además, empleamos el análisis de redes sociales como propuesta para organizar sistemáticamente los términos extraídos en agrupaciones con cierta relevancia. Nuestros hallazgos indican que los modelos de lenguaje tipo transformer son una opción rentable para la identificación de este tipo de términos y muestran cómo su agrupación permite organizar los términos financieros en grupos coherentes y significativos.
Palabras clave: Conceptos financieros, extracción terminológica, detección de comunidades.

1 *Introduction*

Term extraction is the process of identifying and extracting specific terms from a corpus of texts. Traditionally, term extraction was done manually, which is a labor-intensive task. This process begins with thorough text preparation and the identification of key terms in context. It then involves validating these terms against specific criteria, refining the list by removing duplicates and standardizing forms, and categorizing and organizing them for exportation.

Automatic term extraction (ATE) is a natural language processing (NLP) technique that identifies and extracts domain-specific terms from text corpora automatically. Our work fits this category, as it aims to automatically extract and structure Spanish financial terminology from a specific subgenre: financial annual reports (Moreno-Sandoval, 2021). We are particularly interested in scenarios where initial resources are limited. Our approach assumes the availability of a small set of specialized texts and an initial terminology list that may be incomplete and contain noise.

In this paper, terms are expressions that denote concepts in specialized domains. We will use systematically “term” and “keyword” as equivalents. Both words express “concepts” relevant to professional users of NLP in the financial domain, especially economists. This implies that some general concepts with more than one meaning, such as *invertir* (to invest) or *acción* (share),

ISSN 1135-5948 DOI 10.26342/2024-73-10

which some terminologists would not properly consider terms, have been included in our research.

ATE aims to facilitate the work of human experts in creating and maintaining terminological resources, including glossaries and thesauri. This process is essential for improving the accuracy and efficiency of various NLP applications, such as information retrieval by enabling more precise search queries, machine translation by providing domain-specific terminology, text mining by facilitating the discovery of key concepts, and ontology construction by providing the building blocks for knowledge representation.

Several techniques are employed for automatic term extraction (Kageura and Umino, 1996; Tran et al., 2023). These techniques can be broadly categorized into three main approaches:

- Statistical methods: These methods rely on statistical patterns and characteristics of terms to identify potential candidates. They typically employ techniques such as frequency, dispersion, and information gain to identify frequent, widely distributed, and informative terms within the given corpus (Mitkov and Corpas, 2008).
- Linguistic methods: These methods utilize linguistic features and patterns to identify potential terms. They often employ techniques such as part-of-speech tagging, morphology analysis, and chunking to identify words or phrases with specific grammatical or semantic properties indicative of terms. Multimedica's term candidate extractor is a sample (Moreno Sandoval et al., 2019)
- Hybrid methods: These methods combine statistical and linguistic approaches to leverage both strengths. They may involve using statistical methods to initially identify candidate terms and then applying linguistic methods to refine the selection based on contextual information and grammatical features. The Sketch Engine's Keywords tool is a well-known example (Jakubíček et al., 2014).

The choice of ATE technique depends on the corpus's specific characteristics, the desired level of granularity in term extraction, and the application domain.

Some key contributions of this paper are the creation of a comprehensive annotated dataset [1] and a detailed error analysis, providing insights for enhancing annotation practices. Our study also introduces an innovative approach by employing graph-based techniques to manage the extracted terms better.

1.1 Related work

Recent advances in ATE have underscored the significant impact of transformer-based models. In the TermEval 2020 shared task (Rigouts Terryn et al., 2020), the winning team (Hazem et al., 2020) showcased the efficacy of BERT-based models, particularly in English and French. Recurrent Neural Networks (RNNs) that used BERT embeddings were notably effective in handling ambiguous and multi-word terms (Rigouts Terryn, Hoste, and Lefever, 2022). Further evidence of transformer models' effectiveness came from a study (Lang et al., 2021) where three such models, including a token classifier, achieved an F1-score of 69.8% on the ACTER dataset—significantly outperforming the previous BERT-based baseline of 48.1%.

Additional research has explored the utility of the mT5 model. For instance, Medical mT5 (García-Ferrero et al., 2024), a multilingual text-to-text model trained on an extensive medical corpus in English, Spanish, French, and Italian, demonstrates mT5's versatility in specialized fields such as medical terminology, even with limited domain-specific data. Studies have confirmed that mT5 effectively transfers keyword extraction capabilities from scientific texts to news stories (Pundefinedzik et al., 2023) and performs better in detecting unseen, overlapping, and discontinuous keywords (Gotkova and Shvets, 2023).

2 Materials and Methods

2.1 Data Sets

2.1.1 Corpus and Initial Term List

We have used a corpus of 315 annual reports from Spanish IBEX companies with a

[1] The annotated corpus and gold standard terminology will be made available in the UAM repository upon the completion of the project CLARA-FINT.

total of 11,761,460 textual elements (Moreno-Sandoval, Gisbert, and Montoro, 2020), and we have used a list of 13,869 terms collected from this source, extracted automatically and revised manually (Carbajo-Coronado and Moreno-Sandoval, 2023). The method consisted of: a) random selection of 20,000 sentences; b) four linguists manually annotated the sentences and compiled over 5,000 keywords; c) a transformer-based model, mT5 (Xue et al., 2020), was trained as a term extractor with the 20,000 labeled sentences; d) the mT5 fine-tuned model run over 2 large annual reports, and the term candidates were revised by 2 linguists; e) the final list (over 13,000 terms) is the result of the merging of the lists curated by the linguists in two phases.

Sometimes, all the possible variants of a term are collected in the list:

(1) a. área de negocio
b. área de negocios
c. áreas de negocio
d. áreas de negocios

However, not all the variants are represented in the list. For this reason, we lemmatized the terms in the initial list to mark all the possible variants appearing in the corpus. In addition, some noise is introduced when some of the inflected forms does not have a terminological meaning.

2.1.2 Training, Development and Test Sets

To carry out the different experiments, up to ten data sets for sizes 50,000 and 100,000 to 1,000,000 with an increment of 100,000 words were generated by random sampling. Each data set was split into training and development sets following an 80-20% distribution. Table 1 shows the average and the standard deviation of the term overlap of the training sets within each size group. It can be observed that as the dataset size increases, the overlap within different training sets increases from 29,7% to 56,5% with a relatively low standard deviation.

A test set with 697 paragraphs and 17,285 textual elements has been manually annotated, by authors 2 and 3. An IAA calculation was not performed, because we wanted to measure the reliability of the expert annotations under time pressure (10 hours in total divided by 5 days). The number of terms annotated was 1,060 (term tokens), which correspond to 525 different terms (term types) of which 455 were in the initial term list and 70 were not, which represent the 13.3% of the terms.

Data Set Size	Training Set Size	Training Set Overlap
50,000	40,000	0.297 ± 0.008
100,000	80,000	0.349 ± 0.007
200,000	160,000	0.406 ± 0.006
300,000	240,000	0.442 ± 0.005
400,000	320,000	0.467 ± 0.006
500,000	400,000	0.489 ± 0.005
600,000	480,000	0.506 ± 0.005
700,000	560,000	0.524 ± 0.005
800,000	640,000	0.541 ± 0.005
900,000	480,000	0.553 ± 0.006
1,000,000	800,000	0.565 ± 0.004

Table 1: Pairwise training set term overlap average and standard deviation for different sizes.

Table 2 shows the average and the standard deviation of the test set term coverage of different datasets. This coverage ranges from an average of 45.1% for datasets of size 50,000 to 85.1% for size 10^6, having all sizes a very small standard deviation.

Training Set Size	Test Set Coverage Mean ± Stdev
50,000	0.451 ± 0.011
100,000	0.563 ± 0.011
200,000	0.684 ± 0.009
300,000	0.732 ± 0.010
400,000	0.767 ± 0.008
500,000	0.788 ± 0.010
600,000	0.807 ± 0.006
700,000	0.826 ± 0.011
800,000	0.836 ± 0.010
900,000	0.847 ± 0.012
1,000,000	0.851 ± 0.010

Table 2: Training and test set average coverage and standard deviation for different sizes.

2.2 Metrics

The evaluation of an ATE system's performance is essential for determining its efficacy and discovering opportunities for advancement. The field of information extraction provides the most prevalent and straightforward metrics for this evaluation. These metrics rely on using a reference term list (ground truth) constructed from the test set. This reference list is then compared to the system's generated, predicted term list. The following metrics are commonly employed:

- Precision: The proportion of extracted terms that are actually correct terminology.
- Recall: The proportion of correct terminology that is actually extracted.
- F1-score: A harmonic mean of precision and recall, providing a balanced measure of both.

2.3 Models

Deep learning-based models offer a significant advantage over previous models by eliminating the need for manual feature engineering. This process, which involves manually identifying and selecting task-relevant features, is often time-consuming and requires domain expertise. Deep learning models, in contrast, possess the ability to learn features directly from data automatically. Additionally, they can capture long-range dependencies between words, effectively handle out-of-vocabulary (OOV) words, and demonstrate adaptability to various text styles. In the domain of deep learning models, transformer architectures have emerged as particularly good architecture for sequence labeling tasks, such as named entity recognition (NER), exhibiting superior generalization capabilities compared to traditional models. Transformers have demonstrably achieved significantly higher accuracy, precision, and recall rates on a variety of NER datasets. To address the challenge of automatic term extraction, the task is also formulated as a sequence labeling problem utilizing a BIO scheme.

2.4 Experiments

We have experimented with three monolingual and two multilingual pretrained transformer models:

- bert-base-spanish-wwm-uncased (Cañete et al., 2020),
- roberta-base-bne (Fandiño et al., 2022),
- bertin-roberta-base-uncased (De la Rosa et al., 2022)
- bert-base-multilingual-uncased (Devlin et al., 2019),
- deberta-v3-base (He, Gao, and Chen, 2023).

To mitigate overfitting, we employ early stopping as a regularization technique during the training process. This approach stops training when the model's performance on the development set begins to deteriorate. For early stopping, we utilize the F1.5-score, an F-score with more weight towards recall than to precision. This choice aligns well with ATE tasks, where minimizing false negatives (missing relevant terms) is more crucial than minimizing false positives (including irrelevant terms) in the generated term candidate lists. Terminologists generally prefer a more comprehensive list with some noise over missing potentially valuable terms.

Fig. 1 shows precision, recall, and F1 curves for different models and different data sizes and Table 3 the metrics for the models trained with 10^6 text. All metrics improve as the size of the training set increases, but no clear improvement could be attributed to the monolingual or multilingual nature of the model. In addition, significance levels of paired t-tests on systems with close F1 do not allow rejecting that systems perform differently[2]. However, multilingual BERT performs very well at recall and gives the best results at F1.

2.5 Analysis of Results

The generalization ability of ML models refers to their capacity to effectively perform on new, unseen data. A model with good generalization can accurately make predictions on data that it has not encountered during training, demonstrating its ability to learn from the training data and apply its knowledge to new situations. This is crucial for the practical application of ML models, as it ensures that they can be deployed in real-world scenarios without significant performance degradation.

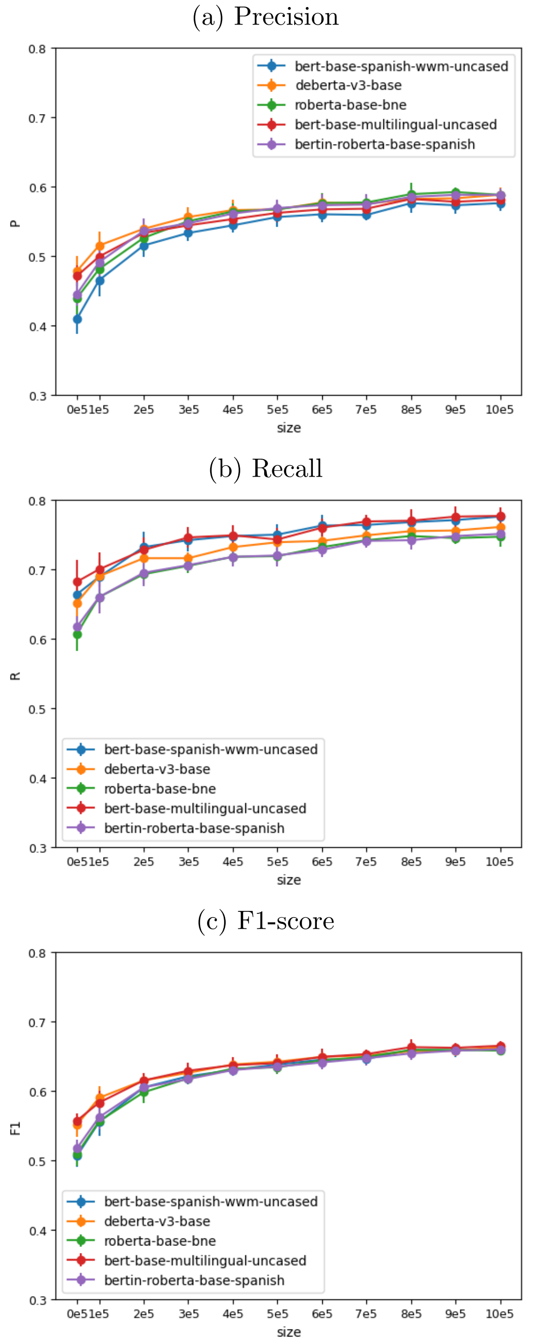
For analyzing the results, we have looked at the term list obtained by applying one of the runs of the multilingual BERT model trained with 10^6 words. The model with the F1-score on the average of its group.

Below, we will analyze qualitatively false positives (FPs), false negatives (FNs), and

[2] The p-value comparing bert-base-multilingual-uncased and deberta-v3-base is 0.665.

Pretrained Model	Precision	Recall	F1-score
bert-base-multilingual-uncased	0.581 ± 0.007	0.777 ± 0.012	0.665 ± 0.006
deberta-v3-base	0.588 ± 0.010	0.761 ± 0.010	0.664 ± 0.007
bert-base-spanish-wwm-uncased	0.576 ± 0.011	0.776 ± 0.009	0.661 ± 0.007
bertin-roberta-base-spanish	0.588 ± 0.009	0.751 ± 0.012	0.660 ± 0.005
roberta-base-bne	0.588 ± 0.008	0.747 ± 0.014	0.658 ± 0.007

Table 3: Precision, recall and F1-score on terminology extraction for models trained with 10^6 text averaged on ten runs with standard deviation.



(a) Precision

(b) Recall

(c) F1-score

Figure 1: Average values and standard deviation of (a) precision, (b) recall, and (c) F1-score for different training sizes and models for ATE.

true positives (TPs) of the test set in depth. Figures are presented in Table 4.

Metric	Total	Percent.
FPs	554	
– Human errors	107	19.31%
– Model errors	252	45.49%
– Partial errors	195	35.20%
FNs	220	
– Human errors	10	4.55%
– Model errors	210	95.45%
TPs	140	
Total tags	**1284**	

Table 4: FPs, FNs and TPs of the test set.

This analysis identifies false positives (FPs) as expressions the model mistakenly classifies as terms in the domain. A sample analysis revealed 554 instances of FPs, encompassing diverse expressions like *cargo al ejercicio* (charge to the fiscal year), *período* (period), *plantillas* (staff), and *compromiso firme* (firm commitment). These FPs arise from various sources, which we have categorized into three distinct classes based on their nature and cause:

- Type 1 (Unlabeled True Positives): These FPs represent financial terms correctly identified by the model but not labeled during the annotation process. Essentially, they are true financial terms missed during annotation. This class constitutes 19.31% of the total FPs.
- Type 2 (Model Errors): These FPs represent actual model performance errors, where non-financial expressions are mistakenly classified as financial terms. This category reflects genuine model limitations and comprises 45.49% of the total FPs.
- Type 3 (Partially Correct Recognitions): These FPs represent terms partially recognized by the model but with either

missing fragments or extraneous additions. This category signifies partial success with limitations and accounts for 35.2% of total FPs.

The multifaceted nature of the annotation task likely contributed to human errors in the first category. This task was designed to serve a dual purpose: supporting the needs of financial professionals and functioning as a general terminology extraction tool. Consequently, the annotation process encompassed a range of general financial terms (e.g., *financiar* (finance), *vender* (sell), *pago* (payment)) deemed valuable for professional analysis, even though they may not qualify as strictly technical terms.

Furthermore, the dynamic nature of the financial sector potentially introduced another source of human error. New terminology and variations constantly emerge within annual reports, particularly in burgeoning fields like corporate social responsibility, with an expanding lexicon ranging from environmental management to social initiatives. This evolution required constant updating of the annotation guide in each revision cycle.

In conclusion, the complexity inherent in annotation tasks does not always guarantee the consistency and completeness of human-curated datasets.

Type 2 errors likely stem from the model's acquisition of incorrect patterns during training, potentially attributable to inconsistencies within the annotated data. Polysemy, the existence of multiple meanings for a single word, further contributes to model errors. The interpretation of meaning heavily relies on context. For instance, the model misinterprets *costa* as a financial term (referencing *a su costa* (at their expense)) when the text refers to a coastline. Similar ambiguities were observed with *fusión* (referring to either a financial entity merger or a chemical process) and *inyección* (interpreted as a monetary contribution instead of a medical procedure). In addition to these examples, we encountered unclassifiable FPs such as *Brazil* or *Latam* whose cause remains unclear.

Type 3 errors are characterized by the model's partial identification of terms. This manifests in two ways: either the model detects only fragments of broader terms or it erroneously extends the existing term boundaries. The most common form of this error involves omitting a portion of the term. This frequently occurs when the head noun of the term is *operación* (operation), *análisis* (analysis), *política* (policy), or *modelo* (model). For example, in the multiword concept *modelo de prevención de riesgos legales* (legal risk prevention model), the system only detects *prevención de riesgos legales* (legal risk prevention). Omission can also occur at the end of the term, affecting modifiers and complements. For instance, in *deuda neta* (net debt), the system extracts only *deuda* (debt). This phenomenon likely arises because these truncated forms frequently appear in the training data.

A false negative (FN) occurs when the model fails by not identifying or recognizing a true term. They have been classified into two main categories:

- Type 1 (Human Errors): These errors encompass the exclusion of relevant terms due to hesitation about their relevance or categorization. Examples include *ahorro de recursos* (resource saving) or *proyecciones de resultados* (result projections). In the latter case, we discussed whether *proyecciones* (projections) should be considered a basic financial term, which in turn generated doubts about the inclusion of *de resultados* (result) in the annotation. They represent 4,55% of the total FNs.
- Type 2: (Model Errors): These errors reflect the model's limitations. Examples include *auditoría interna* (internal audit) and *riesgo de mercado* (market risk). Interestingly, essential concepts from annual reports, such as *consejo de administración* (board of directors) or *informe anual* (annual report), were also not detected. One might consider that these errors arise from the training dataset's features not being sufficiently generalizable across all annual reports or perhaps due to inherent constraints in comprehending the document's structure, especially in headings. They account for 95.45% of the cases.

In both types of FNs we find errors due to extension and omission of part of the term. Specifically, Type 1 errors reveal noticeable inconsistencies in the annotation process, leading to significant discrepancies in ensuring the comprehensive representation of the terms. An example of this is the term *accidentalidad* (accident rate), tagged as *índices de accidentalidad* (accident rate indices) towards the final stages of the project. At the same time, in the initial phases, it was annotated as *accidentalidad*. In a similar case, the term *recompra de acciones* (stock repurchase) was labeled, whereas *planes de recompra de acciones* (stock repurchase plans) would have been a more precise annotation.

The extension and omission of parts in Type 2 errors are also due to the concept structure's inherent complexity, their formulation variability, and the annotation rules' adjustments. A notable example is the detection of *deuda neta* (net debt) instead of *ratio de FFO/Deuda neta* (FFO/Net Debt ratio), which accounts for an omission, or *creación de valor sostenible* (sustainable value creation) instead of *valor sostenible* (sustainable value), which accounts for an extension.

A true positive (TP) is an instance where the model correctly detects a concept, also annotated in the test set. In the model's evaluation, 410 TPs were identified. Interestingly, 10 terms were not previously included in the initial list. Of these terms, 8 had no instances of appearance in either the training or development sets. Furthermore, the term *creación de valor* (value creation) was significantly prevalent, with 88 occurrences in the training set and 16 in the development set. This discovery showcases the model's generalization abilities, as it can accurately detect concepts not encountered during its training phase.

Finally, it should be noted that the analysis of errors has increased the list of terms in the test set from the 1,060 initially labeled by the linguists to a final number of 1,284 terms (224 terms added, 17.44%). The resulting dataset is a gold standard for evaluating other models in the future.

3 Terminology Network and Community Detection

A word co-occurrence network is a graphical representation of relationships between words within a text corpus. This valuable tool facilitates exploring semantic connections between words and identifying patterns and trends in language use. Nodes in the network represent unique words and edges connecting them represent instances where words co-occur within the same context. The weight assigned to an edge reflects the strength of association between the co-occurring words.

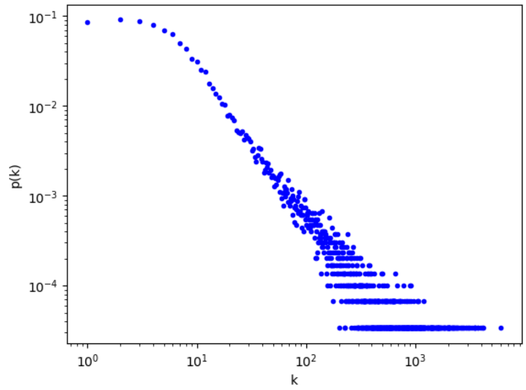


Figure 2: Degree distribution of the term co-occurrence network.

Community detection, also known as network clustering, is a powerful tool for uncovering meaningful substructures in complex network analysis. This technique facilitates data abstraction, unveiling hidden patterns and organizational structures within the network. By identifying communities—groups of nodes with particularly strong internal connections—network clustering offers valuable insights into the relationships between nodes.

We constructed a co-occurrence network of concepts within the same paragraph, based on the terms identified in the corpus of financial reports. The term recognition model employed here was the same multilingual BERT model trained on one million words, as described in Section 2.5. Following the removal of certain invalid terms (e.g., those beginning or ending with prepositions or punctuation symbols), the resulting network comprised 29,562 nodes and 556,714 connections. We calculated their mutual information to quantify the strength of associations between co-occurring terms.

Consistent with observations in other language networks (Solé et al., 2010; Ferrer i Cancho and Solé, 2001), our analysis reveals a small-world network topology. This network class exhibits characteristic properties: a short average shortest path length (L), a high clustering coefficient (C), and a scale-free degree distribution. The latter implies the presence of cliques (fully interconnected node groups), near-cliques (highly interconnected nodes), and numerous sub-networks with dense internal connections. For the term co-occurrence network, L=2.87, C=0.73, and its degree distribution is shown in Fig. 2.

Due to several inherent structural properties, small-world networks present distinct challenges for community detection algorithms. These challenges include the presence of overlapping and nested communities, the need for scalable algorithms, the ability to handle noise effectively, and the development of quality metrics that capture the intricacies of community structures. A diverse array of graph clustering algorithms exists, each offering unique advantages and limitations. Certain algorithms excel at efficiently processing smaller networks, while others are specifically designed to handle the computational demands of massive networks.

Assuming that the community structure of term co-occurrence networks is reflected in both the network's topology and the weights of its connections, we employed the community detection algorithm presented in (Reichardt and Bornholdt, 2006). This algorithm interprets the network's community structure as the spin configuration that minimizes the energy of a spin glass model, where the spin states correspond to community assignments. Simulated annealing is then utilized to achieve high-quality solutions, even if they are not necessarily globally optimal.

In contrast to full-scale community detection, it may be more advantageous to identify the community membership of a specific node within the network. This approach is particularly relevant for very large networks, where comprehensive community detection can be computationally expensive. Here, a fast greedy algorithm can be employed. This algorithm begins by selecting a specific node. It iteratively incorporates nodes with positive adhesion to the growing community, as long as the adhesion between the formed community and the remaining network elements weakens.

To illustrate the process of community detection, let's consider the local community surrounding the subgraph of terms centered on the concept *convenio colectivo* (collective agreement) within a radius of one. This subgraph contains 123 terms and 2,202 connections. The local community for this term comprises 33 members, as detailed below:

1. sistema de previsión social (1)
2. compromisos por pensiones derivados (1)
3. convenio colectivo (0.034693)
4. nivel retributivo (0.0023941)
5. agentes de cambio (0.00089684)
6. política de gestión de capital humano (0.00079714)
7. pacto social (0.00040406)
8. planes de igualdad consolidados (0.00039978)
9. plantilla cubierta (0.00039867)
10. sindicatos independientes (0.0002989)
11. paz social (0.00024157)
12. marco salarial (0.00023942)
13. jubilación anticipada (0.00023912)
14. sindicatos mayoritarios (0.00019927)
15. negociaciones colectivas (0.00017285)
16. plantilla total (0.00013294)
17. subgrupo (9.5797e-05)
18. resultado económico (8.0031e-05)
19. devenir (7.8921e-05)
20. oficinas (7.4725e-05)
21. relaciones internas (6.8122e-05)
22. representación sindical (5.0074e-05)
23. canales de comunicación interna (4.0194e-05)
24. cubierta (3.5813e-05)
25. incrementos salariales (3.1994e-05)
26. agentes sociales (3.1221e-05)
27. actividad profesional (1.8754e-05)
28. bolsa (1.3949e-05)
29. capital humano (1.0405e-05)
30. aportación de valor (1.0366e-05)
31. pagar (9.0434e-06)
32. cubierto (5.9134e-06)
33. directores (4.886e-06)

Notably, each term included in the preceding list possesses an associated score. These scores represent the nodes' eigenvector centrality values. Eigenvector centrality is a network metric that aims to quantify a node's influence or prestige within a connected network.

The "collective agreement" grouping incorporates three clear themes: labor relations, personnel management policies and compensation. *Sistema de previsión social* (welfare system) and *compromisos por pensiones derivados* (derived pension commitments), which are the most central terms are clearly linked to labor relations, retirement systems and compensation.

In this community, there are also some concepts such as *oficinas* (offices) or *cubierto* (covered) that do not belong to the semantic field. The inclusion of these terms may be motivated by the bias introduced by the reports of financial institutions. This detail needs further investigation.

Employing community detection, we can further explore the subgraph centered on terms containing *corrupción* (corruption). This subgraph comprises 123 nodes interconnected by 698 connections. The community detection algorithm was limited to identifying a maximum of 10 communities. The results are presented below, where only the five most central terms within each community are provided:

1. publicidad (1), análisis de los riesgos (1), controles internos (0), anticorrupción (0), órgano de gobierno (0)
2. patrocinios (1), donaciones (0.99488), contribuciones (0.1294), pacto (0.010146), desarrollo sostenible (0.0023177)
3. canal ético (1), incumplimientos (0.70471), comunicar (0.6066), socios comerciales (0.47811), contratistas (0.36318)
4. grupos de trabajo (1), riesgos principales (0.99676), buen gobierno (0.083409), cumplimiento normativo (0.0184), normativa interna (0.014204)
5. gestión de las actividades (1), subcontratadas (0.99993), reputacional (0.027046), marco de control interno (3.2605e-24), demandas (3.2605e-24)
6. reputacionales (1), relaciones de negocio (1), asesores (0), incentivos (0), prevención de riesgos penales (0)
7. control de gestión (1), circulares (0.95455), órganos de gobierno (0.37559), fiscalidad (0.30352), unidad de negocio (0.24498)
8. sistema de control interno (1), sciif (0.67642), auditoría interna (0.59007), control interno (0.47232), riesgos operacionales (0.45346)
9. estado de información no financiera (1), evolución de los negocios (0.90215), soborno (0.83708), informe de gestión (0.75523), corrupción (0.73715)
10. conflictos de intereses (1), conflictos de interés (1), riesgo reputacional (6.1355e-17), valores éticos (6.1355e-17), propiedad intelectual (6.1355e-17)

A quick analysis of these groupings around the concept of "corruption" would be:

- The central idea of this category is "internal control and corporate governance".
- The numbering of the groupings has no meaning of preference or importance.
- The subgroupings most clearly related to this topic would be:

  8. *sistema de control interno* (internal control system) where the company's mechanisms to control corruption appear: *auditoría interna* (internal audit), *control interno* (internal control), and *riesgos operacionales* (operational risks).
  9. *estado de información no financiera* (non-financial reporting status) where *soborno* (bribery) and *corrupción* (corruption) explicitly appear. This category is directly linked to accountability and transparency through the statement of non-financial information and the management report.

- Groups 2 (sponsorships and donations), 3. (ethical channel), 5 (management of activities and subcontracting), 6 (reputation and business relations), and 7 (management control) deal with different aspects related to corruption.

The project's financial expert qualitatively assesses that the groupings and concepts shown cover at least 80% of the categories.

4 Conclusions

The evaluation of various pre-trained models shows consistent concept detection with F1 scores between 0.658 and 0.665. Models like deberta-v3-base and bertin-roberta-base-spanish excel in precision at 0.588, indicating slightly better accuracy with minimal false positives. This precision score is not uncommon for these types of tasks, suggesting both the challenge and potential for improvement.

The analysis of false positives (FPs), false negatives (FNs), and true positives (TPs) identifies 554 FPs, wrongly detected terms, divided into unlabeled true positives (19.31%), genuine model errors (45.49%), and partially correct recognitions (35.2%). FNs occur when the model overlooks actual financial terms, categorized into human errors and model errors, with the latter constituting a significant portion (95.45%). TPs demonstrate the model's capability to accurately identify financial terms, even those not encountered during training. Several factors contribute to these errors, including the evolving complexity of financial terminology, the ongoing need to refine guidelines, and the dedicated annotation time.

The application of network clustering in concept communities has provided a basis for helping terminologists and financial specialists classify and relate concepts. In future work, we would like to explore the separation of banks from the rest of the companies in the corpus of financial reports. Our finance expert has recommended this action, as the conceptual specificities of banks introduce distortions and noise in the concept analysis of other industries.

Acknowledges

This publication is part of the project "Computational linguistic methods for readability and simplification of financial narratives." CLARA-FINT (PID2020-116001RB-C31), funded by the Spanish Ministry of Science and Innovation and the State Research Agency.

We would like to express our sincere appreciation to the project's financial expert, Dr. Ana Gisbert, for her invaluable assistance in evaluating the results of the community detection algorithm.

References

Carbajo-Coronado, B. and A. Moreno-Sandoval. 2023. Financial concepts extraction and lexical simplification in Spanish. *Revista Electrónica de Lingüística Aplicada*, 22(1):164–180.

Cañete, J., G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, and J. Pérez. 2020. Spanish Pre-Trained BERT Model and Evaluation Data. In *PML4DC at ICLR 2020*.

De la Rosa, J., E. Ponferrada, M. Romero, P. Villegas, and P. G. y María Grandury. 2022. Bertin: Efficient pre-training of a spanish language model using perplexity sampling. *Procesamiento del Lenguaje Natural*, 68(0):13–23.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*.

Fandiño, A. G., J. A. Estapé, M. Pàmies, J. L. Palao, J. S. Ocampo, C. P. Carrino, C. A. Oller, C. R. Penagos, A. G. Agirre, and M. Villegas. 2022. MarIA: Spanish Language Models. *Procesamiento del Lenguaje Natural*, 68.

Ferrer i Cancho, F. and R. V. Solé. 2001. The small world of human language. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 268(1482):2261–2265.

García-Ferrero, I., R. Agerri, A. Atutxa Salazar, E. Cabrio, I. de la Iglesia, A. Lavelli, B. Magnini, B. Molinet, J. Ramirez-Romero, G. Rigau, J. M. Villa-Gonzalez, S. Villata, and A. Zaninello. 2024. MedMT5: An open-source multilingual text-to-text LLM for the medical domain. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11165–11177, Torino, Italy, May. ELRA and ICCL.

Gotkova, T. and A. Shvets. 2023. Key environmental lexicon extraction using generative transformer (short paper). In *MDTT*.

Hazem, A., M. Bouhandi, F. Boudin, and B. Daille. 2020. TermEval 2020: TALN-LS2N system for automatic term extraction. In B. Daille, K. Kageura, and A. R. Terryn, editors, *Proceedings of the 6th International Workshop on Computational Terminology*, pages 95–100, Marseille, France, May. European Language Resources Association.

He, P., J. Gao, and W. Chen. 2023. DeBERTav3: Improving deBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*.

Jakubíček, M., A. Kilgarriff, V. Kovář, P. Rychlý, and V. Suchomel. 2014. Finding terms in corpora for many languages with the Sketch Engine. In *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 53–56.

Kageura, K. and B. Umino. 1996. Methods of Automatic Term Recognition: A review. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 3(2):259–289.

Lang, C., L. Wachowiak, B. Heinisch, and D. Gromann. 2021. Transforming term extraction: Transformer-based approaches to multilingual term extraction across domains. In C. Zong, F. Xia, W. Li, and R. Navigli, editors, *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3607–3620, Online, August. Association for Computational Linguistics.

Mitkov, R. and G. Corpas. 2008. Mutual terminology extraction using a statistical framework. *Procesamiento del lenguaje Natural*, (41):107–112.

Moreno-Sandoval, A. 2021. *Financial Narrative Processing in Spanish.* Tirant lo Blanch.

Moreno-Sandoval, A., A. Gisbert, and H. Montoro. 2020. Fint-esp: A corpus of financial reports in Spanish. In *Multiperspectives in analysis and corpus design*, pages 89–102. Comares.

Moreno Sandoval, A., J. Díaz, L. C. Llanos, and T. Redondo. 2019. Biomedical term extraction: NLP techniques in computational medicine. *IJIMAI*, 5(4):51–59.

Pundefinedzik, P., A. Mikołajczyk, A. Wawrzyński, F. Żarnecki, B. Nitoń, and M. Ogrodniczuk. 2023. Transferable keyword extraction and generation with text-to-text language models. In *Computational Science – ICCS 2023: 23rd International Conference, Prague, Czech Republic, July 3–5, 2023, Proceedings, Part II*, page 398–405, Berlin, Heidelberg. Springer-Verlag.

Reichardt, J. and S. Bornholdt. 2006. Statistical mechanics of community detection. *Phys. Rev. E*, 74:016110, Jul.

Rigouts Terryn, A., V. Hoste, P. Drouin, and E. Lefever. 2020. TermEval 2020: Shared task on automatic term extraction using the annotated corpora for term extraction research (ACTER) dataset. In B. Daille, K. Kageura, and A. R. Terryn, editors, *Proceedings of the 6th International Workshop on Computational Terminology*, pages 85–94, Marseille, France, May. European Language Resources Association.

Rigouts Terryn, A., V. Hoste, and E. Lefever. 2022. Tagging terms in text: A supervised sequential labelling approach to automatic term extraction. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 28(1):157–189, March.

Solé, R., B. Corominas-Murtra, S. Valverde, and L. Steels. 2010. Language networks: Their structure, function, and evolution. *Complexity*, 15:20–26, 07.

Tran, H. T. H., M. Martinc, J. Caporusso, A. Doucet, and S. Pollak. 2023. The recent advances in automatic term extraction: A survey. *arXiv:2301.06767.*

Xue, L., N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, and C. Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934.*