International Journal of Combinatorial Optimization Problems and Informatics, 15(4), Nov 2024, 199-210. ISSN: 2007-1558.
https://doi.org/10.61467/2007.1558.2024.v15i4.510

www.editada.org

Multi-Label Topic Classification in a Twitter Corpus of Public Communication of Science in Mexican Spanish

Alec Sánchez-Montero,[1] Gemma Bel-Enguix,[1] Sergio-Luis Ojeda-Trueba[1]
[1] Universidad Nacional Autónoma de México.
alecm@comunidad.unam.mx, gbele@iingen.unam.mx, sojedat@iingen.unam.mx

Abstract. In the context of Mexico, comprehensive studies on the public communication of science (PCS) through social networks remain an unaddressed area of research. To address this gap, the present work is conducted from the perspective of natural language processing (NLP). The objective of this study is to develop and evaluate an automatic multilabel topic classification system for PCS tweets published in Mexico. This is achieved by training various machine learning models, which include traditional algorithms and transformer-based models. Utilizing a manually labeled corpus that identifies eighteen distinct areas or themes of science, the study evaluates and compares several approaches for the automatic identification and classification of thematic areas within PCS tweets. The findings indicate that transformer-based models, such as XLM-RoBERTa, demonstrate superior performance compared to classic algorithms, while the emerging LLM models, such as BLOOM, present a promising alternative for a range of NLP tasks.

Keywords: Natural Language Processing, Multi-Label Text Classification, Public Communication of Science, Machine Learning, Transformers, Large Language Models.

Article Info
Received Jul 24, 2024
Accepted Sep 24, 2024

1 Introduction

In the current era, vast amounts of information are generated and shared across multiple digital platforms on the internet, such as news, blogs, podcasts and multimedia websites. Amid this surge of continuous new data, particularly in mass media, several instances of unreliable content can be found; often, such content is intended to harm or mislead the target audience, as is notably the case with fake news and pseudoscience. In response, automatic filtering tools for unreliable information have been developed, while the public has been warned about the dangers of fallacious information. Nonetheless, one of the most effective ways to prevent the spread of misinformation has been somewhat overlooked: disseminating scientific information through broad and accessible communication channels. Knowledge can be made available in many ways, and one effective method is through non-specialized and accessible discourse, such as public communication of science (PCS).

PCS aims to disseminate and communicate scientific knowledge to the general public in a bidirectional process, contrary to the specialized language conventions used by peers or experts. Consequently, science communicators address individuals who do not possess specialized training in sciences to build a dialogue around discoveries, advancements, and scientific debates. PCS is conveyed through a variety of media channels, including print publications, broadcast media, and digital platforms. Among the latter, social networks are particularly effective for broad outreach due to their accessibility and extensive number of users. Micro-blogging platforms like Twitter/𝕏 enable the spread of scientific information within both specialized communities and the general public. The presence of PCS on social networks also provides an opportunity to counteract misinformation and deception, as such websites are frequently characterized by the prevalence of pseudoscience and misleading content. Therefore, social networks are a key medium for PCS, as they enable the democratization of knowledge, even when they can also promote misinformation and false knowledge, as observed during major global events like the COVID-19 pandemic (Claasen, 2021).

In Mexico, the study of scientific communication, both from linguistic and computational perspectives, is sparse, and PCS on digital platforms has been even less explored. On the other hand, a simple approach could be carried out based on a few simple research questions: What topics are discussed the most? Which science is the most popular? How accessible is the information on these media? However, there is a lack of annotated corpora that can favor comprehensive studies of PCS and the work of identification, topicalization, and classification of their content.

In general terms, one of the purposes of PCS is to bring scientific knowledge closer to a broad, non-specialized audience; thus, the adjective «public» in PCS specifies the target audience of the communicated discourse. Discussing PCS emphasizes the bidirectionality of the communication process. According to Sánchez-Mora (2016), PCS activities are a multi, inter, and transdisciplinary field that brings together knowledge from various areas such as natural sciences, exact sciences, health sciences, technologies, engineering, and recently social sciences and humanities, as well as the management of different media and the understanding of various audiences.

The impact of key figures in scientific communication (Denia, 2020) and the role of educational communities in promoting scientific communication (Déchène, 2024) have also been analyzed. These studies highlight how social media dynamics adapt to the scientific world, such as the emergence of scientific influencers or pages dedicated to PCS within educational communities. Lastly, studies indicate that in recent years, the use of platforms like Twitter/𝕏 for scientific communication has blurred the boundaries between the «specialized» and the «public» community (Peters et al., 2014), highlighting the democratizing potential of social networks when used with good intentions.

Notably, there is an increasing emphasis on the importance of social networks as a venue for circulating scientific information and interaction between scientists and the non-scientific public. As can be seen, research on PCS on Twitter/𝕏, and social networks in general, is a novel and highly interesting area, prompting the scientific community to pay more attention to this social network as it represents an opportunity to investigate, extend, and discuss scientific knowledge (Cheplygina et al., 2020; Daneshjou et al., 2021; Milbourne, 2022).

In this paper, we examine the role of PCS on social networks in Mexico through tools and models provided by Natural Language Processing (NLP) and data science. Specifically, we focus on Twitter/𝕏, a prominent medium for sharing opinions and information. From an NLP perspective, studying PCS in Mexico via Twitter allows for the analysis of various facets of the phenomenon, such as information coverage based on topics. In particular, the objective of this study is to develop and evaluate an automatic classification model for scientific disciplines based on a corpus of PCS tweets published in Mexico, manually annotated with a multi-label system according to the thematic areas addressed in each text.

The structure of the paper is as follows: In the second section, we discuss the significance of PCS on Twitter/𝕏, while also reviewing related work in the field of scientific communication. In the third section, we detail the characteristics of the dataset created for our research and the general methodology, from data collection to the experimentation phase. Section 4 covers the training and evaluation process of the experiments. Finally, Section 5 presents the conclusions and outlines directions for future research.

2 Related works

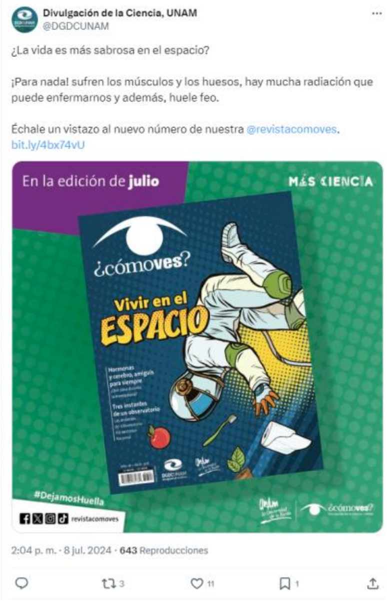
Models of scientific communication, such as public participation, which aim to generate dialogue and engagement with the non-scientific public (Lewenstein, 2003), find a suitable place for their implementation on Twitter/𝕏. Several studies have analyzed the characteristics and qualities of PCS on this social network. Researchers have sought to measure public interaction with specific scientific topics (Guenther et al., 2023), as well as the degree of influence among users in the public's interest and understanding of science (Denia, 2020). Another interesting aspect studied is the increased frequency of scientific vocabulary usage thanks to PCS on Twitter (Sundström, 2021), a phenomenon also common in meme culture.

The language of PCS has not been extensively approached from the perspective of NLP and computation. Regarding the general area of science communication in English, August et al. (2020) offer a quantitative study of writing strategies based on a corpus of science documents. Rakedzon et al. (2017) implement a jargon identifier for scientists and educators who must communicate science. Additionally, there are contributions to the automatic classification of science writings (Joorabchi et al., 2011; Machado & Oliveira, 2021; Rendón Miranda et al., 2014). In contrast, PCS has not received much attention (Macedo-Rouet et al., 2003; Pilkington, 2019; Kyvik, 2005), and has been linked more to journalistic analysis (Lermann et al., 2023).

Artificial Intelligence has an impact on PCS. The use of ChatGPT to generate scientific communication, for example, has the handicap of the acknowledged difficulty of this system to differentiate fact from fiction Rawte, Sheth & Das, 2023), besides being unable to handle different references and citations (Athaluri et al., 2023). Even so, the great capacity of these models to perform tasks such as automatic summarization or automatic image generation, among other capabilities, hints at the great possibilities of this interaction (Henke, J., 2024). Könneker (2024) highlights that LLM-based applications can summarize scientific publications and make them texts that can be understandable by non-expert people.

Concerning the Mexican context, the use of Twitter/𝕏 in Mexico is quite significant, ranking among the top 10 countries with the highest number of active users, surpassing 17 million. Twitter/𝕏 has become a popular space for mass digital communication, where scientific knowledge is shared, discoveries are discussed, and dialogue on scientific topics from various disciplines is promoted. In the case of PCS on Twitter/𝕏, the texts published, known as «tweets» or posts, are brief and fragmented, influenced by factors such as platform interactivity and space limitations (Aguilar-Tello & Angulo-Giraldo, 2022; Barajas-Galindo & Rodríguez Carnero, 2020).

Regarding the PCS on Mexican social networks, our initial exploration has shown a high diversity and breadth of communicated topics, such as physics, biology, and astronomy. It should be noted that PCS publications complement other activities or resources intended for the non-specialized public, whether physical or virtual elements. To illustrate some of the main characteristics of these tweets, a prototypical example is presented in Figure 1.


Divulgación de la Ciencia, UNAM
@DGDCUNAM

¿La vida es más sabrosa en el espacio?

¡Para nada! sufren los músculos y los huesos, hay mucha radiación que puede enfermarnos y además, huele feo.

Échale un vistazo al nuevo número de nuestra @revistacomoves.
bit.ly/4bx74vU

2:04 p. m. · 8 jul. 2024 · 643 Reproducciones

Fig. 1. Example of a PCS tweet in Mexican Spanish (Source: Twitter/𝕏).

In terms of the agents involved in PCS, the network of science communicators shows a homogeneous character. Both institutions and individuals participate in communicating scientific knowledge. Among these communicators, the efforts of higher education institutions, research organizations and centers, scientific dissemination publications, researchers, students, and individual science communicators stand out.

3 Methodology and dataset characteristics

The methodology of this work has been developed based on a pipeline for text classification (Kowsari et al., 2019). This pipeline consists of the following stages: 1) data collection and selection, 2) data annotation, 3) text preprocessing, 4) feature extraction and selection of classification techniques, and 5) evaluation, as depicted in Figure 2. In the first stage, the Twitter API v2 was used to extract the timelines from a predefined list of users using the Tweepy library in a Python environment. This list of users was the result of an analysis to select Twitter profiles related to PCS in the Mexican context. These profiles belong to 19 self-described «disseminators», «communicators» or «science journalists», from both individual and institutional accounts on general scientific areas. In other words, the accounts selected for this study represent a variety of general and non-specialized scientific topics, in accordance with the areas of scientific knowledge covered in the Mexican context.

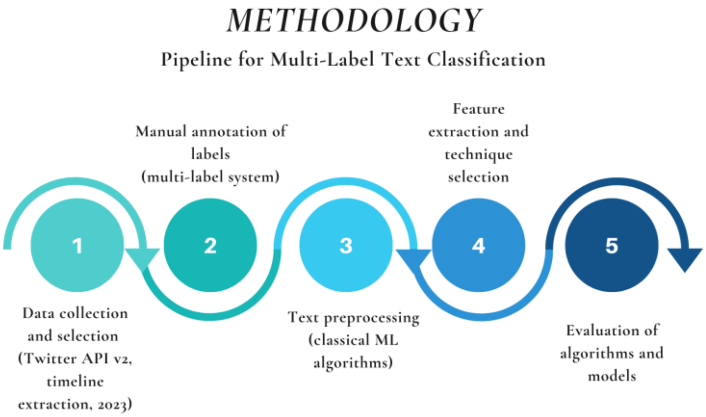

METHODOLOGY
Pipeline for Multi-Label Text Classification
1
Data collection and selection (Twitter API v2, timeline extraction, 2023)
2
Manual annotation of labels (multi-label system)
3
Text preprocessing (classical ML algorithms)
4
Feature extraction and technique selection
5
Evaluation of algorithms and models

Fig. 2. Pipeline for Multi-Label Text Classification in NLP.

It should be noted that we collected these data without specific preferences for a particular scientific field, since we started from a scenario with very little quantitative information regarding the context of the object of study. This situation resulted in a wide range of topics in the corpus, from astronomy and general physics to genetics and history of science, among other areas. In terms of corpus classes, this type of random subject distribution could imply an imbalance in the categories we used for training our model. Adopting such a strategy is justifiable to more accurately reflect the natural distribution of topics in the context of PCS tweets in Mexico. By constructing a dataset that captures the true distribution of classes, one can identify specific thematic areas where models may perform optimally or sub-optimally, a fundamental data-driven feature for research in this novel area of study.

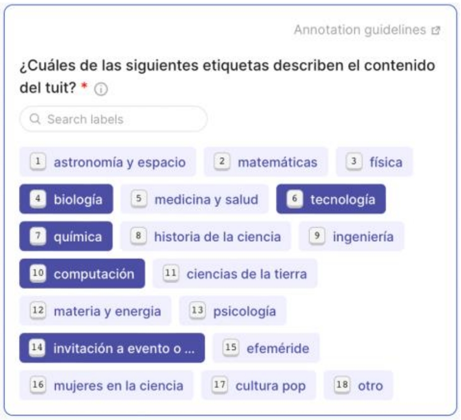
To constitute a consistent and appropriate dataset for the classification task, the fundamental homogeneity criterion was that all tweets should have been written in Spanish and published in Mexico between January 2020 and May 2023, when we collected the data. After gathering this information with the Twitter API v2, significant modifications occurred in the platform, derived from the change of ownership. The transition from the name «Twitter» to «𝕏» was one of them. Another important update, related to platform data collection was the elimination of free access to the academic research API, along with the introduction of a monthly payment model starting at US$100 to use an API with more limited functionalities, compared to the now non-existent academic version. It is relevant to consider this modification, since, in the future, the development of research related to this platform could be conditioned by economic constraints.

After eliminating duplicate messages, texts of a personal nature or non-scientific opinions by the authors, the dataset comprised 3,733 tweets. This dataset was taken as the research corpus to be labeled according to the thematic content of each text. Based on the tokenization function of the SpaCy library, the total number of tokens in the corpus is 144,375 and the number of unique tokens is 21,830.

Argilla, an open-source platform specialized in the development of LLM, was selected to carry out the corpus annotation. Through this platform users can load datasets via a programming code through a Python environment. In this task, we aimed to accurately identify and classify the thematic areas present in the tweets of the corpus, rather than ensuring balanced classes, based on a predefined list of tags. For annotating each tweet, a multi-label text classification approach was used, i.e. each tweet could be labeled with one or more thematic categories, represented by the list of labels corresponding to its content. For this purpose, we opted for the «Feedback Dataset» feature in Argilla, which allows flexible and straightforward multi-label classification of individual records in a dataset.

As a result of previous qualitative data analysis, a set of 18 labels was proposed, aiming to be as exhaustive and representative as possible. These labels were: «astronomy and space» («astronomía y espacio»), «matemáticas» («mahematics»), «physics» («física»), «biología» («biology»), «medicina y salud» («medicine and health»), «technology» («tecnología»), «química» («chemistry»), «history of science» («historia de la ciencia»), «engineering» («ingeniería»), «computation» («computación»), «earth science» («ciencias de la tierra»), «matter and energy» («materia y energía»), «psychology» («psicología»), «external resources or events» («invitación a evento o a recursos»), «anniversary or "day of"» («efeméride»), «women in science» («mujeres en la ciencia»), «pop culture» («cultura pop»), and «other» («otro»). The visualization of these 18 labels or tags in the Argilla interface (https://argilla.io/) is shown in Figure 3, in Spanish. As an instruction for manual labeling, it was indicated to read each of the tweets separately and, based on the information in the text and the contextual features of each record, select all the labels that described the content of each tweet.

.



Annotation guidelines

¿Cuáles de las siguientes etiquetas describen el contenido del tuit? *

Search labels

1 astronomía y espacio
2 matemáticas
3 física
4 biología
5 medicina y salud
6 tecnología
7 química
8 historia de la ciencia
9 ingeniería
10 computación
11 ciencias de la tierra
12 materia y energía
13 psicología
14 invitación a evento o ...
15 efeméride
16 mujeres en la ciencia
17 cultura pop
18 otro

Fig. 3. Multi-label visualization for the annotation of the corpus thematic areas in Argilla (in Spanish).

Once the corpus annotation task was completed, we obtained the results shown in the graph of Figure 4. Based on this graph, we observe that each tweet can be associated with one or more labels due to the multi-label nature of the annotation, so the sum of the frequencies is not equal to the total number of records in the corpus. The labeling results reveal the distribution of predominant thematic areas in the corpus, which does not correspond to a balanced class distribution, as some labels have a broad representation in contrast to other labels with fewer identified cases.

Among the most frequent labels are «biology» with 1705 instances, «external resources or events» with 1665, «physics» with 1485, «astronomy and space» with 1223, and «medicine and health» with 756. In contrast, the least represented areas in the corpus are «computation» with 98 instances, «pop culture» with 79, and «other» with 26. This distribution indicates the prevalence of general areas (such as biology and physics) and specific areas (such as astronomy and space topics) that may correspond to the most relevant topics in the Mexican context of the PCS. At the same time, the less represented areas may be the least disseminated (such as computation), complementary to another more general non-scientific thematic area (such as pop culture), or not relevant for the study of the PCS on Twitter (such as the «other» category).

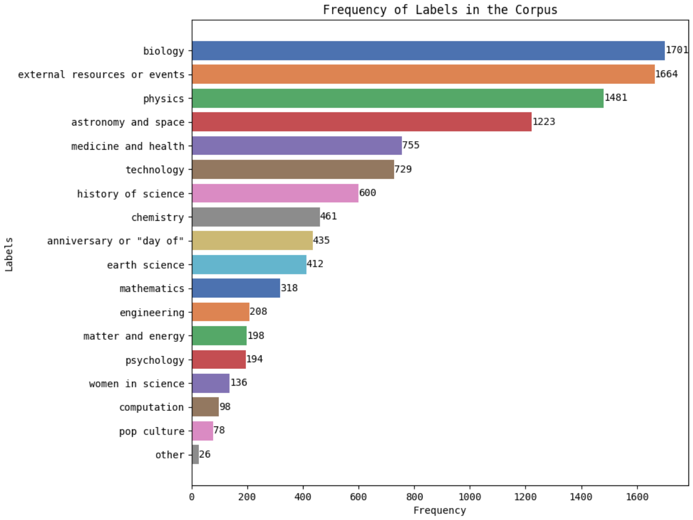


Fig. 4. Distribution of the thematic areas in the corpus.

With this corpus as the basis for training the automatic tweet classification model, we proceeded with the following phases of the methodology. Classic machine learning algorithms were selected because some previous works have demonstrated their usefulness for text classification with small datasets (Riekert et al., 2021). In turn, we aimed to contrast the performance of these algorithms with transformer-based models, which represent a more advanced approach in NLP tasks. When using the classic algorithms, it was necessary to preprocess the corpus texts. At this stage, various text cleaning and normalization techniques were used, such as removing punctuation marks, emojis, hashtags, stop words, hyperlinks, email addresses, and other non-lexical elements, as well as lemmatizing words and converting tokens to lowercase. Then, features of the preprocessed tweets were extracted using the TF-IDF technique to obtain numerical representations of the semantic and contextual features of the texts, to train the selected classic machine learning algorithms.

Due to the characteristics of transformer-based models, the process of training and evaluating texts from the dataset was simplified compared to traditional machine learning methods. In other words, since transformer architectures are designed to work efficiently with text sequences, performing an exhaustive preprocessing stage becomes optional. In this case, no text preprocessing was performed and, for feature extraction, the Hugging Face Transformers library was used based on the representation of the data in the form of raw text tokens, which were fed as input for the models to automatically learn the feature representations during the training process, provided as input to the models to automatically learn the feature representations during the training process, a standard practice with this type of models in NLP.

Moreover, since we worked with an annotated dataset comprising 18 different labels, a label encoding approach known as multilabelbinarizer was applied, using the Scikit-learn library. This technique allows the proper representation of multi-label annotations in the form of binary vectors composed of ones and zeros, where each element of the vector indicates the presence (one) or absence (zero) of a particular tag. This approach was used in both the classic algorithms and the transformer models to enable the training of the model in the classification of tweets with multiple labels.

For the selection of classic algorithms, we opted for a series of widely used classification algorithms that have demonstrated solid performance in previous classification tasks. These algorithms are Support Vector Machine (SVM), Random Forest Classifier (RFC), Linear Regression (LR), and Decision Tree Classifier (DTC). Regarding the pre-trained transformer-based models, the following pre-trained models were selected: distilbert-base-multilingual-cased (Sanh et al., 2020), distilroberta-base (Sanh et al., 2020), twitter-xlm-roberta-base-sentiment (Barbieri et al., 2022), bert-base-multilingual-cased (Devlin et al., 2019), xlm-roberta-large (Conneau et al., 2020), gpt2-spanish (Radford et al., 2019), and bloom-560m (BigScience Workshop, 2022). For the selection of the models, we made sure that they were pre-trained in Spanish and that they did not have high requirements in terms of computational resources. To evaluate all these models, standard classification evaluation metrics were used: precision, recall, and F1 score.

4 Training and evaluation experiments of automatic classification models

As mentioned in the previous section, for the training of classic machine learning algorithms (i.e., SVM, RFC, LR, and DTC), we used a method that included text preprocessing and feature extraction with the TF-IDF vectorizer. We trained two versions of the SVM algorithm: one without class weight adjustment and one with balanced class weights. This dual approach allowed us to compare the impact of class balancing on the performance of the SVM models, particularly in handling class imbalance in our dataset. RFC was chosen for its robustness and versatility in handling non-linear relationships and feature importance, while LR served as a simple baseline model, providing a straightforward comparison for evaluating the performance of more complex algorithms; finally, the DTC algorithm offers interpretability and can capture non-linear relationships. Additionally, we employed a 5-fold cross-validation to evaluate the models' performance with greater accuracy. This technique divides the dataset to obtain five performance measures, which are averaged to provide a general estimate of the model's performance. Table 1 presents the detailed results of the ML algorithms on the evaluation dataset, which corresponds to 20% of the dataset with a random state of 42 for reproducibility.

Table 1. Results of Machine Learning algorithms on the test dataset based on precision, recall and F1 score metrics

ML Algorithm	Precision	Recall	F1 Score
Support Vector Machine (linear without class weight)	0.9	0.61	0.72
Support Vector Machine (linear with balanced class weight)	0.83	**0.69**	**0.75**
Random Forest Classifier	0.9	0.57	0.7
Linear Regression	**0.92**	0.46	0.61
Decision Tree Classifier	0.71	**0.69**	0.7

As can be seen from the data, the LR algorithm achieved the highest precision (0.92) among all the options, although the SVM without class weights and the RFC offered a close value (0.9). In contrast, the SVM model with balanced class weight had both the highest recall (0.69) and F1 score (0.75), which indicates its superior performance in identifying positive examples and maintaining a good balance between precision and recall. Nonetheless, the DTC algorithm demonstrated a balanced performance across all metrics, also achieving the highest recall (0.69). On the other hand, the RFC algorithm, while having a high precision, showed a lower recall (0.57) and F1 score (0.7), while the LR's highest precision came at the cost of having both the lowest recall (0.46) and F1 score (0.61). It is important to note that these results refer to the overall performance of the models and not to metrics by specific labels.

Regarding the transformer-based models, we decided to train and evaluate them based on their size, which led to two categories for our text classification experiments. First, we examined DistilBERT and DistilRoBERTa. These are efficient versions of BERT and RoBERTa, respectively, pre-trained in multiple languages, including Spanish. Additionally, we used the twitter-xlm-roberta-base-sentiment model, which is specifically designed for sentiment analysis in social media text, such as Twitter, and is also pre-trained in Spanish. We set up both distilbert-base-multilingual-cased and distilroberta-base for sequence classification using the Transformers library, addressing the multi-label classification problem. The twitter-xlm-roberta-base-sentiment model was fine-tuned for the classification task with 18 labels, leveraging its base architecture designed for predicting three sentiment tags: positive, negative, and neutral. The same data partition method used for the ML algoirthms was applied to all transformer-based models. Specifically, the dataset was split into an 80-20 training-test ratio with a random state of 42. Additionally, for predictions, a threshold of 0.5 was set for label assignment, meaning a label was assigned to a text instance if the predicted probability for that label was equal to or greater than 0.5.

The results of this first set of transformer-based models, evaluated on the test dataset, are shown in Table 2. This table presents precision, recall, and F1 score for each model trained with 5 and 10 epochs. Compared to the classic ML algorithms, the lightweight transformer-based models generally show lower performance in terms of precision, recall, and F1 score.

Table 2. Results of lightweight transformer-based models with different numbers of epochs in the test dataset

Model	Epochs	Precision	Recall	F1 Score
distilbert-base-multilingual-cased	5	**0.83**	**0.6**	**0.64**
	10	**0.83**	0.62	**0.67**
distilroberta-base	5	0.74	0.48	0.52
	10	0.73	0.56	0.6
twitter-xlm-roberta-base-sentiment	5	0.82	0.59	0.63
	10	0.81	**0.63**	0.65

Among these evaluated models, the DistilBERT model maintained consistent precision at 0.83 across both 5 and 10 epochs, with improvements in recall and F1 score at 10 epochs. This model exhibits higher precision, and F1-score compared to the other models from this category. However, the twitter-xlm-roberta model also demonstrates competitive results, with performance very close to DistilBERT across all evaluation metrics, and even showing higher recall values in the 10-epoch training. This model saw a minor drop in precision, but better recall and F1-score with 10 epochs. In contrast, the distilroberta-base model presents notably lower performance across all evaluated metrics, with a slight decrease in precision but improved recall and F1 score with more training epochs. Notably, the results suggest that while the models exhibit solid performance - although a little lower than ML algorithms-, variations in epoch numbers impact their precision and recall differently, highlighting the importance of epoch tuning for optimizing performance.

So far, in all cases, both in all general classification experiments, it was observed that recall remained the lowest value among the evaluation metrics, due to the imbalance situation between classes. This could indicate that the models struggle to retrieve all positive cases related to the correct tags. A low recall means that the model is missing some positive cases. In the context of multi-label classification, low recall can arise for various reasons, such as the complexity of label relationships, class imbalance, the quality of the training dataset, or the model configuration.

To address this issue, we investigated whether training larger transformer-based models could improve recall and overall performance. In particular, we turned to bert-base-multilingual-cased, xlm-roberta-large, gpt2-spanish, and bloom-560m. As we mentioned before, we ensured that all these models were pre-trained in Spanish and did not have high computational resources requirements. Even though we are aware of newer, more advanced, and complex LLMs, our goal was to determine if we could improve recall by using models that are more accessible and feasible for limited computational capabilities. In other words, our objective was to find an optimal balance between achieving high model performance and maintaining resource efficiency.

These larger general-purpose models, with their greater capacity and advanced architectures in comparison to what we had tried so far, were trained on the same dataset to evaluate their performance metrics. Our hypothesis was that the increased parameters and training capabilities of these larger models would enhance the overall performance, potentially addressing the deficiencies observed with previous evaluated models. To test this hypothesis, we conducted a series of experiments with each of the larger models, carefully monitoring their precision, recall, and F1 scores. The results were then compared to those obtained from the lightweight models to assess any improvements in handling class imbalance and retrieving positive cases.

For the fine-tuning process, we utilized 10 epochs based on results from our experiments with the lightweight models. Specific adjustments to model architecture were made for models such as GPT-2 and BLOOM-560m, which are primarily designed for text generation. During fine-tuning, we paid close attention to optimizing hyperparameters and ensuring that the models could effectively handle the multi-label classification task. Table 3 displays the results of fine-tuning these larger transformer-based models on the multi-label classification task.

Table 3. Results of larger transformer-based models evaluated on the test dataset

Model	Epochs	Precision	Recall	F1 Score
bert-base-multilingual-cased	10	0.82	0.69	0.69
xlm-roberta-large	10	0.79	**0.77**	**0.77**
gpt2-spanish	10	0.83	0.65	0.65
bloom-560m	10	**0.85**	0.7	0.75

Among the models evaluated, xlm-roberta-large achieved the highest F1 score (0.77) of all models evaluated, both ML algorithms and transformer-based models. This highest F1 score reflects a strong balance between precision and recall, i.e. the model effectively handles both false positives and false negatives. This model also demonstrated its effectiveness in retrieving relevant labels, demonstrating the highest recall (0.77). These preliminary results could support our hypothesis that larger, more advanced models can better address the challenges of multi-label classification, particularly in retrieving positive cases and managing class imbalance in datasets such as ours. Even though the precision of xlm-roberta-large remained the lowest (0.79) among larger models, its overall performance in handling the multi-label classification task was remarkable.

Furthermore, the BLOOM model we trained —a smaller version than the base model— achieved a runner-up performance with the highest precision among transformer-based models (0.85). It also demonstrated a high recall (0.7) and a notable F1 score (0.75). This particular model's strength was precision, indicating its accuracy in predicting positive labels. In contrast, the BERT model had a precision of 0.82 and a recall of 0.69, resulting in an F1 score of 0.69, a higher value than its distilled version. Finally, the GPT-2 model showed a high precision (0.83), but had a lower recall (0.65), although this recall was slightly higher compared to lightweight models. Consequently, the GPT-2 model's F1 score was 0.65.

Despite being an older model published in 2020, xlm-roberta-large achieved the highest F1 score (0.77) among all models evaluated, both classic ML algorithms and transformer-based. This highlights its continued relevance and effectiveness in multi-label classification tasks, showing that a strong and decent performance can be achieved without resorting to the most recent and computationally intensive models. Older models such as xlm-roberta-large can still deliver strong performance when properly fine-tuned, making them a practical choice for text classification tasks in languages like Spanish, which means efficient solutions may be available without always needing the latest large language models.

To better understand the performance of xlm-roberta-large at a more granular level, we compared its metrics per label with those of the SVM with balanced class weight, i.e. the ML algorithm that achieved the highest balanced performance for the same classification task. This comparison could help us delve into the automatic classification on each individual label in more detail, so we can identify where the model is performing better, allowing us to pinpoint both the well-performing and the challenging classes. Table 4 presents the comparison of per-label metrics (precision, recall, and F1) for xlm-roberta-large and the SVM with balanced class weights.

Table 4. Comparison of per-label metrics for SVM (ML algorithm) and XLM-RoBERTa (transformer-based)

Label	Precision SVM	Precision XLM-RoBERTa	Recall SVM	Recall XLM-RoBERTa	F1 SVM	F1 XLM-RoBERTa
astronomy and space	**0.93**	**0.93**	0.88	**0.95**	0.9	**0.94**
biology	0.87	**0.92**	0.87	**0.93**	0.87	**0.93**
earth science	0.64	**0.67**	0.53	**0.82**	0.58	**0.74**
computation	1	0.75	0.43	**0.65**	0.61	**0.69**
pop culture	1	0.71	0.07	**0.33**	0.12	**0.45**
anniversary or "day of"	0.75	**0.93**	0.6	**0.95**	0.66	**0.94**
physics	0.87	0.85	0.83	**0.89**	0.85	**0.87**
history of science	0.65	**0.8**	0.53	**0.82**	0.58	**0.81**
engineering	0.56	**0.65**	0.21	**0.5**	0.31	**0.56**
external resources or events	0.76	**0.82**	0.72	**0.84**	0.74	**0.83**
mathematics	**0.98**	0.96	0.66	**0.85**	0.79	**0.9**
matter and energy	**0.57**	0.52	0.39	**0.4**	**0.46**	**0.46**
medicine and health	0.8	**0.88**	0.76	**0.9**	0.78	**0.89**
women in science	0.78	**0.81**	0.35	**0.9**	0.48	**0.85**
other	1	0.66	0.33	**0.66**	0.5	**0.66**
psychology	0.75	**0.85**	0.47	**0.78**	0.58	**0.82**
chemistry	0.56	**0.75**	0.51	**0.7**	0.53	**0.72**
technology	0.78	**0.81**	0.65	**0.85**	0.71	**0.83**

The comparison between SVM and XLM-RoBERTa —the top-performing classification models from their respective categories— reveals how XLM-RoBERTa generally outperforms traditional ML algorithms, particularly in terms of F1 score for individual labels. In our comparison, XLM-RoBERTa consistently achieved the highest F1 score across most labels, except for a tie in the F1 score of 0.46 for the label «matter and energy». It should be noted that XLM-RoBERTa generally achieves F1 scores well above chance levels, with the exceptions of only two classes: «pop culture» an «matter and energy», the first of them not being a science-exclusive thematic area. While SVM showed higher precision for certain labels, XLM-RoBERTa excelled in recall, a crucial factor given the class imbalance in our dataset. The significant improvement in recall with XLM-RoBERTa is particularly noteworthy, as it highlights its effectiveness in identifying relevant instances across various labels.

Regarding the individual labels, there are some, e.g. «astronomy and space», for which XLM-RoBERTa and SVM both achieved high precision and demonstrated strong near-perfect F1 scores. This indicates that for certain labels —especially for those more represented in the corpus—, both models perform exceptionally well. However, it is important to note that the SVM's precision values of 1.0 for some labels are not entirely reliable, as it results from zero division in the metrics calculation. This limitation affects the interpretation of these values, and thus no single model is highlighted as a definitive winner for these labels. When it comes to the lower-performing labels, such as «pop culture», «engineering» and «matter and energy», the evaluation metrics indicate that those classes might be particularly difficult to classify. This difficulty could stem from various factors, such as a lack of sufficient samples in the training dataset or inherent complexities in the data associated with these labels. Future work on this topic could benefit from focusing on those classes to build a targeted augmented dataset with additional samples for these specific labels.

Nonetheless, we are confident that our fine-tuned model is well-suited for identifying and classifying PCS tweets across several key scientific and thematic areas within the Mexican context. Specifically, based on evaluation metrics, we present a model that demonstrates strong performance in areas such as astronomy and space, biology, scientific anniversaries, physics, mathematics, medicine and health, and technology. While there are areas for improvement, the model's overall performance suggests it is a valuable tool for processing and categorizing scientific discourse on social media.

5 Conclusions and future work

In this paper, we conducted an automatic text classification task within the context of Public Communication of Science on Twitter/𝕏, based on a multi-label system to categorize the thematic areas identified in the research corpus. Our objective was to develop and evaluate an automatic classification model, from a supervised learning approach. For this purpose, we used an annotated corpus comprising tweets related to PCS published in Mexico as a training and test dataset. The annotation process aimed to identify and categorize the thematic areas related to the tweets' content, using a multi-label system where each tweet could be labeled with one or more of the 18 predefined thematic categories.

Based on this annotated corpus, several machine learning models were trained and evaluated, including both classic algorithms (Support Vector Machine, Random Forest Classifier, Linear Regression, and Decision Tree Classifier) and transformer-based models (DistilBERT, DistilRoBERTa, XLM-RoBERTa, BERT, GPT-2, and BLOOM). These models were tested to determine their efficacy in classifying tweets into the predefined thematic categories, with a focus on achieving high precision, recall, and F1 scores. The evaluation aimed to compare the performance of traditional machine learning techniques with that of transformer-based approaches.

During the dataset evaluation, particular attention was given to recall, which became a significant concern due to the class imbalance in the dataset. In this study, improving recall was crucial for ensuring that all relevant thematic categories were accurately captured. While the classic ML algorithms demonstrated certain strengths, especially in terms of precision, the larger transformer-based models, notably XLM-RoBERTa, showed notable improvements in recall, posing an effective and computationally non-intensive approach to address the challenge of missing positive cases. This finding is particularly useful for managing class imbalance and capturing a broader range of thematic categories in annotated datasets with plenty of labels in NLP.

According to the results reported in this paper, larger pre-trained transformer-based models like XLM-RoBERTa are an effective means to improve recall and overall text classification performance for multi-label annotated datasets. Despite not being the latest advancements in NLP, these models, when properly fine-tuned, can significantly enhance the detection and classification of thematic areas in social media text. By opting not to use the most recent large language models (LLMs), our study emphasizes the value of using well-established transformer-based models in Spanish to ensure computational efficiency and resource management. While newer LLMs might offer even greater advancements, our approach effectively balances performance with practical constraints.

For future work, efforts should focus on addressing the lower-performing labels identified in this study, such as «pop culture», «engineering», and «matter and energy», if future researchers consider them to be PCS-relevant. These areas may benefit from targeted data augmentation or more specialized model fine-tuning to improve classification accuracy. However, as noted, one limitation regarding the corpus data type is the restriction imposed by collecting Twitter data via the API due to changes made on the platform over the past year. Further research is encouraged to explore advanced text processing and machine learning approaches that could enhance performance in multi-label classification tasks for PCS tweets. Overall, this study represents one of the first approaches from NLP and machine learning to classify tweets related to PCS in Mexico. The results should be interpreted as an initial exploration of multi-label text classification in PCS tweets.

References

Aliev, R., Pedrycz, W., Guirimov, B., & Huseynov, O. (2020). Clustering method for production of Z-number based if-then rules. *Information Sciences*, *520*, 155-176.

Bose, M., & Mali, K. (2019). Designing fuzzy time series forecasting models: A survey. *International Journal of Approximate Reasoning, 111*, 78-99.

Castillo, O., Castro, J. R., Pulido, M., & Melin, P. (2022). Interval type-3 fuzzy aggregators for ensembles of neural networks in COVID-19 time series prediction. *Engineering Applications of Artificial Intelligence, 114*, 1-11.

Chen, G. T., & Chen, L. (2010). Forecasting financial crises for an enterprise by using the grey Markov forecasting model. *Springer Science+Business Media B.V.*

Dixit, A., & Jain, S. (2022). Intuitionistic fuzzy time series forecasting method for nonstationary time series data with suitable number of clusters and different window size for fuzzy rule generation. *Information Sciences, 623*, 132-145.

Dixit, A., & Jain, S. (2022). Intuitionistic fuzzy time series forecasting method for non-stationary time series data with suitable number of clusters and different window size for fuzzy rule generation. *Information Sciences, 623*, 132-145.

He, Y., & Huang, M. (2005). A grey-Markov forecasting model for the electric power requirement in China. *IEEE International Conference on Mechatronics and Automation*, 574-582.

Hu, Y.-C. (2017). Predicting foreign tourists for the tourism industry using soft computing-based grey–Markov models. *Sustainability, 9*(1228), 1-14.

Li, M. (2016). Water demand prediction of grey Markov model based on GM(1,1). *3rd International Conference on Mechatronics and Information Technology*, 6.

Lin, L.-C., & Wu, S.-Y. (2013). Analyzing Taiwan IC assembly industry by grey-Markov forecasting model. *Hindawi Publishing Corporation*.

Lin, Y.-H., Lee, P.-C., & Chang, T.-P. (2009). Adaptive and high-precision grey forecasting model. *Expert Systems with Applications, 35*, 9658-9662.

Ma, X., & Liu, Z. (2017). Application of a novel time-delayed polynomial grey model to predict the natural gas consumption in China. *Journal of Computational and Applied Mathematics, 324*, 17-24.

MATLAB. (2023). *Fuzzy Logic Toolbox: User's Guide* (R2023b). Natick, MA: MathWorks.

Mi, J., Fan, L., Duan, X., & Qiu, Y. (2018). Short-term power load forecasting method based on improved exponential smoothing grey model. *Mathematical Problems in Engineering, 2018*, 1-10.

Mierzwiak, R., Xie, N., & Dong, W. (2019). Classification of research problems in grey system theory based on grey space concept. *The Journal of Grey System, 31*(1), 100-111.

Nemati, A., Hashemkhani Zolfani, S., & Khazaelpour, P. (2023). A novel gray FUCOM method and its application for better video games experiences. *Expert Systems with Applications, 234*, 1-20.

Shah, M. (2012). Fuzzy based trend mapping and forecasting for time series data. *Expert Systems with Applications, 39*(7), 6351-6358.
Surowiecki, J. (2005). *The Wisdom of Crowds*. Anchor Books.

Takagi, T., & Sugeno, M. (1985). Fuzzy identification of systems and its applications to modeling and control. *IEEE Transactions on Systems, Man, and Cybernetics, 15*(1), 116-132.

Tanaka, H. (1987). Fuzzy data analysis by possibility linear models. *Fuzzy Sets and Systems, 24*, 363-375.

Tanaka, H., Uejima, S., & Asai, K. (1982). Linear regression analysis with fuzzy model. *IEEE Transactions on Systems, Man, and Cybernetics, 12*(6), 903-907.

Wang, X., Hyndman, R. J., Li, F., & Kang, Y. (2022). Forecast combinations: An over 50-year review. *International Journal of Forecasting*.

Wang, Z.-X., Hipel, K. W., Wang, Q., & He, S.-W. (2011). An optimized NGBM(1,1) model for forecasting the qualified discharge rate of industrial wastewater in China. *Applied Mathematical Modelling, 35*, 5524-5532.

Xiang, L., Wenda, T., & Guangsheng, Z. (2007). Forecast of flood in Chaohu Lake Basin of China based on grey-Markov theory. *Chinese Geographical Science, 17*(1), 5-12.

Zhan-li, M., & Jin-hua, S. (2011). Application of grey-Markov model in forecasting fire accidents. *Procedia Engineering, 11*, 314-318.

Zhou, Y., Ren, H., Zhao, D., Li, Z., & Pedrycz, W. (2022). A novel multi-level framework for anomaly detection in time series data. *Applied Intelligence*, 1-18.