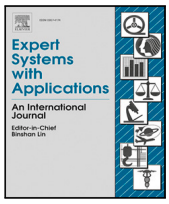
Expert Systems With Applications 281 (2025) 127503

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa



ALPET: Active few-shot learning for citation worthiness detection in low-resource Wikipedia languages

Aida Halitaj *, Arkaitz Zubiaga

School of Electronic Engineering and Computer Science, Queen Mary University of London, 327 Mile End Rd, London E1 4NS, United Kingdom

ARTICLE INFO

Keywords:
Low-resource languages
Wikipedia
Few-shot learning
Active learning
Fact checking
Citation worthiness detection

ABSTRACT

Citation Worthiness Detection (CWD) consists in determining which sentences, within an article or collection, should be backed up with a citation to validate the information it provides. This study, introduces ALPET, a framework combining Active Learning (AL) and Pattern-Exploiting Training (PET), to enhance CWD for languages with limited data resources. Applied to Catalan, Basque, and Albanian Wikipedia datasets, ALPET outperforms the existing CCW baseline while reducing the amount of labeled data in some cases above 80%. ALPET's performance plateaus after 300 labeled samples, showing it suitability for low-resource scenarios where large, labeled datasets are not common. While specific active learning query strategies, like those employing K-Means clustering, can offer advantages, their effectiveness is not universal and often yields marginal gains over random sampling, particularly with smaller datasets. This suggests that random sampling, despite its simplicity, remains a strong baseline for CWD in constrained resource environments. Overall, ALPET's ability to achieve high performance with fewer labeled samples makes it a promising tool for enhancing the verifiability of online content in low-resource language settings.

1. Introduction

The rise of misinformation in the digital age has made automated fact-checking an essential tool in ensuring the reliability of information (Guo, Schlichtkrull, & Vlachos, 2022; Thorne & Vlachos, 2018). Automated fact-checking are complex systems including tasks that involve the identification of information worthy of checking, linking it with evidence and the subsequent verification step (Zeng, Abumansour, & Zubiaga, 2021). While substantial progress has been made in fact-checking systems for major languages, low-resource languages remain largely underexplored and are recently receiving more attention (Halitaj & Zubiaga, 2024; Le, To, Nguyen, & Van Nguyen, 2024).

Misinformation has a global impact (Quelle, Cheng, Bovet, & Hale, 2023), so contributing towards advancement of fact-checking systems for low-resource languages is necessary to maintain information integrity worldwide. However, in doing so, the scarcity of labeled datasets for low-resource languages poses a significant challenge (Gupta & Srikumar, 2021). Many of the existing methods for developing automated fact-checking systems rely on extensive labeled data, which is often unavailable for low-resource languages. Furthermore, in high-resource languages, like English, fact-checking systems have the advantage of focusing on specific domains, such as political discourse, which generates abundant data for verification (Konstantinovskiy, Price, Babakar, & Zubiaga, 2021). In contrast, low-resource languages often lack this advantage, as the volume of digital information available in specialized topics is more limited, affecting the ability to build effective systems within a single domain of a language (Gupta & Srikumar, 2021; Shahi & Nandini, 2020); as a result, we cannot limit ourselves to one specific domain for a low-resource language.

In this research we address this challenge by using Wikipedia data, which spans a wide range of topics and domains. The goal is to develop a method for citation worthiness detection (CWD), where the task is to determine whether a sentence needs to be accompanied with a citation to back it up; an ability to detect sentences needing citation in turn supports the integrity of Wikipedia. Research in CWD has been limited to date, with the majority of efforts focused on English (Redi, Fetahu, Morgan, & Taraborelli, 2019); when it comes to low-resource languages, the only such effort to date was with CCW (Halitaj & Zubiaga, 2024), where we introduced a context-aware model that leverages surrounding sentences and topic categories to help determine if a sentence should be accompanied with a citation.

Furthering the limited research of CWD in low-resource language, here we are the first to propose an active learning strategy specifically tailored to address data scarcity in the task. We introduce a novel method called ALPET (Active Learning with Pattern Exploiting Training), which integrates active learning (AL) strategies with few-shot

* Corresponding author.
E-mail addresses: a.halitaj@qmul.ac.uk (A. Halitaj), a.zubiaga@qmul.ac.uk (A. Zubiaga).

https://doi.org/10.1016/j.eswa.2025.127503
Received 24 January 2025; Received in revised form 13 March 2025; Accepted 29 March 2025
Available online 9 April 2025

learning (FSL) techniques using pre-trained language models (PLMs). This approach is designed to maximize performance with minimal labeled data, addressing the limitations of existing methods for low-resource languages. ALPET strategically selects the most informative samples for labeling, thus reducing the need for extensive labeled datasets, which are often unavailable for low-resource languages. While few-shot learning (FSL) enables models to generalize from a small number of examples (Wang, Yao, Kwok, & Ni, 2020), AL further optimizes this process by focusing on high-value samples, reducing the cost and time associated with manual labeling. Despite the promise of both AL and PET, their integration for low-resource languages in CWD tasks has not been explored until now.

The goal of ALPET is to investigate whether AL strategies, when combined with PET, could improve model performance as opposed to the existing models like CCW (Halitaj & Zubiaga, 2024). Furthermore, in this research we investigate if ALPET can select more informative samples than random selection. While we initially hypothesized that this approach would consistently outperform random sampling, the results indicate that the effectiveness of AL strategies is more nuanced and context-dependent. Despite this, the study provides valuable insights into the challenges of applying AL and FSL methods in low-resource language settings, offering a foundation for future work in optimizing data selection and model performance in these environments.

Our study makes the following contributions:

1. We introduce ALPET, a novel active few-shot learning method that integrates active learning (AL) strategies with few-shot learning (FSL) techniques using pre-trained language models (PLMs) through Pattern Exploit Training (PET). ALPET leverages Wikipedia data to address the data scarcity issue in low-resource languages by strategically selecting the most informative samples for labeling, reducing the dependency on large, labeled datasets.
2. We provide a comprehensive analysis of multiple variants of our proposed ALPET framework, including with various AL query strategies for CWD, comparing their performance against random sampling in a few-shot learning environment. The analysis focuses on the effectiveness of these strategies in identifying informative data points for labeling and its suitability for CWD in low-resource settings, with a particular emphasis on K-Means clustering techniques.
3. The research quantifies the reduction in labeled data achieved by ALPET compared to existing methods like CCW while maintaining comparable performance. This reduction in manual annotation not only improves efficiency but also significantly reduces development costs, making it particularly valuable in low-resource contexts.
4. Within the largely unexplored realm of CWD for low-resource languages, ours is the first effort to incorporate and investigate active learning, in turn comparing with the state-of-the-art method, CCW, and providing a comprehensive analysis of the potential benefits of the proposed strategy.

These contributions offer a deeper understanding of the challenges and potential solutions for CWD in low-resource languages and pave the way for further advancements in the field by establishing a foundation for building effective CWD systems for under-resourced languages.

2. Research objectives and hypotheses

The scarcity of labeled data in low-resource languages presents a major challenge for developing automated fact-checking systems. In this research, we aim to address this issue by exploring how the combination of AL and PET can enhance citation check-worthiness detection for low-resource languages, which in our case we investigate with Albanian, Basque, and Catalan.

The specific objectives (O) of this research, and the hypotheses (H) we set forth in line with the objectives, are:

- **O1:** To investigate whether ALPET can achieve comparable or superior citation worthiness detection performance in low-resource languages (Albanian, Basque, and Catalan) compared to the CCW baseline model, while utilizing fewer labeled examples. This objective directly addresses hypothesis H1, which posits that ALPET will outperform CCW in data efficiency and F1 score in low-resource languages.
  H1: ALPET outperforms the CCW baseline model in terms of data efficiency (fewer labeled examples) and performance (F1 Score) in low-resource languages while using the same AL query strategies.
- **O2:** To determine the optimal number of labeled examples required by ALPET to achieve peak performance in citation worthiness detection for low-resource languages, analyzing whether the model's performance plateaus after a certain number of labeled samples. This objective is directly related to hypothesis H2, which suggests that ALPET's performance will plateau after a certain number of samples, highlighting its effectiveness with small datasets.
  H2: ALPET's performance improves with increasing labeled data but it may reach a plateau at a certain point, suggesting that the method is effective with small sized datasets in low-resource settings.
- **O3:** To quantify the reduction in labeled data achieved by ALPET compared to the CCW baseline model while maintaining comparable citation worthiness detection performance in low-resource languages, assessing the robustness of ALPET's performance with reduced training data. This objective is linked to hypothesis H3, which predicts that ALPET will achieve competitive performance with an average of 58%–72% fewer labeled examples than CCW, demonstrating its robustness in low-resource settings.
  H3: ALPET can match the performance of CCW with far fewer labeled examples in low-resource languages. Its performance stays stable even as the number of labeled examples decreases, showing its robustness in these settings.
- **O4:** To compare the performance of various active learning query strategies against random sampling in selecting informative data points for citation worthiness detection in low-resource languages, evaluating their effectiveness based on the F1 Score. This objective directly addresses hypothesis H4, which states that active learning query strategies generally achieve higher F1 scores than random sampling in low-resource languages.
  H4: Active learning query strategies generally achieve higher F1 scores than random sampling in low-resource language datasets.

3. Background and related work

3.1. Citation worthiness detection (CWD)

Citation worthiness detection, often referred to as the "citation needed" task in the literature, involves identifying whether a sentence within a given corpus requires a citation (Bonab, Zamani, Learned-Miller, & Allan, 2018; Redi et al., 2019). This task is particularly crucial on collaborative platforms like Wikipedia, where maintaining information credibility is essential. By ensuring that unsourced statements are flagged to be properly supported by reliable references, CWD helps prevent the spread of misinformation. CWD in Wikipedia simplifies and speeds up the process of verifiability policy[1] by prioritizing unsourced sentences to be reviewed by editors. This process is vital for maintaining academic and public trust in Wikipedia, which serves as a widely-used reference for both educational and general purposes. It is also critical in fact-checking systems to evaluate the veracity of claims (Sathe, Ather, Le, Perry, & Park, 2020), and in various social

[1] https://en.wikipedia.org/wiki/Wikipedia:Verifiability.

media platforms to combat misinformation (Baigutanova et al., 2023; Halitaj & Zubiaga, 2024).

CWD in low-resource languages has several challenges. One major issue is the availability of credible labeled datasets, as these languages often lack the extensive digital content needed to develop such resources. As a result, most NLP tools and pre-trained models, like BERT and GPT, are optimized for high-resource languages, leading to suboptimal performance in low-resource contexts. Specifically, considering that low-resource languages have variety of dialects that are underrepresented or not captured at all by PLMs. The scarcity of scientific research focused on low-resource languages further amplifies the problem, as most advancements in the field are designed and tested on larger languages.

Existing CWD approaches in Wikipedia are usually defined as supervised learning text classification task. The pioneering work for this task started from the assessment of Wikipedia verifiability policy (Redi et al., 2019) where they used recurrent neural networks (RNN) with GRU cells and GloVe pre-trained word embeddings to identify sentences that needed citation. However, they heavily relied on featured articles,[2] citation needed tag,[3] and manual annotation efforts to create the dataset.

Building upon Redi's work, another approach was proposed utilizing positive unlabeled learning where they aimed to develop an unified approach to check-worthiness detection tasks including claim detection, rumour detection and citation needed (Wright & Augenstein, 2020). While aiming to create a unified solution for these three distinct tasks, authors also aimed to reduce manual labeling effort by asking annotators to mark only sentences that were clear-cut check-worthy and the rest to be handled through positive unlabeled method. They used transfer-learning to transfer the knowledge from one task to another and different from more traditional neural networks used in Redi's approach, in this research they used pre-trained BERT model.

Both studies (Redi et al., 2019; Wright & Augenstein, 2020) focused primarily on English, leveraging featured articles and *{citation needed}* tags added by active editors. These methods, however, are not applicable to low-resource languages due to the scarcity of featured articles and the absence of such tags, resulting from a lack of active editorial communities. To overcome this challenge, a recent study proposed an approach to use quality score of articles to automatically build a credible datasets for low-resource languages (Halitaj & Zubiaga, 2024). Unlike previous work, which relied solely on the sentence text and its section placement within an article, this study used adjacent sentences as contextual information and employed mBERT for the final classification of sentences needing citations. While their approach advances CWD in low-resource languages, it relies on substantial amount of data due to the need for contextual information from adjacent sentences; and the automated large-scale labeling process, although innovative, carries the challenge of potential inaccuracies, as it cannot fully ensure the correctness of every label. Incorporating an oracle in the labeling process through active learning could address these issues by enhancing the credibility of the dataset. Building on this foundation, we propose an approach that integrates cold-start pool-based active learning with few-shot learning using Pattern Exploit Training (PET), which reduces the data requirements for effective CWD and minimizes dependence on additional contextual information.

3.2. Active learning in NLP

Active learning (AL) is a machine learning approach where the algorithm uses a querying strategy to identify the most informative data points for labeling by an oracle, to improve model's performance (Settles, 2009). The goal is to overcome the labeling bottleneck of traditional passive learning systems by optimizing the model's performance with a smaller set of labeled examples, making the learning process more efficient and cost-effective. This approach is particularly valuable in scenarios with limited labeled data, such as low-resource Natural Language Processing (NLP) tasks.

AL involves two key concepts: problem scenarios and query strategies. Scenarios define the learning environment, including how data is presented and how the model interacts with it, while query strategies determine which data points to label within that scenario. Some query strategies can be applied across multiple scenarios, while others are scenario-specific. For an AL system to be effective, it is essential to match the appropriate query strategy with the right scenario. According to the existing literature some of the main AL scenarios are membership query synthesis (Angluin, 1988), stream-based selective sampling (Cohn, Atlas, & Ladner, 1994), pool-based sampling (Lewis & Gale, 1994), batch AL (Brinker, 2003), and multi-task AL (Acharya, Mooney, & Ghosh, 2014; Harpale & Yang, 2010; Reichart, Tomanek, Hahn, & Rappoport, 2008). Each scenario offers distinct advantages and has seen different levels of application in NLP. In what follows we will briefly introduce the above mentioned AL scenarios.

1. *Membership Query Synthesis (MQS)*. It is one of the earliest AL scenarios (Baum & Lang, 1992) which enables the generation of artificial examples to expand datasets based on defined feature dimensions. Initially endorsed for automated oracles (King et al., 2009, 2004) due to the challenges humans faced in interpreting synthetic examples, MQS has recently been applied to simple text classification tasks (Quteineh, Samothrakis, & Sutcliffe, 2020; Schumann & Rehbein, 2019). Nonetheless, its use remains limited, specifically for more complex NLP tasks with unbalanced data (Schumann & Rehbein, 2019; Zhang, Strubell and Hovy, 2022).
2. *Stream-Based Selective Sampling*. Also known as online AL (Cacciarelli & Kulahci, 2024), this scenario involves data arriving in a continuous stream, requiring the learner to decide in real-time whether to request the label (Settles, 2009). Query strategies like uncertainty sampling, threshold-based decision, query by committee are commonly employed to make this decision effective. This scenario has been applied in several NLP tasks, including part-of-speech tagging (Dagan & Engelson, 1995), named entity recognition (Van Tran, Nguyen, Hoang, Hwang, & Nguyen, 2017), sentiment analysis (Kranjc et al., 2015; Smailović, Grčar, Lavrač, & Žnidaršič, 2014) etc.
3. *Pool-Based Sampling*. This scenario, perhaps the most extensively researched in NLP (Settles, 2012), involves selecting the most informative samples from a large, static pool of unlabeled data (Lewis & Gale, 1994). It has been successfully applied in a wide range of NLP tasks, including text classification (Lewis & Gale, 1994; Siddhant & Lipton, 2018; Tong & Koller, 2001; Zhang, Feng and Tan, 2022; Zhang, Lease, & Wallace, 2017), speech recognition (Tur, Hakkani-Tür, & Schapire, 2005), named entity recognition (Dossou et al., 2022; Radmard, Fathullah, & Lipani, 2021; Tsvigun et al., 2022), part of speech tagging (Mendonça, Sardinha, Coheur, & Santos, 2020; Ringger et al., 2007; Stratos & Collins, 2015), word sense disambiguation (Alagić & Šnajder, 2015; Dligach & Palmer, 2011; Zhu, Wang, Yao, & Tsou, 2008), machine translation (Chimoto & Bassett, 2022; Zhao, Zhang, Zhou, & Zhang, 2020), language understanding (Grießhaber, Maucher, & Vu, 2020), and prompt engineering (Qian et al., 2024). The versatility of this approach is further enhanced by combining it with deep learning techniques like transfer-learning (Mamooler, Lebret, Massonnet, & Aberer, 2022; Zhang, Feng et al., 2022), semi-supervised learning (Imamura, Takayama, Kaji, Toyoda, & Kitsuregawa, 2009; Tsvigun et al., 2023), weak supervision (Brantley, Sharaf, & Daumé III, 2020; Qian, Raman, Li, & Popa, 2020; Zhang, Yu,

[2] https://en.wikipedia.org/wiki/Wikipedia:Featured_articles.
[3] https://en.wikipedia.org/wiki/Wikipedia:Citation_needed.

Shetty, Song and Zhang, 2022), data augmentation (Dossou et al., 2022; Zhao et al., 2020), and few-shot learning (Bayer & Reuter, 2024; Müller, Pérez-Torró, Basile, & Franco-Salvador, 2022; Qian et al., 2024; Zeng & Zubiaga, 2022; Zhu, Yadav, Afzal, & Tsatsaronis, 2022), resulting in development of various query strategies adhering to deep neural network (DNN) characteristics (Gal & Ghahramani, 2016). However, its application in more specialized tasks, such as citation worthiness detection, remains underexplored.

4. *Batch Active Learning.* Unlike traditional AL, where data points are queried one by one, batch AL queries multiple data points simultaneously to increase efficiency and reduce the number of iterations needed for model improvement (Citovsky et al., 2021; Settles, 2009). This approach has been applied in various NLP tasks, including text classification (Beatty, Kochis, & Bloodgood, 2018), machine translation (Ananthakrishnan, Prasad, Stallard, & Natarajan, 2010; Shi, Benton, Malioutov, & Irsoy, 2021), rumor detection (Farinneya, Pour, Hamidian, & Diab, 2021), and named entity recognition (NER) (Shen, Zhang, Su, Zhou, & Tan, 2004; Shi et al., 2021). However, ensuring diversity among selected examples remains a challenge (Ash, Zhang, Krishnamurthy, Langford, & Agarwal, 2019; Hoi, Jin, & Lyu, 2006), particularly in complex NLP tasks with unbalanced data, where similar sentences and repetition can reduce information gain.
5. *Multi-Task Active Learning (MTAL).* It handles multiple tasks simultaneously, improving efficiency by allowing related tasks to share data and enhancing performance; for distinct tasks by ensuring comprehensive annotations (Reichart et al., 2008; Settles, 2009). It has been applied in NLP tasks such as role labeling (Ikhwantri et al., 2018), dependency parsing (Rotman & Reichart, 2022), named entity recognition (Rotman & Reichart, 2022; Zhou, Cai, Zhang, Guo, & Yuan, 2021), and natural language understanding (Zhu, Ye, Luo, & Zhang, 2020). While challenges like data scarcity and annotation complexity exist across different AL scenarios, they are magnified in MTAL due to the need to manage multiple tasks simultaneously. This adds layers of complexity and resource demands that are less noticeable in single-task scenarios, making it less suited for complex tasks like those in fact-checking.

Pool-based sampling is one of the most used AL query strategy in NLP tasks due to its effectiveness in minimizing labeling costs by selecting only the most informative data points. This approach allows modern NLP models, specifically PLMs, such as transformers (e.g., BERT (Devlin, 2018), GPT (Radford, Narasimhan, Salimans, Sutskever, et al., 2018), T5 (Raffel et al., 2020)), to achieve high performance with limited resources (Dor et al., 2020; Li et al., 2024; Lu et al., 2023; Yao et al., 2023; Zeng, 2024), particularly in low-resource languages and domains (Grießhaber et al., 2020; Kasai, Qian, Gurajada, Li, & Popa, 2019; Li et al., 2024; Maekawa, Zhang, Kim, Rahman, & Hruschka, 2022; Öhman, 2021; Ye, Liu, Pavani, & Dasgupta, 2023; Zhou & Waibel, 2021). It also facilitates rapid domain adaptation (Lu et al., 2023) and enhances human-in-the-loop systems by focusing efforts on the most impactful examples (Yao et al., 2023). Yet, its application in more specialized tasks, such as citation worthiness detection, remains underexplored – a gap we aim to address in our research.

While AL has been used in related areas like misinformation detection (Barnabò et al., 2023), rumour detection (Farinneya et al., 2021), claim verification (Zeng & Zubiaga, 2022), its use in CWD in Wikipedia settings is limited. These studies primarily focus in specific domains like political fact-checking and English language, rather than the broader, domain-agnostic context of Wikipedia. The gap in the literature suggests a need for further exploration in applying AL to CWD task, particularly with an emphasis on language diversity and resource constraints.

In the context of fact-checking, AL has been combined with models like PET to improve the accuracy of claim verification (Zeng & Zubiaga, 2022). However, similar integration for citation worthiness detection or claim detection (as a related task), especially within low-resource language settings, is not common. Our work seeks to address this gap by exploring different pool-based AL query strategies in conjunction with PET models to improve CWD in low-resource settings.

In the next section we explain in more details our proposed methodology.

4. Methodology

The proposed methodology, referred to as ALPET, integrates Active Learning (AL) and Pattern Exploit Training (PET) to create an efficient approach for data selection and model training in low-resource setting. It is structured into four key steps:

1. Data selection: the process begins with applying pool-based active learning strategies to select informative data points from a large pool of unlabeled sentences (see Section 4.1).
2. Data processing: the selected data is processed to remove redundancy when it exists such as duplicates and highly similar sentences, ensuring dataset diversity (see Section 4.2).
3. Multi-round dataset preparation for FSL: multiple rounds of datasets are then created with incremental sample sizes for each active learning strategy (see Section 4.3).
4. Model training: the PET model with mBERT is used to train and classify sentences as needing citations (see Section 4.4).

Fig. 1 provides a detailed flowchart of the ALPET model architecture, illustrating each of the steps mentioned above including an example of doing CWD with PET.

4.1. Data selection

The CWD task is framed within a pool-based AL scenario using various query strategies. Central to these strategies is the acquisition function, a term commonly used in mathematical definitions to describe the mechanism that assigns a score to each unlabeled data point based on specific criteria. In the context of AL, this function is often referred to as a query framework or query strategy (Settles, 2009). Traditionally, acquisition functions in AL are based on uncertainty or diversity, with more recent approaches incorporating hybrid methods that combine elements of both (Margatina, Vernikos, Barrault, & Aletras, 2021). In this work, we employ uncertainty, diversity, and hybrid acquisition functions as part of our data subset selection process. Fig. 2 illustrates the flowchart of data selection strategies.

Since ALPET automates the data selection process using active learning query strategies, it minimizes the need for direct human annotation. Instead of relying on manual labels, we use multilingual BERT as a proxy oracle, where the model assigns labels to the selected samples. While this eliminates user annotation efforts, it is important to recognize that the proxy model may introduce biases inherent to its pretraining and fine-tuning data. Unlike traditional AL setups where a human annotator verifies each query, our approach fully automates the labeling pipeline. This ensures efficiency but also raises considerations regarding potential biases and label reliability. Future extensions of ALPET could explore integrating human-in-the-loop mechanisms to assess the trade-offs between automation and annotation quality.

Next we elaborate these query strategies in more details.

4.1.1. Diversity sampling

The goal here is to select data points that are as different from each other as possible ensuring a representative sample of the overall data distribution. Three diversity sampling methods that we use in this work are geometry-based, corset-based, and cluster samplings.

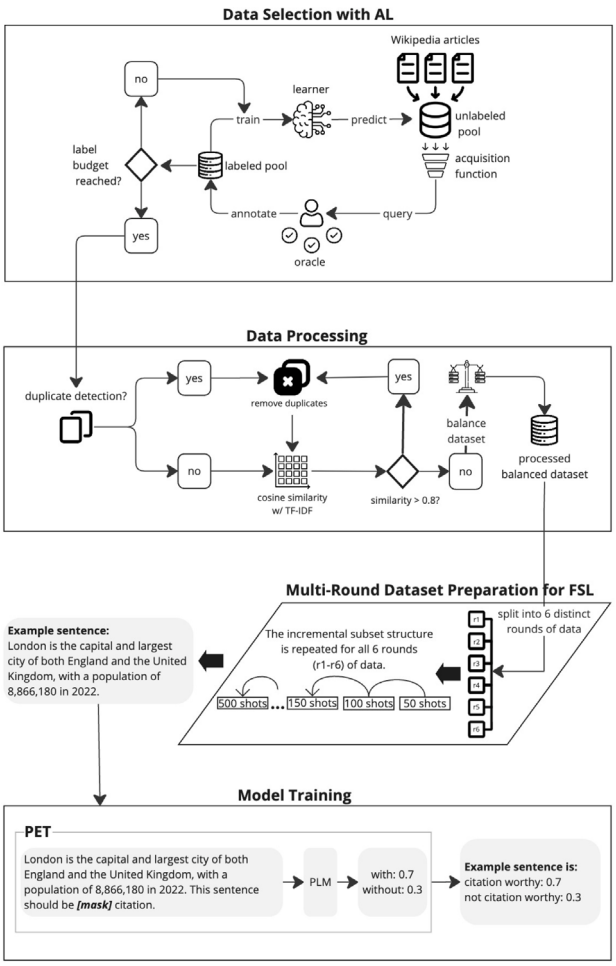


Fig. 1. ALPET model architecture.

A. **Geometry-based sampling.** Distance metrics belong in the geometry-based metrics, they play a fundamental role in identifying data points for labeling in the active learning scenarios. Among most widely used distance metrics are cosine and euclidean distance. We have used both of them in ALPET methodology and more details are presented below.

1. **Cosine distance** is derived from cosine similarity and measures dissimilarity between data points. In our task, sentences are represented by embeddings, or high-dimensional vectors, to capture the semantic content of the sentence. Cosine distance between two vectors **A** and **B** is mathematically defined as:

$$\text{Cosine Distance} = 1 - \cos(\theta) = 1 - \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \times \|\mathbf{B}\|} \quad (1)$$

where $\mathbf{A} \cdot \mathbf{B}$ is the dot product of the vectors, and $\|\mathbf{A}\|$ and $\|\mathbf{B}\|$ are their magnitudes.

Our goal in using cosine distance is to select sentences for labeling that represent different contexts where a citation might be needed. Sentences that are highly dissimilar from those already labeled might represent different styles, topics, or structures that the model has not yet encountered. We avoided using cosine similarity because we did not want to end up selecting sentences that are very similar to those already labeled as this could lead to redundancy in the training data, where the model continues to see variations of the same sentence structure or topic. In contrast, cosine distance encourages diversity in the selected data points, which we believe is crucial for effectively training the model in an active learning setting for the complex task of CWD.

2. **Euclidean distance** measures the straight-line distance between two points in a high-dimensional space. In our task, this metric is used to quantify the absolute difference between sentence embeddings, providing a direct measure of dissimilarity. The Euclidean distance between two vectors **A** and **B** is mathematically defined as:

$$\text{Euclidean Distance} = \|\mathbf{A} - \mathbf{B}\| = \sqrt{\sum_{i=1}^{n} (a_i - b_i)^2} \quad (2)$$

where a_i and b_i are the components of vectors **A** and **B** in an n-dimensional space.

Using euclidean distance, we aim to select sentences that are different from those already labeled, thus exploring underrepresented areas of the feature space. Unlike cosine distance, which focuses on the angular relationship between vectors, euclidean distance accounts for the overall magnitude of differences, making it particularly useful for detecting sentences that are not only contextually but also substantially different in their feature representations (such as length, complexity, or intensity). This diversity in selection might help minimizing redundancy and enhancing the model's ability to generalize across different types of citation-worthy content.

Building on distance metrics, in this research, we have used several custom data selection strategies to guide the active learning process. Each method leverages either cosine or euclidean distance to identify the most informative data points.

i. **Maximum average distance selection**: In this method, data points are selected based on the maximum average distance from already selected instances. The process begins with a cold start, where the first sentence is selected randomly from the unlabeled pool. For the next selection, the average distance of each candidate sentence (that has not been selected yet) to the already selected sentence is calculated. The sentence with the maximum average distance to the already selected sentence is chosen next. With two sentences selected, we recalculate the average distance of the remaining sentences to both selected sentences. The candidate with the highest average distance is added to the selection pool. This loop continues until the desired number of sentences is selected. By focusing on instances that are farthest from those previously selected, this approach aims to select points that are diverse and representative of different regions of the feature space.

ii. **Minimum average distance selection**: This method selects data points with the minimum average distance from the already selected instances. The selection process is similar to the one above that consideres the maximum average distance, but instead, here we consider the minimum average distance. The rationale here is to focus on data points that are similar to those already chosen, effectively reinforcing the model's understanding of dense or well-represented areas of the feature space.

iii. **Combined maximum and minimum average distance selection**: This method alternates between selecting data points based on the maximum and minimum average distances from the already selected instances. The process begins with a cold start, where the first sentence is randomly chosen from the unlabeled pool. For the next selection, the average distance of each candidate

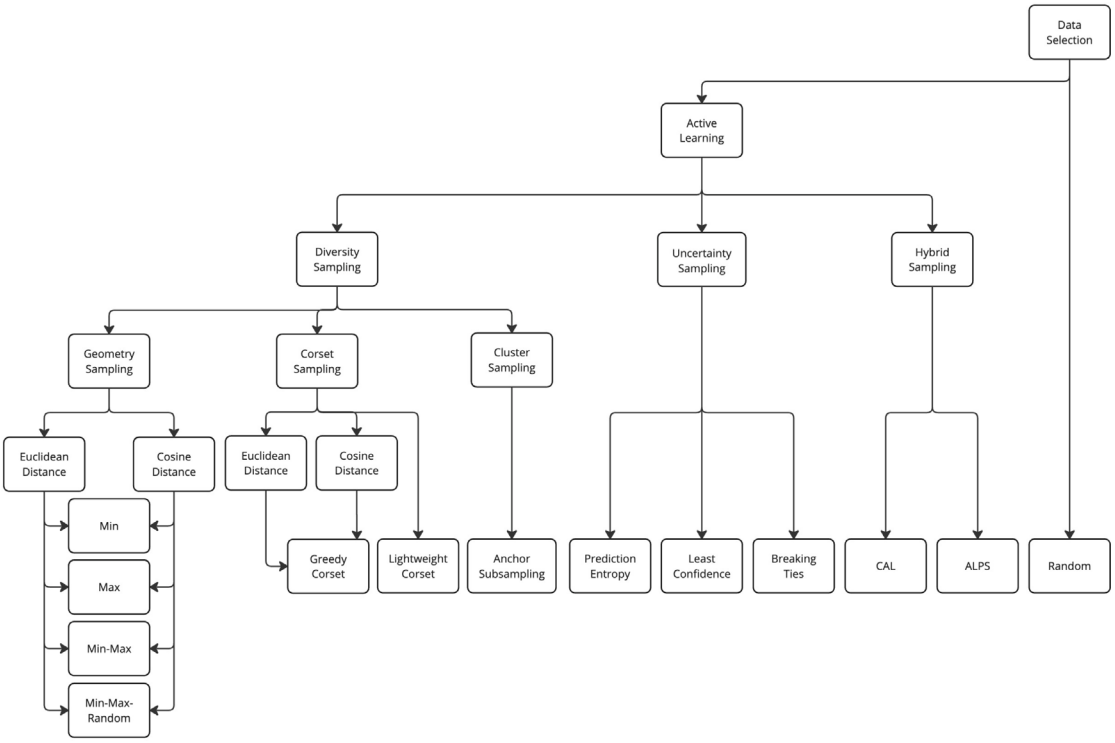


Fig. 2. Data selection flowchart.

sentence to the already selected sentence is calculated. The sentence with the maximum average distance to the selected instance is chosen first. Once this sentence is added to the selection pool, the next sentence is selected based on the minimum average distance from the already selected sentences. With three sentences now in the pool, the average distance of the remaining candidates to the selected sentences is recalculated. The selection alternates between choosing the sentence with the maximum average distance and the sentence with the minimum average distance, creating a balance between diversity and representativeness. This loop continues until the desired number of sentences is selected. By combining both maximum and minimum average distances, it ensures that both diverse and representative instances are included in each iteration, covering a broad spectrum of the data space while also reinforcing existing knowledge.

iv. **Maximum, minimum, and random selection**: In this method, we introduce an element of randomness. Each iteration involves selecting one data point with the maximum average distance, one with the minimum average distance, and one randomly chosen instance. This approach adds an exploratory component, allowing the model to occasionally consider unexpected instances that may not fit neatly into the established patterns, leading to accidental discoveries.

For each custom distance-metric selection strategy, the data points are initially selected from the unlabeled pool following the methodologies outlined above considering either cosine or euclidean distance. Once the selection process is complete, the selected data points are submitted to the oracle for labeling, with the labels being generated based on existing data. After the annotation has been finalized, the labeled data points are then input into the PET model to determine whether each sentence requires a citation.

B. **Corset-based strategies:** In AL a corset is a small, representative subset of the entire dataset that, when used to train a model, can approximate the performance of a model trained on the full dataset (Sener & Savarese, 2017). In our methodology we have used lightweight corset and greedy corset (with cosine and euclidean metrics).

i. **Greedy coreset**: Originally proposed to address the data labeling bottleneck for deep convolutional neural networks (Sener & Savarese, 2017), this approach has been adapted for text data in the small-text library (Schröder, Müller, Niekler, & Potthast, 2023). It constructs a greedy coreset over text embeddings, solving a k-center problem through a greedy approximation. The method aims for precise coverage and diversity by selecting points that minimize the maximum distance between any data point and its closest coreset point. While more computationally intensive, it is particularly useful when a highly representative subset is crucial, such as in smaller datasets.

ii. **Lightweight coreset:** This strategy selects a representative subset of data points using K-Means clustering, designed for computational efficiency by approximating the selection process (Bachem, Lucic, & Krause, 2018). It works by choosing data points that are farthest from the already selected points, minimizing redundancy. While this approach is efficient, it may not capture the full diversity of the data as precisely as more computationally intensive methods like GreedyCoreset (Schröder et al., 2023). The method's effectiveness depends on the quality of the feature representations, such as text embeddings.

C. **Clustering-based strategies:**

i. **Anchor subsampling:** This strategy addresses pool-based AL challenges of selecting minority class in large imbalanced datasets (Lesci & Vlachos, 2024). The term anchor here refers to the chosen class-specific instances from the

labeled set. The process begins by selecting anchor points from the labeled dataset, representing different classes of the data space. Each unlabeled instance is then scored based on its average distance from the anchors, forming a subpool of the most similar instances. The next step is selecting the instances from the subpool to be labeled by an oracle. Once labeled, instances are added to the labeled dataset and the process repeats in subsequent AL iterations.

4.1.2. Uncertainty sampling

Uncertainty-based selection focuses on identifying data points where the model is most unsure of its predictions.

i. **Prediction entropy:** This query strategy selects instances with the largest prediction entropy (Schröder et al., 2023). As a method it was initially proposed to reduce the labeling efforts in an object recognition task (Holub, Perona, & Burl, 2008), however, in small-text library it was adapted for text-based data. Given a pool of unlabeled data, for each instance the model outputs a probability distribution over all possible classes. Then the entropy of the predicted probability distribution is calculated for each instance which are then ranked by the highest entropy scores indicating cases where the model is most uncertain about the prediction. The top k instances with the highest entropy scores are selected for labeling. Once labeled, instances are added to the training set, and the model is retrained on the updated labeled dataset.

ii. **Least confidence**: Is one of the earliest uncertainty-based strategies (Lewis, 1995). It selects instances with the least prediction confidence (regarding the most likely class) (Schröder et al., 2023). Specifically, for each instance in the unlabeled pool, the model assigns a probability distribution over the possible classes. Then it identifies instance where the model is least confident about the correct class. This instance is then selected for labeling because it represents a point of high uncertainty, where the model might benefit most from additional labeled data. However, this method may sometimes overlook instances where the model is uncertain between several classes because it only considers the confidence of the most likely class.

iii. **Breaking ties**: This function is designed to select data points which have a high uncertainty in classification, specifically those where the margin between the most likely and second most likely predicted class is minimal. The small margin indicates that the model is uncertain about which class the sentence belongs to, making these instances particularly valuable for AL. This strategy was originally proposed in the context of image data (Luo et al., 2005) but in small-text library it has been adapted to work with text data under the name BreakingTies (Schröder et al., 2023). The core idea to target ambiguous instances remains the same.

4.1.3. Hybrid sampling

Hybrid acquisition functions were developed to combine the best aspects of both uncertainty and diversity sampling. In this research, we have used Contrastive Active Learning (CAL) and ALPS (Active Learning by Processing Surprisal), two state-of-the-art hybrid methods, to maximize the efficiency of the AL process.

i. **Contrastive active learning (CAL)**: This function combines uncertainty and diversity sampling for warm-start AL. Using ContrastiveActiveLearning from small-text library (Schröder et al., 2023), we implement CAL (Margatina et al., 2021), which fine-tunes BERT with an initial labeled dataset and applies a KNN algorithm to identify the closest labeled examples for each data point in the unlabeled pool. A contrastive score, based on Kullback–Leibler (KL) divergence between predicted probabilities of the unlabeled candidate and its labeled neighbors, is calculated to select high-divergence examples for labeling by a proxy. The labeled batch is then removed from unlabeled pool and added to the training (labeled) dataset. This loop repeats until all unlabeled data have been labeled. This method aims to effectively identify sentences with similar vocabulary but differing predictions, enhancing the selection of informative examples.

ii. **Active learning by processing surprisal (ALPS)**: This method combines uncertainty and diversity sampling for cold-start AL. Implemented as EmbeddingKMeans in small-text library (Schröder et al., 2023), ALPS (Yuan, Lin, & Boyd-Graber, 2020) uses surprisal embeddings derived from the masked language modeling loss in PLMs like BERT to estimate uncertainty, bypassing the need for unreliable model confidence scores in the cold-start scenario. After computing surprisal embeddings for each sentence in the unlabeled pool, K-Means is applied to cluster these embeddings and the sentence closest to each cluster center is selected. Thus ALPS identifies data points that are both surprising (indicating high uncertainty) and representative of diverse, underexplored areas in the data space, making it particularly effective in early stages where labeled data is scarce.

4.2. Data processing

4.2.1. Duplicate and similarity removal

In cold-start AL iterations, after each data subset selection, we observed duplicate sentences or sentences with high structural similarity. To mitigate redundancy in the dataset, we applied a cosine similarity with TF-IDF weights. Sentences with a cosine similarity score above 0.8 were considered too similar and were excluded from the final dataset. This threshold was selected based on our empirical analysis of the dataset, where we observed that sentences with a similarity score above 0.8 were almost identical except for minor variations, such as differing in only the last one or two words.

To assess the impact of this filtering step, we ran ALPET without cosine similarity filtering. In this case, we observed a slight increase in model performance, which aligns with our expectation that the model benefits from learning patterns in structurally similar sentences. Without filtering, the model consistently predicted the same label for these near-duplicate sentences, reinforcing learned patterns. However, after removing such cases, performance declined, likely due to the reduced exposure to repetitive structures. For our CWD task, we aim to capture diverse sentence structures and contexts, so such highly similar sentences were unnecessary thus we opted for filtering.

4.2.2. Data balancing

This research was conducted using Wikipedia articles where each was split into individual sentences, forming the unlabeled pool of data for AL. Although the sentences were pre-labeled, we temporarily removed the labels to simulate an active learning environment where labels are obtained iteratively from an oracle.

Each languages' dataset (ca, eu, and sq) is imbalanced, with higher proportion of sentences that do not contain citations compared to those that do. In pool-based AL scenarios, imbalance can become more pronounced due to the model's tendency to favor majority class examples, leading to a loop of oversampling the majority class. To address this, we applied random undersampling to the majority class post-selection, reducing its size to match the minority class. In this way the minority class remained sufficiently represented throughout the learning process which is critical for our task.

4.3. Multi-round dataset preparation for few-shot learning

After data processing steps, each query strategy yielded a dataset with 3000 data points per class. We then constructed six distinct rounds of datasets. Each round comprises ten subsets, with the number of data points per class varying from 50 to 500, increasing in increments of 50. The dataset preparation process involved the following steps:

1. **Round-based data partitioning**: We randomly separated the 3000 data points per class into six distinct groups, each containing 1000 total data points (500 per class), to allow for multiple experimental iterations. The six rounds were chosen in line with standard practices in machine learning experiments, where 5 to 10 rounds are typically used to ensure reliable and generalizable results.
2. **Incremental sample sizes**: For each group created in the first step, we generated ten cumulative subsets, with sample sizes ranging from 50 to 500 data points per class with increments of 50. This incremental approach is commonly used in few-shot learning experiments, as it allows for a fine-grained analysis of the model's performance across different levels of data availability.

4.4. Model training with Pattern Exploit Training (PET)

Pattern-Exploiting Training (PET) (Schick & Schütze, 2020) is a semi-supervised method for few-shot learning in NLP tasks like text classification and natural language inference. The core idea behind PET is to reformulate input examples as cloze-style questions (fill-in-the-blank), that help PLMs better understand the task. This method uses a concept called Pattern-Verbalizer Pair (PVP) which includes two elements:

1. **Pattern** where the input is transformed into a fill-in-the-blank questions. This is done by inserting a masked token into the input text, which the model will later try to predict.
2. **Verbalizer** maps task labels to actual words in the language model's vocabulary. These words are what the model predicts to fill in the blank created by the pattern.

Example:

- **Input**: "This movie was amazing".
- **Pattern**: The input is transformed into "This movie was [mask]".
- **Verbalizer**: The words from model's vocabulary "great" (positive), "bad" (negative), and "okay" (neutral) are mapped to the task labels (positive, negative, neutral).

In this example, the model predicts the word that should replace *[mask]* based on the context of the sentence. The selected word is then compared to the verbalizer to determine the sentiment classification (e.g., if the model predicts "great", the sentiment is classified as positive).

5. Experimental settings

5.1. Datasets

To validate our hypothesis, we use real-world data sourced from Wikipedia articles. Specifically, we employ three datasets in different languages: ca-citation-needed, eu-citation-needed, and sq-citation-needed (Halitaj & Zubiaga, 2024) in Catalan, Basque and Albanian, respectively. These datasets contain contextual information beyond individual sentences and labels. For this study, however, we focused exclusively on two components: the text of sentences from Wikipedia articles and their labels indicating the presence of inline citations. The sentence text was used for data selection through various AL query strategies, while the labels served primarily to simulate the annotation process with oracles when necessary.

Table 1
Distribution of datasets and their usage across three languages. Numbers present the number of sentences (r1–r6 denote six rounds of distinct train sets).

Dataset	Data partition	Citation	No citation
ca	Unlabeled pool	335,538	802,052
eu		73,086	232,372
sq		29,928	77,105
ca/eu/sq	Train Sets r1–r6	500 each r1–r6	500 each r1–r6
	Dev Set	250	250
	Test Set	250	250

For each dataset — ca-citation-needed, eu-citation-needed, and sq-citation-needed — we applied a consistent labeling budget and data split, dividing the data into training, development, and testing sets.

In the context of FSL, where models are trained on very limited data, even slight changes in the test or development set sizes can sometimes affect performance (Wang et al., 2020). Thus, in our experiments, we tested different sizes of the development and test sets, even though improvements in learning performance were not necessarily expected. The primary purpose of this approach was to assess how varying the size of these sets might influence the stability of our model's evaluation metrics. Although the stability and performance remained largely unchanged, we observed an increased in time and resource consumption when the test and development sets were used at their maximum capacity. Therefore, in our results, we report only the experiments where the number of shots in the test and development sets were limited. The data selection for these reduced sets was done randomly. Table 1 presents the details of the datasets used for the three languages, including their splits into training, development, and test sets. As described in Section 4.3 for training PET models of each language we have created 6 distinct training datasets, each used to train a separate model. But we have evaluated each model using the same development and test datasets. The results of each language for all models per specific shots are then averaged and reported.

5.2. PET with active samples

Models that can be used with PET tasks are PLMs and in our CWD task we employed a multilingual BERT (mBERT) to calculate probabilities of candidate tokens that could replace *[mask]* in predefined patterns for each of the datasets we used. Since we are working with datasets of three languages, we had to manually pre-define patterns for each language. In order to avoid introducing any bias in any of the languages we decided to use the same pattern structure for three languages but we translated them accordingly to match the language. Even though the goal of this research is not to find the most optimized patterns and verbalizer for PET, we experimented with a couple of patterns and we choose the best performing ones to report the final results on. It is worth mentioning that we started experimenting with patterns in Albanian language, then we translated them into Catalan which was quite straight forward. More challenging was translation of patterns into Basque language due to the grammar rules and structure of the language. We could not automatically just translate them with tools like Google Translate, instead we had to amend them manually in order for the patterns to make sense and to make sure that it follows grammatical rules of the language. The patterns we used for each language are presented in Table 2.

5.3. Baseline model

This study seeks to evaluate the efficiency of citation worthiness detection (CWD) within few-shot learning settings to accommodate languages with low-resources using the ALPET method.

We benchmark our work against the Contextualized Citation Worthiness (CCW) model (Halitaj & Zubiaga, 2024) due to its SOTA approach in addressing CWD in the Wikipedia domain. The baseline

Table 2
Patterns used for PVP of PET model in three of the datasets ca-citation-needed, eu-citation-needed, sq-citation-needed.

Language	Verbalizer	Pattern
ca	0 : sense 1: amb	1. *[text_a]* Aquesta frase va *[mask]* citació. 2. *[text_a]* Aquesta frase s'hauria d'escriure *[mask]* citació. 3. *[text_a]* s'hauria d'escriure *[mask]* citar les fonts. 4. *[text_a]* En un article de Viquipèdia, aquesta frase sería *[mask]* citació. 5. Si *[text_a]* fós part de Viquipèdia, els editors requeririen que s'escrigués *[mask]* citació per a que fosi verificable.
eu	0 : gabe 1: barne	1. *[text_a]* Esaldi honetan erreferentzia *[mask]*. 2. *[text_a]* Esaldi hau erreferentzia *[mask]* idatzi beharko litzateke. 3. *[text_a]* idaztean erreferentzia *[mask]*. 4. *[text_a]* Wikipediako artikulu batean, esaldi honetan erreferentzia *[mask]*. 5. *[text_a]* Wikipedian balego, egiaztagarritasuna mantentzeko editoreek gomendatuko lukete erreferentzia *[mask]* idaztea.
sq	0 : pa 1: me	1. *[text_a]* Kjo fjali duhet të jetë *[mask]* citim. 2. *[text_a]* Kjo fjali duhet të shënohet *[mask]* citim.' 3. *[text_a]* Duhet të shënohet *[mask]* burim informacioni. 4. *[text_a]* Në një artikull të Wikipedias, kjo fjali duhet të jetë *[mask]* citim. 5. Nëse *[text_a]* do të ishte fjali e Wikipedias, editorët do të kërkonin të ishte *[mask]* citim për shkak të verifikueshmërise.

Table 3
Model percentage improvement by instance range count.

Dataset	Model	50–100	100–300	300–500
EU	ALPET	3.0%	5.3%	0.9%
SQ	ALPET	2.5%	1.2%	1.2%
CA	ALPET	1.4%	5.3%	1.9%

Table 4
Summary of reduction percentages in labeled examples for ALPET compared to CCW.

	Reduction (ca)	Reduction (eu)	Reduction (sq)
mean	70%	58%	72%
std	10%	24%	21%
median	67%	67%	78%
75p	75%	75%	83%

incorporates a transformer-based architecture, utilizing contextual embeddings from mBERT to capture sentence-level features. To ensure methodological consistency and comparability, we adapted CCW by adding an AL step for data selection. This modification aligns with our ALPET approach, and it allow for a more direct performance comparison between the models in terms of both accuracy and efficiency across languages like Albanian, Basque, and Catalan.

5.4. Evaluation metrics

To assess and compare the performance of the proposed ALPET method against the baseline CCW, we employ the macro F1 score as the primary evaluation metric. Given the balanced nature of our dataset and the equal importance of both classes — positive (sentences requiring inline citations) and negative (sentences not requiring inline citations) — the macro F1 score is well-suited for this task as it ensures that the performance across both classes is equally represented.

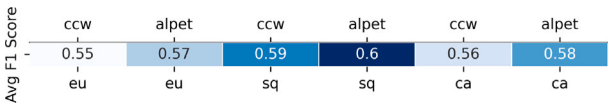


Fig. 3. F1 Score averaged across all query strategies and instances count.

5.5. Training details

Hyperparameters. The ALPET method was trained using the following hyperparameters. We utilized the multilingual BERT with a maximum sequence length of 256 tokens and a batch size of 4 for training and 8 for evaluation. The model was trained for 3 epochs, with a learning rate of 1e–5 and a weight decay of 0.01. Optimization was handled by the Adam optimizer with an epsilon value of 1e–8 and a maximum gradient norm of 1.0 to prevent gradient explosion. The same hyperparameters, where applicable, were used for the baseline CCW model. All experiments were executed on a GPU setup to handle the computational demands of training the BERT model.

Checkpoints. Training was repeated for 3 iterations to ensure robustness in the results. Checkpoints were saved after every epoch allowing the model to resume training from the last saved checkpoint.

Labeling budget. Our final labeling budget was 3000 samples per class for each AL query strategy. From which then we created six distinct rounds of datasets as described in Section 4.3. For each dataset the maximum labeling budget was set to 500 instances per class, and we experimented with scenarios ranging from 50 to 500, increasing by steps of 50.

The task was conducted in a pool-based AL scenario, utilizing various AL query strategies as described in Section 4.1. AL labeling was done in iterations, with 60 samples selected per iteration across 100 iterations.

Data split. All labeled samples obtained through AL were used for training, while additional data, consisting of 250 instances per class, was reserved for test and development sets, as shown in Table 1.

Hardware/Software setup. All experiments were conducted using PyTorch on a single NVIDIA A100 40 GB GPU.

6. Experiment results

In this section we present the evaluation of our ALPET model alongside the CCW baseline model on three datasets CA-citation-needed, EU-citation-needed, and SQ-citation-needed. Results are organized as answers to hypotheses.

6.1. ALPET outperforms baseline models in low-resource languages (H1)

This section aims to evaluate Hypothesis *H1*. We hypothesized that ALPET would outperform the baseline CCW model in terms of data efficiency (i.e., achieving comparable performance with fewer labeled examples) and predictive performance (F1 Score) in low-resource languages, while utilizing the same AL query strategies.

Our experiments conducted on the CA-citation-needed, EU-citation-needed, and SQ-citation-needed datasets provide empirical evidence that supports this hypothesis. Fig. 3 offers a summarized view of the performance comparison between the ALPET model and the CCW baseline model. This figure presents the average F1 scores achieved by both models across all data selection strategies and instance counts for each of the three datasets. As figure indicates, ALPET, on average, achieves better performance than CCW baseline across all three datasets.

In the EU-citation-needed dataset as illustrated in the top subplot of Fig. 4(a), ALPET achieved an average F1 Score of 53% across all query strategies with only 50 labeled examples, whereas CCW required 150 labeled examples to reach the same F1 Score. This indicates that ALPET achieved comparable predictive performance with 66.67% reduction

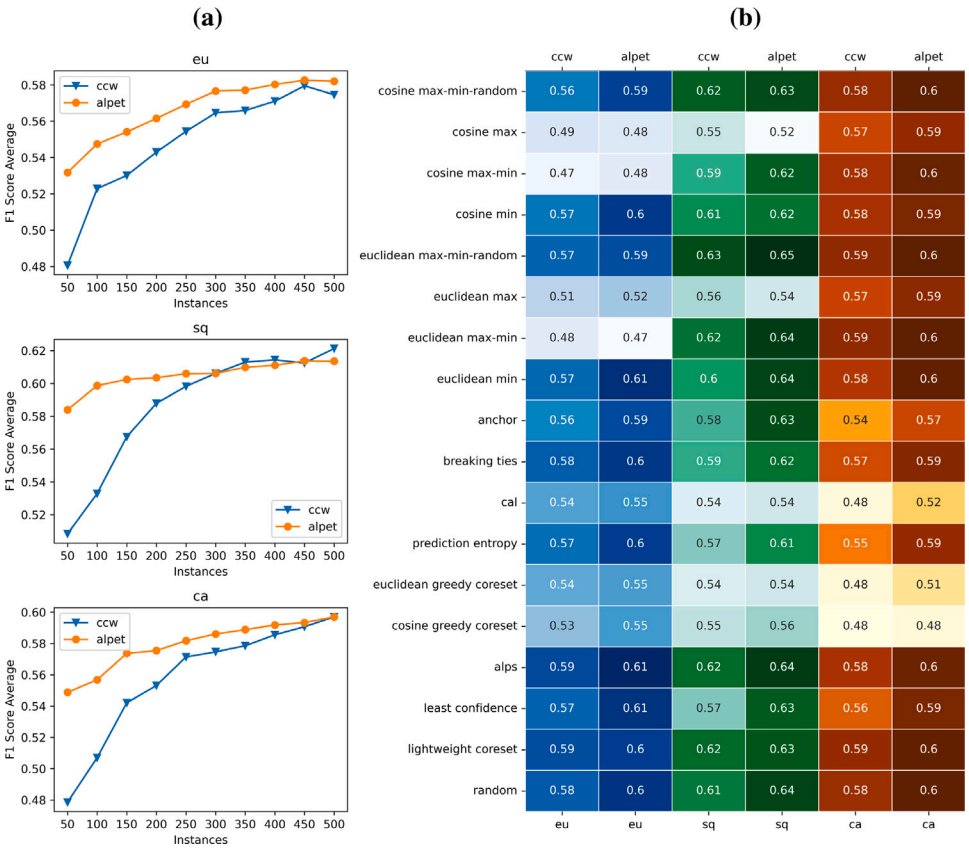


Fig. 4. Comparison between the **ALPET** model and the **CCW** baseline. Subplots in figure (a) represents the average F1 Score for various active learning query strategies across different number of instances for three datasets: eu, sq, and ca. The orange line with circle markers represents the alpet model, while the blue line with triangle markers represents the ccw baseline. Each subplot in (a) corresponds to a specific dataset, with eu at the top, sq in the middle, and ca at the bottom. In figure (b), the heatmap displays the average F1 Score across all instances for each query strategy. The distinct colors in the heatmap (blue, green, and orange) correspond to the eu, sq, and ca datasets, respectively, while the x-axis distinguishes between the ccw and alpet models for each dataset.

in labeled examples. In the SQ-citation-needed dataset, shown in the middle subplot of Fig. 4(a), ALPET achieved an average F1 Score of 58% across all query strategies with only 50 labeled examples, whereas CCW required 200 labeled examples to reach similar performance level. This represents a 75% reduction in labeled examples, reinforcing ALPET's superior data efficiency in this dataset.

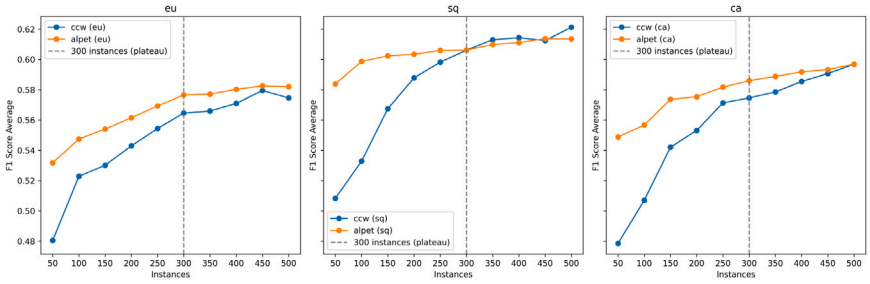
Similarly, in the CA-citation-needed dataset, as presented in the bottom subplot of Fig. 4(a), ALPET achieved an average F1 Score of 55% across all query strategies with only 50 labeled examples, whereas CCW needed above 150 labeled examples to match this performance. This demonstrates that ALPET not only outperforms CCW in terms of predictive performance but also does so with at least 66.67% fewer labeled examples. The reduced annotation cost achieved by ALPET, showed by the smaller number of labeled examples required, is particularly valuable in low-resource settings where labeled data is scarce and expensive.

The heatmap in Fig. 4(b) provides a detailed comparison of AL query strategies, showing the average F1 scores across all instance counts for both ALPET and CCW in the three datasets (eu, sq, and ca). The heatmap demonstrates that ALPET generally outperforms CCW in most query strategies across all datasets. For example, in the pool-bt strategy, ALPET achieves an average F1 Score of 60% in the eu dataset, compared to 58% for CCW, and similarly outperforms in the sq and ca datasets with an average F1 Score difference of 2%–3%. However, the models achieve comparable performance for pool-cal and pool-greedyc in the sq dataset, and pool-greedyc-cosine for ca dataset. ALPET underperforms CCW in strategies such as cosine-max in the eu and sq datasets, euclidean-max–min in the eu dataset and euclidean-max in the sq dataset. In conclusion we can see that the results presented in this section demonstrate ALPET's ability to achieve comparable or higher F1 scores with significantly fewer labeled examples than the CCW baseline, supporting Hypothesis *H1*. ALPET shows a clear advantage in both data efficiency and performance for CWD in low-resource languages.

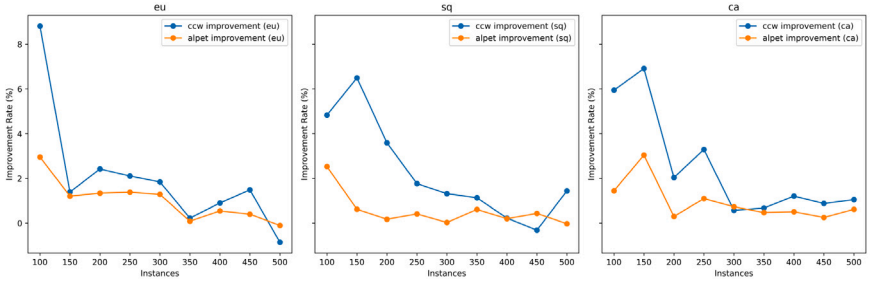
6.2. ALPET's performance plateau and data efficiency (H2)

This section aims to evaluate Hypothesis *H2*. We hypothesized that ALPET's performance improves with increasing labeled data but plateaus after a certain number of samples, making it effective in low-resource settings. The subplots in Fig. 5 confirm this hypothesis by showing that ALPET maintains a good performance up to 300 labeled examples, beyond which its improvement is minimal.

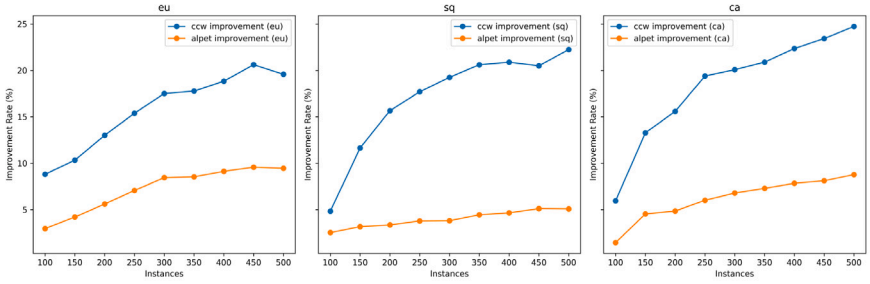
Looking closely at Fig. 5(a), at 50 shots, ALPET starts with a higher averaged F1 score across all three datasets (53%, 58%, and 55%) compared to CCW (48%, 51%, and 48%), showing a clear initial advantage in performance. As the number of labeled instances increases, ALPET continues to maintain its lead. The dashed vertical line marks the point where we hypothesized the performance plateau to occur (at 300 labeled examples per class). Specifically, in the EU-citation-needed dataset as shown in Fig. 5(a) (left plot), ALPET's performance rises from an F1 Score of 53% at 50 labeled examples to 58% at 300 labeled examples. Beyond 300 examples, the F1 Score stabilizes, showing minimal improvement with F1 score slightly passing 58% up to 500 instances. A similar trend is observed in the SQ-citation-needed dataset illustrated in Fig. 5(a) (middle plot), where ALPET's F1 score increases to 61% at 300 labeled examples, after which further improvement flattens. In the CA-citation-needed dataset Fig. 5(a) (right plot), ALPET reaches an F1 score of 59% at 300 labeled examples,



(a) F1 Score comparison between ALPET and CCW models across three datasets with plateau observed at 300 instances. Graphs show the average F1 score across all AL query strategies for ALPET and CCW models as the number of training instances increases.



(b) Incremental F1 Score improvement between successive batches for ALPET vs. CCW across three datasets.



(c) Cumulative F1 Score improvement for ALPET vs. CCW across three datasets with varying numbers of training instances.

Fig. 5. Comparison of F1 Score performance between ALPET and CCW models across three datasets (EU, SQ, CA). Subfigures show the average F1 scores, incremental improvements across batches, and cumulative F1 score improvements with increasing training instances.

beyond 300 instances and up to 500 the F1 Score improves to 60%. This trend is consistent across datasets, showing that ALPET is particularly efficient at leveraging smaller datasets in low-resource settings and that the extra effort or resources spent on labeling beyond this point does not lead to much further improvement. This is significant in low-resource settings, where obtaining large amounts of labeled data is often expensive and time-consuming. ALPET's ability to achieve good performance with only up to 300 labeled samples suggests its potential to overcome the data bottleneck and facilitate CWD in languages where labeled data is scarce.

A detailed comparison of performance gains in Fig. 5(b) shows the F1 score improvement between successive labeled example counts. For EU dataset Fig. 5(b) (left plot) ALPET improves almost 3% from 50–100 instances, and 1%–2% with each additional batch of labeled examples, but after 300 samples the gain diminishes to 0.1%. Similarly for CA datasets. An overall percentage improvement of instance count ranges is presented in the Table 3 which shows that from 300–500 instances the model improves about 1%–2% across languages.

The subplots in Fig. 5(c) show the cumulative improvement in F1 Score relative to the initial performance at 50 instances for both ALPET and CCW models across the three datasets. ALPET shows steady improvement, with a maximum of 9% cumulative gain by 500 instances; with the majority of gain achieved up to 300 instances as seen in the column **100–300** of Table 3. In contrast, CCW achieves 25% gain by 500 instances reinforcing the need for more data to achieve comparable results.

ALPET's performance has direct implications for resource efficiency because it reduces the annotation effort required compared to approaches like CCW. Furthermore, training on a smaller dataset reduces the computational time and resources required, making ALPET a more efficient approach for CWD in environments with constrained resources.

The analysis presented in this section, demonstrating a clear performance plateau after 300 labeled examples across all three datasets, supporting Hypothesis *H2*. ALPET's ability to achieve and maintain good performance with a limited number of samples underscores its suitability for low-resource CWD tasks.

6.3. Efficiency and robustness of ALPET with reduced labeled data in low-resource languages (H3)

This section aims to evaluate Hypothesis *H3*. We hypothesized that ALPET would achieve comparable performance to the baseline CCW model while requiring significantly fewer labeled examples. Specifically as presented in Fig. 5(a), ALPET achieves an average F1 Score of 55% across the three datasets with 50 labeled examples per class, whereas CCW requires 200 examples for the same performance. We also evaluate whether ALPET's performance remains robust with fewer labeled examples to demonstrate its data efficiency in low-resource

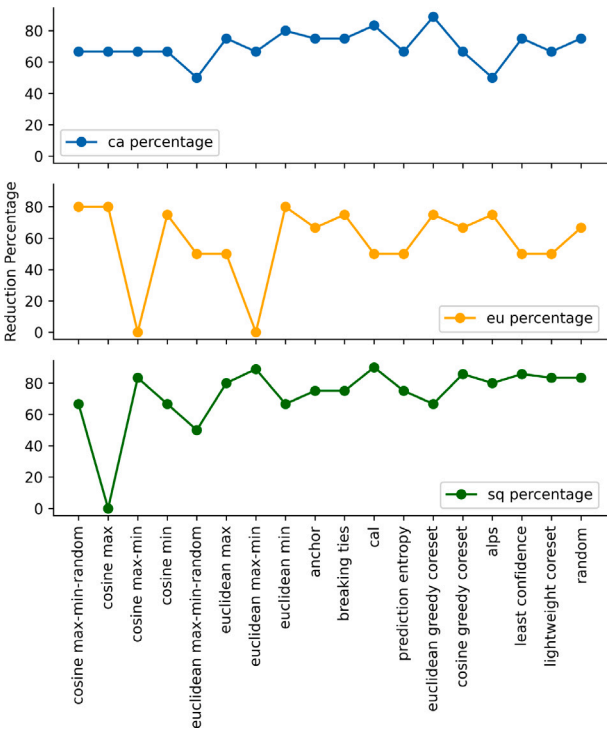


Fig. 6. Percentage reduction in labeled examples needed by ALPET compared to CCW across query strategies for CA, EU, and SQ datasets.

settings. The percentage reduction in labeled examples between ALPET and CCW for each query strategy was calculated by comparing the 50-shot performance of ALPET to the performance of CCW at the point where it matched ALPET's F1 Score. We consider an F1 Score of 55%, the average score achieved by ALPET with 50 labeled samples, as a competitive performance benchmark. A summary of reductions across three datasets is visualized in Fig. 6 and presented in Table 4.

The reduction percentages across the three datasets reveal that ALPET achieves an average reduction of 70% in CA, 58% in EU, and 72% in SQ (mean in Table 4) while maintaining a competitive F1 Score of 55% compared to CCW performance. In at least 50% of the query strategies, ALPET reduces the labeled data requirement by 67% in CA and EU, and by 78% in SQ (median in Table 4). In 25% of the query strategies, ALPET achieves a reduction of 75%–83% (75th percentile in Table 4). These results demonstrate that ALPET's reduction performance is consistently high across various query strategies. The variability (as indicated by std in Table 4) differs slightly across datasets. The EU dataset shows a standard deviation of 24%, suggesting higher variability in reduction percentages compared to SQ (21%) and CA (10%). This indicates that the reduction in labeled examples may depend more on the chosen query strategy for some languages than others.

Fig. 6 visualizes the percentage reduction in labeled examples needed by ALPET compared to CCW across various query strategies for the CA, EU, and SQ datasets. Overall, ALPET demonstrates strong performance, with most query strategies achieving substantial reductions in labeling requirements, ranging between 58% and 83% across datasets. While the CA dataset shows consistently high reductions, the EU and SQ datasets exhibit some variability, including a few query strategies that fail to reduce labeling at all. However, the majority of strategies still perform effectively, highlighting the robustness of ALPET's approach in minimizing labeled data across different languages.

The findings presented in this section, demonstrating ALPET's ability to achieve competitive performance (F1 Score of 55%) with above 80% reductions in labeled data strongly support Hypothesis *H3*. ALPET's efficiency and robustness in low-resource settings make it a promising approach for CWD in such contexts. From the results above we saw that ALPET is efficient in three low-resource language, this might suggest that the model's data efficiency can generalize to other languages with similar resource constraints, however, further research will need to be conducted to confirm this.

6.4. Effectiveness of active learning query strategies (H4)

Initially we hypothesized that AL query strategies typically achieve higher F1 scores than random sampling in low-resource language datasets. This expectation was based on the belief that effective AL query strategies, which select data strategically, would outperform the simple random sampling. However, the results obtained from the comparative analysis of the ALPET model across three low-resource languages using the F1 Score as the evaluation metric did not support this hypothesis across all AL strategies employed.

As described in the dataset section, we created ten datasets for each language, with instance counts ranging from 50 to 500. In each case, we compared random sampling with one of the AL query strategies, ensuring that both methods used the same number of shots. In Fig. 7 each dataset has its own plot (marked with ca, eu, sq at the top), whereas the fourth plot with notation (combined) holds the average performance of three datasets. The x-axes represent shot/instance sizes (50–500) while *y*-axis show AL strategies. To evaluate the performance, we calculated the difference in F1 scores between random sampling and AL strategies. Before generating the combined plot of Fig. 7 we found the average F1 Score of three datasets, then subtracted random sampling's F1 from that of the AL strategies. We define an AL query strategy as better than random sampling if the difference between their F1 Scores is a negative number. The results were visualized using heatmaps, where red shades represent better performance by AL strategies relative to random sampling, the blue shades represent better performance from random sampling, and finally the light blue almost white squares with zero values represent similar performance of both models (see Fig. 7).

The experiments conducted to evaluate this hypothesis revealed a more nuanced scenario than initially anticipated. While certain AL strategies did exhibit some advantages, the results only partially support H4. Strategies such as ALPS, LightweightCoreset, euclidean-min, and euclidean-cycle outperformed random sampling in certain instances, particularly when slightly larger shot sizes provided more samples to guide the clustering or distance-based selection. Their effectiveness however varied by language and shot size and often yielded only marginal improvements over random data selection. These more effective AL strategies, have in common usage of K-Means clustering which in essence works through euclidean distance calculation.

By contrast, AL strategies that failed to outperform random sampling, systematically across all shot sizes and languages, such as cosine-max, euclidean-max, GreedyCoreset, tend to select points based on their maximum distance from one another. In low-resource scenarios, this approach is struggling to identify sufficiently representative samples, making it no more effective (or sometimes less effective) than a random selection. Another key factor is the size of the unlabeled data pool from which AL strategies draw. Prior studies (Sorscher, Geirhos, Shekhar, Ganguli, & Morcos, 2022) have shown that advanced samplers begin to outperform random sampling when a dataset contains around one million rows and approximately 1% of data are selected form there — a condition satisfied by Catalan dataset only. In contrast, for smaller datasets (e.g., around 500 selected samples), random sampling tends to be just as effective, or even better. In conclusion, the hypothesis that AL strategies consistently outperform random sampling in low-resource language datasets for CWD was not fully supported. The effectiveness of specific AL strategies was contingent upon factors such as the dataset, language, and the number of labeled instances. As a result, random sampling remains a competitive baseline, particularly when dealing with smaller pool of unlabeled data.

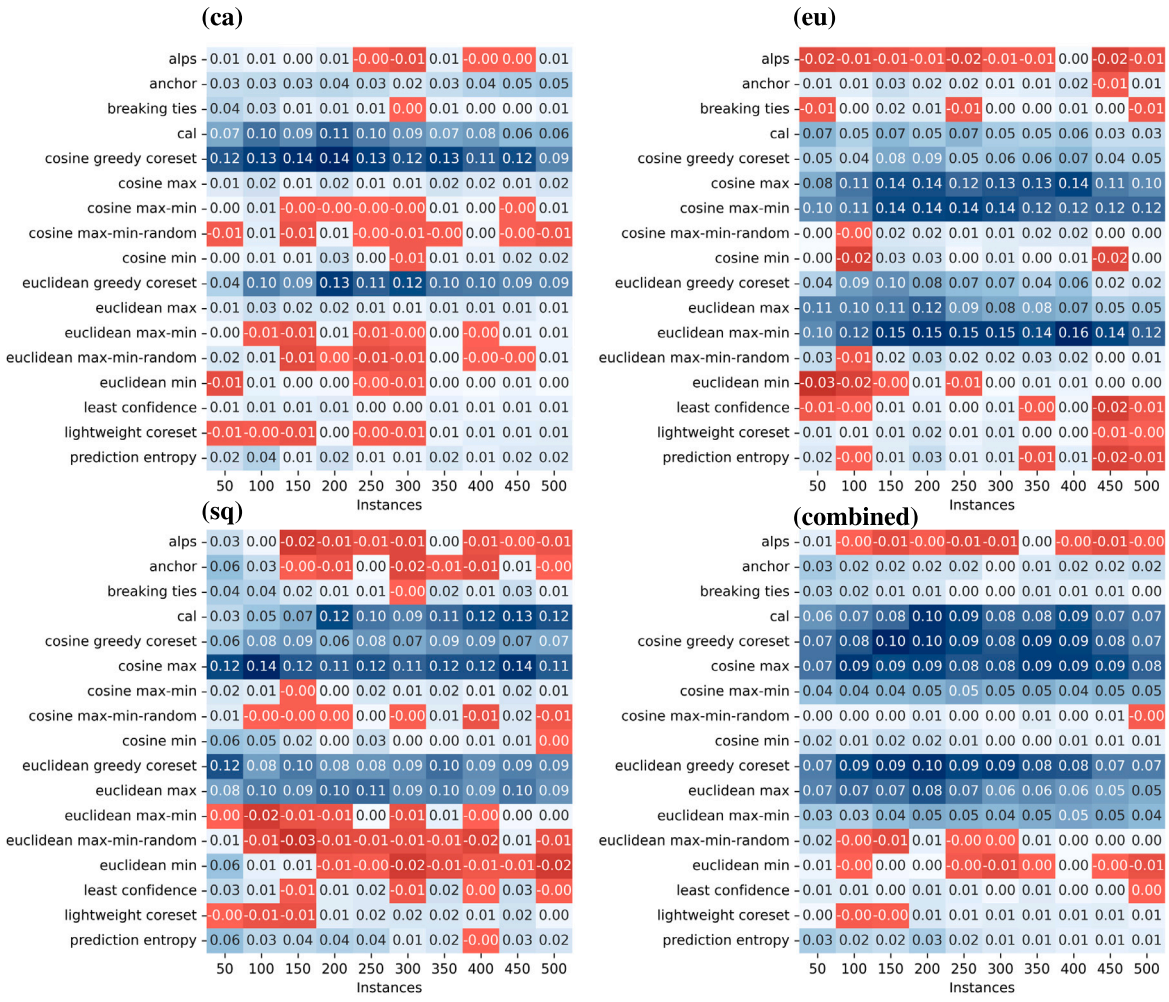


Fig. 7. Comparing AL strategies to random data sampling. Heatmap displaying the cross-differences in F1 scores between active learning (AL) strategies and random sampling across all datasets and shot sizes.

6.4.1. Linguistic features and AL performance

The red squares in the heatmaps of Fig. 7 show that performance of AL query strategies compared to random sampling varies across languages, which may be due to factors like language complexity. This variability signifies that further investigation is needed to understand the reason behind it.

In the ca and eu datasets, usually AL strategies outperform random sampling when the labeled data size is smaller (up to 250 shots) as opposed to when the instances increase up to 500. This may be attributed to the larger original unlabeled data pools in these two datasets, which give AL strategies access to a more diverse and varied set of examples to choose from. The higher variance in the data pool allows AL strategies to better identify informative samples, which leads to improved performance in low-shot scenarios. On the other hand, the sq dataset, with its smaller unlabeled data pool, shows more limited gains from AL strategies early on compared to when more labeled data instances are added. This can be an indication that AL's advantage is more pronounced when there is a larger pool of unlabeled data to choose from.

Ultimately, we were interested to find out in which cases AL query strategies are more consistent in beating random data selection across three languages. We found out that only in a limited number of cases this happened, ALPS with 250 and 300 instances and euclidean-min with 250 instances. Showing that the effectiveness of these strategies is more conditional than initially hypothesized. ALPS performance can be attributed to the strategy's ability to group semantically similar sentences based on dense embeddings from BERT, which encapsulates rich semantic information. By clustering these semantically dense sentences, K-Means increases the likelihood of selecting informative examples. In contrast, random sampling may pick redundant or less informative samples, making BERT embeddings and K-Means especially effective in low-resource settings.

To gain deeper insights into the linguistic features that influence the performance of AL strategies, we compared K-Means and random sampling across six linguistic metrics: Unique Word Count, Type-Token Ratio (TTR), Vocabulary Richness, Total Tokens, Average Words Per Sentence, and Average Word Length, visualized in Fig. 8. The reduced TTR for K-Means suggest a trade-off between sentence length and lexical diversity because longer sentences often contain repeated words, which lower the TTR. However, the higher Unique Word Count and Vocabulary Richness indicate that the K-Means captures diverse vocabulary. By selecting longer sentences as presented by Average Words Per Sentence, K-Means samples provide more complex language structures. While TTR may be lower, the presence of more tokens suggests that the sentences are capable of conveying more information. K-Means, when paired with BERT embeddings, effectively captures semantically similar sentences. Average Word Length provides additional insight into linguistic complexity, particularly in Basque (eu), which, as an agglutinative language, naturally has longer words due to its morphological structure. K-Means' ability to work well with longer word forms in Basque suggests that it can effectively capture morphologically complex languages. In contrast, Albanian (sq) and Catalan (ca), which have shorter average word lengths, benefit more from K-Means' ability

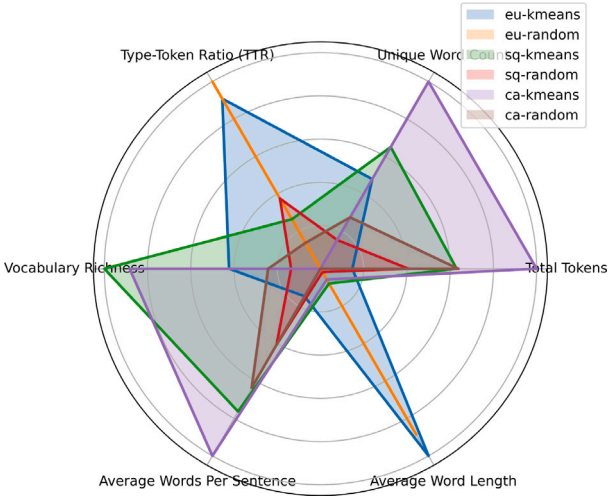


Fig. 8. Linguistic features that could influence the performance of data selection through ALPS versus random selection.

to select longer, more informative sentences rather than focusing on individual word length. In conclusion, K-Means performance across three languages indicates that this strategy as opposed to random has helped the model to better understand the language nuances that likely contribute to improved model training. This is crucial in low-resource contexts, where the richness of the training data influences the performance of models.

While not uniformly outperforming random sampling, analysis in this section reveal that specific AL query strategies, particularly those employing semantic clustering techniques like K-Means, can offer advantages in low-resource language settings. These advantages become more pronounced with larger unlabeled pool of data, where AL strategies have a greater opportunity to identify and leverage informative samples allowing random sampling to remain a competitive baseline for smaller datasets. These findings highlight the importance of carefully considering both the AL strategy and dataset characteristics when working with low-resource languages.

7. Conclusion

We introduced ALPET, an active few-shot learning approach integrating Active Learning (AL) and Pattern-Exploiting Training (PET) for Citation Worthiness Detection (CWD) in low-resource languages. Evaluating ALPET on Catalan, Basque, and Albanian confirmed its superior data efficiency and accuracy over the CCW baseline.

Our findings show that ALPET achieves the same or better performance than CCW while requiring significantly fewer labeled examples. With just 50 labeled examples per class, ALPET attained an average F1 score of 55%, whereas CCW needed 200 examples for the same performance. Across datasets, ALPET reduced annotation needs by 58%–78%, making it a cost-effective solution for low-resource settings. Moreover, its performance plateaus after 300 labeled samples, indicating its ability to achieve optimal performance with minimal data, a crucial advantage for under-resourced languages.

A key factor behind ALPET's superior performance is PET's ability to generalize well with minimal labeled data. Active Learning further enhances this by selecting the most informative samples, refining model learning. However, our analysis also revealed that the effectiveness of AL strategies varies by dataset and language, with some strategies (e.g., those incorporating K-Means clustering in the backend) performing better when applied to large unlabeled pools. This suggests that AL strategy selection should be tailored to the specific task and dataset characteristics.

Beyond CWD, ALPET's data efficiency and robustness make it a promising approach for other NLP tasks facing similar constraints, such as claim detection and rumor verification. This research contributes to the development of more effective and scalable CWD systems for under-resourced languages, enhancing the reliability of information in these languages. Future work could explore the impact of real-time human annotation in Active Learning settings, further strengthening the applicability of this approach in real-world scenarios.

8. Limitation and future work

A limitation of this study is the reliance on simulated AL rather than involving real-time human annotators. While AL is designed to iteratively select the most informative examples for human labeling to outperform random sampling, conducting such experiments with live human feedback is time-consuming and costly. This makes it challenging to implement in academic research. To address this, we simulated the process by using already labeled datasets, treating them as if they were unlabeled. Although this approach is widely adopted to bypass the logistical difficulties of real-time annotation, it may not fully replicate the dynamic interactions found in practical, human-in-the-loop AL environments. Human annotators bring their own subjective interpretations, biases, and inconsistencies to the labeling process, factors that a simulated environment cannot fully replicate.

Additionally, in this study, we focused on a specific set of AL query strategies. While this provides valuable insights, it does not cover the entire spectrum of available techniques. Exploring a wider range of AL methods, such as committee-based strategies, could provide a more comprehensive understanding of AL's potential and limitations for CWD in low-resource languages.

Furthermore, while ALPET has been evaluated on three distinct low-resource languages, its effectiveness in extremely low-resource languages, where labeled data is even scarcer and linguistic structures may be more complex, remains an open question. Certain languages with rich morphological structures or limited written corpora could present additional challenges that require further investigation.

Real-world deployment scenarios and practical applications of ALPET remain unexplored. Evaluating its feasibility in citation recommendation systems or automated fact-checking pipelines would provide insights into its practical utility. Moreover, conducting user studies to assess how human annotators interact with ALPET's outputs could help validate its practical applicability. Finally, benchmarking ALPET against multilingual CWD models would offer a broader perspective on its competitiveness across a wider range of languages.

Future research could prioritize addressing these limitations by:

- Incorporating small-scale experiments with human annotators to validate the findings derived from the simulated AL setting. This would provide valuable insights into the real-world performance of ALPET and help refine the data selection process.
- Exploring and evaluating a more diverse set of AL strategies, focusing on their robustness and adaptability to different languages and data characteristics. This could involve investigating methods that specifically address the challenges posed by imbalanced datasets and noisy labels, which are common in low-resource scenarios.
- Investigate ALPET's effectiveness in languages with highly complex morphology or extremely limited annotated data to assess its adaptability across a wider range of low-resource settings.
- Investigating real-world deployment scenarios to assess ALPET's practicality in citation recommendation systems or automated fact-checking pipelines.
- Conducting user studies to evaluate ALPET's applicability, measuring how well human annotators interact with the system and whether its recommendations align with expert judgments.
- Expanding benchmarking efforts by comparing ALPET against broader multilingual CWD models to assess its competitiveness across a wider set of languages.

CRediT authorship contribution statement

Aida Halitaj: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing — original draft, Writing — review & editing, Visualization, Project administration. **Arkaitz Zubiaga:** Conceptualization, Supervision, Writing — review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

Acharya, A., Mooney, R. J., & Ghosh, J. (2014). Active multitask learning using both latent and supervised shared topics. In *Proceedings of the 2014 SIAM international conference on data mining* (pp. 190–198). SIAM.

Alagić, D., & Šnajder, J. (2015). Experiments on active learning for Croatian word sense disambiguation. In *The 5th workshop on balto-slavic natural language processing* (pp. 49–58).

Ananthakrishnan, S., Prasad, R., Stallard, D., & Natarajan, P. (2010). A semi-supervised batch-mode active learning strategy for improved statistical machine translation. In *Proceedings of the fourteenth conference on computational natural language learning* (pp. 126–134).

Angluin, D. (1988). Queries and concept learning. *Machine Learning*, *2*, 319–342.

Ash, J. T., Zhang, C., Krishnamurthy, A., Langford, J., & Agarwal, A. (2019). Deep batch active learning by diverse, uncertain gradient lower bounds. arXiv preprint arXiv:1906.03671.

Bachem, O., Lucic, M., & Krause, A. (2018). Scalable k-means clustering via lightweight coresets. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 1119–1127).

Baigutanova, A., Myung, J., Saez-Trumper, D., Chou, A.-J., Redi, M., Jung, C., et al. (2023). Longitudinal assessment of reference quality on wikipedia. In *Proceedings of the ACM web conference 2023* (pp. 2831–2839).

Barnabò, G., Siciliano, F., Castillo, C., Leonardi, S., Nakov, P., Da San Martino, G., et al. (2023). Deep active learning for misinformation detection using geometric deep learning. *Online Social Networks and Media*, *33*, Article 100244.

Baum, E. B., & Lang, K. (1992). Query learning can work poorly when a human oracle is used. *vol. 8*, In *International joint conference on neural networks* (p. 8). Beijing China.

Bayer, M., & Reuter, C. (2024). ActiveLLM: Large language model-based active learning for textual few-shot scenarios. arXiv preprint arXiv:2405.10808.

Beatty, G., Kochis, E., & Bloodgood, M. (2018). Impact of batch size on stopping active learning for text classification. In *2018 IEEE 12th international conference on semantic computing* (pp. 306–307). IEEE.

Bonab, H., Zamani, H., Learned-Miller, E., & Allan, J. (2018). Citation worthiness of sentences in scientific reports. In *The 41st international ACM SIGIR conference on research & development in information retrieval* (pp. 1061–1064).

Brantley, K., Sharaf, A., & Daumé III, H. (2020). Active imitation learning with noisy guidance. arXiv preprint arXiv:2005.12801.

Brinker, K. (2003). Incorporating diversity in active learning with support vector machines. In *Proceedings of the 20th international conference on machine learning* (pp. 59–66).

Cacciarelli, D., & Kulahci, M. (2024). Active learning for data streams: a survey. *Machine Learning*, *113*(1), 185–239.

Chimoto, E. A., & Bassett, B. A. (2022). COMET-QE and active learning for low-resource machine translation. arXiv preprint arXiv:2210.15696.

Citovsky, G., DeSalvo, G., Gentile, C., Karydas, L., Rajagopalan, A., Rostamizadeh, A., et al. (2021). Batch active learning at scale. *Advances in Neural Information Processing Systems*, *34*, 11933–11944.

Cohn, D., Atlas, L., & Ladner, R. (1994). Improving generalization with active learning. *Machine Learning*, *15*, 201–221.

Dagan, I., & Engelson, S. P. (1995). Committee-based sampling for training probabilistic classifiers. In *Machine learning proceedings 1995* (pp. 150–157). Elsevier.

Devlin, J. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Dligach, D., & Palmer, M. (2011). Good seed makes a good crop: accelerating active learning using language modeling. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies* (pp. 6–10).

Dor, L. E., Halfon, A., Gera, A., Shnarch, E., Dankin, L., Choshen, L., et al. (2020). Active learning for BERT: an empirical study. In *Proceedings of the 2020 conference on empirical methods in natural language processing* (pp. 7949–7962).

Dossou, B. F., Tonja, A. L., Yousuf, O., Osei, S., Oppong, A., Shode, I., et al. (2022). AfroLM: A self-active learning-based multilingual pretrained language model for 23 African languages. arXiv preprint arXiv:2211.03263.

Farinneya, P., Pour, M. M. A., Hamidian, S., & Diab, M. (2021). Active learning for rumor identification on social media. In *Findings of the association for computational linguistics: EMNLP 2021* (pp. 4556–4565).

Gal, Y., & Ghahramani, Z. (2016). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *International conference on machine learning* (pp. 1050–1059). PMLR.

Grießhaber, D., Maucher, J., & Vu, N. T. (2020). Fine-tuning BERT for low-resource natural language understanding via active learning. arXiv preprint arXiv:2012.02462.

Guo, Z., Schlichtkrull, M., & Vlachos, A. (2022). A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, *10*, 178–206.

Gupta, A., & Srikumar, V. (2021). X-fact: A new benchmark dataset for multilingual fact checking. arXiv preprint arXiv:2106.09248.

Halitaj, A., & Zubiaga, A. (2024). Providing citations to support fact-checking: Contextualizing detection of sentences needing citation on small wikipedias. *Natural Language Processing Journal*, Article 100093.

Harpale, A., & Yang, Y. (2010). Active learning for multi-task adaptive filtering.

Hoi, S. C., Jin, R., & Lyu, M. R. (2006). Large-scale text categorization by batch mode active learning. In *Proceedings of the 15th international conference on world wide web* (pp. 633–642).

Holub, A., Perona, P., & Burl, M. C. (2008). Entropy-based active learning for object recognition. In *2008 IEEE computer society conference on computer vision and pattern recognition workshops* (pp. 1–8). IEEE.

Ikhwantri, F., Louvan, S., Kurniawan, K., Abisena, B., Rachman, V., Wicaksono, A. F., et al. (2018). Multi-task active learning for neural semantic role labeling on low resource conversational corpus. arXiv preprint arXiv:1806.01523.

Imamura, M., Takayama, Y., Kaji, N., Toyoda, M., & Kitsuregawa, M. (2009). A combination of active learning and semi-supervised learning starting with positive and unlabeled examples for word sense disambiguation: an empirical study on Japanese web search query. In *Proceedings of the ACL-IJCNLP 2009 conference short papers* (pp. 61–64).

Kasai, J., Qian, K., Gurajada, S., Li, Y., & Popa, L. (2019). Low-resource deep entity resolution with transfer and active learning. arXiv preprint arXiv:1906.08042.

King, R. D., Rowland, J., Oliver, S. G., Young, M., Aubrey, W., Byrne, E., et al. (2009). The automation of science. *Science*, *324*(5923), 85–89.

King, R. D., Whelan, K. E., Jones, F. M., Reiser, P. G., Bryant, C. H., Muggleton, S. H., et al. (2004). Functional genomic hypothesis generation and experimentation by a robot scientist. *Nature*, *427*(6971), 247–252.

Konstantinovskiy, L., Price, O., Babakar, M., & Zubiaga, A. (2021). Toward automated factchecking: Developing an annotation schema and benchmark for consistent automated claim detection. *Digital Threats: Research and Practice*, *2*(2), 1–16.

Kranjc, J., Smailović, J., Podpečan, V., Grčar, M., Žnidaršič, M., & Lavrač, N. (2015). Active learning for sentiment analysis on data streams: Methodology and workflow implementation in the ClowdFlows platform. *Information Processing & Management*, *51*(2), 187–203.

Le, H. T., To, L. T., Nguyen, M. T., & Van Nguyen, K. (2024). ViWikiFC: Fact-checking for Vietnamese wikipedia-based textual knowledge source. arXiv preprint arXiv:2405.07615.

Lesci, P., & Vlachos, A. (2024). AnchorAL: Computationally efficient active learning for large and imbalanced datasets. arXiv preprint arXiv:2404.05623.

Lewis, D. D. (1995). A sequential algorithm for training text classifiers: Corrigendum and additional data. *vol. 29*, In *Acm sigir forum* (pp. 13–19). ACM New York, NY, USA.

Lewis, D. D., & Gale, W. A. (1994). A sequential algorithm for training text classifiers. URL https://api.semanticscholar.org/CorpusID:260481767.

Li, D., Wang, Z., Chen, Y., Jiang, R., Ding, W., & Okumura, M. (2024). A survey on deep active learning: Recent advances and new frontiers. *IEEE Transactions on Neural Networks and Learning Systems*.

Lu, Y., Yao, B., Zhang, S., Wang, Y., Zhang, P., Lu, T., et al. (2023). Human still wins over llm: An empirical study of active learning on domain-specific annotation tasks. arXiv preprint arXiv:2311.09825.

Luo, T., Kramer, K., Goldgof, D. B., Hall, L. O., Samson, S., Remsen, A., et al. (2005). Active learning to recognize multiple types of plankton. *Journal of Machine Learning Research*, *6*(4).

Maekawa, S., Zhang, D., Kim, H., Rahman, S., & Hruschka, E. (2022). Low-resource interactive active labeling for fine-tuning language models. In *Findings of the association for computational linguistics: EMNLP 2022* (pp. 3230–3242).

Mamooler, S., Lebret, R., Massonnet, S., & Aberer, K. (2022). An efficient active learning pipeline for legal text classification. arXiv preprint arXiv:2211.08112.

Margatina, K., Vernikos, G., Barrault, L., & Aletras, N. (2021). Active learning by acquiring contrastive examples. arXiv preprint arXiv:2109.03764.

Mendonça, V., Sardinha, A., Coheur, L., & Santos, A. L. (2020). Query strategies, assemble! active learning with expert advice for low-resource natural language processing. In *2020 IEEE international conference on fuzzy systems (FUZZ-iEEE)* (pp. 1–8). IEEE.

Müller, T., Pérez-Torró, G., Basile, A., & Franco-Salvador, M. (2022). Active few-shot learning with fasl. In *International conference on applications of natural language to information systems* (pp. 98–110). Springer.

Öhman, J. (2021). Active learning for named entity recognition with swedish language models.

Qian, K., Raman, P. C., Li, Y., & Popa, L. (2020). Learning structured representations of entity names using active learning and weak supervision. arXiv preprint arXiv:2011.00105.

Qian, K., Sang, Y., Bayat, F. F., Belyi, A., Chu, X., Govind, Y., et al. (2024). APE: Active learning-based tooling for finding informative few-shot examples for LLM-based entity matching. arXiv preprint arXiv:2408.04637.

Quelle, D., Cheng, C., Bovet, A., & Hale, S. A. (2023). Lost in translation–multilingual misinformation and its evolution. arXiv preprint arXiv:2310.18089.

Quteineh, H., Samothrakis, S., & Sutcliffe, R. (2020). Textual data augmentation for efficient active learning on tiny datasets. In *Proceedings of the 2020 conference on empirical methods in natural language processing* (pp. 7400–7410).

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al. (2018). Improving language understanding by generative pre-training.

Radmard, P., Fathullah, Y., & Lipani, A. (2021). Subsequence based deep active learning for named entity recognition. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: long papers)* (pp. 4310–4321).

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., et al. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, *21*(140), 1–67.

Redi, M., Fetahu, B., Morgan, J., & Taraborelli, D. (2019). Citation needed: A taxonomy and algorithmic assessment of wikipedia's verifiability. In *The world wide web conference* (pp. 1567–1578).

Reichart, R., Tomanek, K., Hahn, U., & Rappoport, A. (2008). Multi-task active learning for linguistic annotations. In *Proceedings of ACL-08: HLT* (pp. 861–869).

Ringger, E., McClanahan, P., Haertel, R., Busby, G., Carmen, M., Carroll, J., et al. (2007). Active learning for part-of-speech tagging: Accelerating corpus annotation. In *Proceedings of the linguistic annotation workshop* (pp. 101–108).

Rotman, G., & Reichart, R. (2022). Multi-task active learning for pre-trained transformer-based models. *Transactions of the Association for Computational Linguistics*, *10*, 1209–1228.

Sathe, A., Ather, S., Le, T. M., Perry, N., & Park, J. (2020). Automated fact-checking of claims from wikipedia. In *Proceedings of the twelfth language resources and evaluation conference* (pp. 6874–6882).

Schick, T., & Schütze, H. (2020). Exploiting cloze questions for few shot text classification and natural language inference. arXiv preprint arXiv:2001.07676.

Schröder, C., Müller, L., Niekler, A., & Potthast, M. (2023). Small-text: Active learning for text classification in python. In *Proceedings of the 17th conference of the European chapter of the association for computational linguistics: system demonstrations* (pp. 84–95). Dubrovnik, Croatia: Association for Computational Linguistics, URL https://aclanthology.org/2023.eacl-demo.11.

Schumann, R., & Rehbein, I. (2019). Active learning via membership query synthesis for semi-supervised sentence classification. In *Proceedings of the 23rd conference on computational natural language learning* (pp. 472–481).

Sener, O., & Savarese, S. (2017). Active learning for convolutional neural networks: A core-set approach. arXiv preprint arXiv:1708.00489.

Settles, B. (2009). Active learning literature survey.

Settles, B. (2012). *Synthesis lectures on artificial intelligence and machine learning*. Morgan & Claypool Publishers.

Shahi, G. K., & Nandini, D. (2020). FakeCovid–a multilingual cross-domain fact check news dataset for COVID-19. arXiv preprint arXiv:2006.11343.

Shen, D., Zhang, J., Su, J., Zhou, G., & Tan, C. L. (2004). Multi-criteria-based active learning for named entity recognition. In *Proceedings of the 42nd annual meeting of the association for computational linguistics (ACL-04)* (pp. 589–596).

Shi, T., Benton, A., Malioutov, I., & Irsoy, O. (2021). Diversity-aware batch active learning for dependency parsing. arXiv preprint arXiv:2104.13936.

Siddhant, A., & Lipton, Z. C. (2018). Deep bayesian active learning for natural language processing: Results of a large-scale empirical study. arXiv preprint arXiv:1808.05697.

Smailović, J., Grčar, M., Lavrač, N., & Žnidaršič, M. (2014). Stream-based active learning for sentiment analysis in the financial domain. *Information Sciences*, *285*, 181–203.

Sorscher, B., Geirhos, R. ., Shekhar, S., Ganguli, S., & Morcos, A. (2022). Beyond neural scaling laws: beating power law scaling via data pruning. *Advances in Neural Information Processing Systems*, *35*, 19523–19536.

Stratos, K., & Collins, M. (2015). Simple semi-supervised POS tagging. In *Proceedings of the 1st workshop on vector space modeling for natural language processing* (pp. 79–87).

Thorne, J., & Vlachos, A. (2018). Automated fact checking: Task formulations, methods and future directions. arXiv preprint arXiv:1806.07687.

Tong, S., & Koller, D. (2001). Support vector machine active learning with applications to text classification. *Journal of Machine Learning Research*, *2*(Nov), 45–66.

Tsvigun, A., Lysenko, I., Sedashov, D., Lazichny, I., Damirov, E., Karlov, V., et al. (2023). Active learning for abstractive text summarization. arXiv preprint arXiv:2301.03252.

Tsvigun, A., Shelmanov, A., Kuzmin, G., Sanochkin, L., Larionov, D., Gusev, G., et al. (2022). Towards computationally feasible deep active learning. arXiv preprint arXiv:2205.03598.

Tur, G., Hakkani-Tür, D., & Schapire, R. E. (2005). Combining active and semi-supervised learning for spoken language understanding. *Speech Communication*, *45*(2), 171–186.

Van Tran, C., Nguyen, T. T., Hoang, D. T., Hwang, D., & Nguyen, N. T. (2017). Active learning-based approach for named entity recognition on short text streams. In *Multimedia and network information systems: proceedings of the 10th international conference MISSI 2016* (pp. 321–330). Springer.

Wang, Y., Yao, Q., Kwok, J. T., & Ni, L. M. (2020). Generalizing from a few examples: A survey on few-shot learning. *ACM Computing Surveys (Csur)*, *53*(3), 1–34.

Wright, D., & Augenstein, I. (2020). Claim check-worthiness detection as positive unlabelled learning. arXiv preprint arXiv:2003.02736.

Yao, B., Jindal, I., Popa, L., Katsis, Y., Ghosh, S., He, L., et al. (2023). *Beyond labels: Empowering human with natural language explanations through a novel active-learning architecture*. Association for Computational Linguistics.

Ye, Z., Liu, D., Pavani, K., & Dasgupta, S. (2023). LAMM: Language aware active learning for multilingual models. In *Proceedings of the 32nd ACM international conference on information and knowledge management* (pp. 5255–5256).

Yuan, M., Lin, H.-T., & Boyd-Graber, J. (2020). Cold-start active learning through self-supervised language modeling. arXiv preprint arXiv:2010.09535.

Zeng, X. (2024). *Few-shot claim verification for automated fact checking* (Ph.D. thesis), Queen Mary University of London.

Zeng, X., Abumansour, A. S., & Zubiaga, A. (2021). Automated fact-checking: A survey. *Language and Linguistics Compass*, *15*(10), Article e12438.

Zeng, X., & Zubiaga, A. (2022). Active PETs: active data annotation prioritisation for few-shot claim verification with pattern exploiting training. arXiv preprint arXiv:2208.08749.

Zhang, Y., Feng, S., & Tan, C. (2022). Active example selection for in-context learning. arXiv preprint arXiv:2211.04486.

Zhang, Y., Lease, M., & Wallace, B. (2017). Active discriminative text representation learning. *vol. 31*, In *Proceedings of the AAAI conference on artificial intelligence*.

Zhang, Z., Strubell, E., & Hovy, E. (2022). A survey of active learning for natural language processing. arXiv preprint arXiv:2210.10109.

Zhang, R., Yu, Y., Shetty, P., Song, L., & Zhang, C. (2022). Prboost: Prompt-based rule discovery and boosting for interactive weakly-supervised learning. arXiv preprint arXiv:2203.09735.

Zhao, Y., Zhang, R. H., Zhou, S., & Zhang, Z. (2020). Active learning approaches to enhancing neural machine translation. In *Findings of the association for computational linguistics: EMNLP 2020* (pp. 1796–1806).

Zhou, B., Cai, X., Zhang, Y., Guo, W., & Yuan, X. (2021). MTAAL: multi-task adversarial active learning for medical named entity recognition and normalization. *vol. 35*, In *Proceedings of the AAAI conference on artificial intelligence* (pp. 14586–14593).

Zhou, Z., & Waibel, A. (2021). Active learning for massively parallel translation of constrained text into low resource languages. arXiv preprint arXiv:2108.07127.

Zhu, J., Wang, H., Yao, T., & Tsou, B. K. (2008). Active learning with sampling by uncertainty and density for word sense disambiguation and text classification. In *22nd international conference on computational linguistics, coling 2008* (pp. 1137–1144).

Zhu, Z. L., Yadav, V., Afzal, Z., & Tsatsaronis, G. (2022). Few-shot initializing of active learner via meta-learning. In *Findings of the association for computational linguistics: EMNLP 2022* (pp. 1117–1133).

Zhu, H., Ye, W., Luo, S., & Zhang, X. (2020). A multitask active learning framework for natural language understanding. In *Proceedings of the 28th international conference on computational linguistics* (pp. 4900–4914).