Advancing Active Learning with Ensemble Strategies

Naif Alatrush[1], **Sultan Alsarra**[2(✉)], **Afraa Alshammari**[1], **Luay Abdeljaber**[1],
Niamat Zawad[1], **Latifur Khan**[1], **Patrick T. Brandt**[1], **Javier Osorio**[3], **Vito J. D'Orazio**[4]
[1]University of Texas at Dallas, [2]King Saud University, [3]University of Arizona, [4]West Virginia University
{naif.alatrash,afraa.alshammari,luay.abdeljaber,
niamat.zawad,lkhan,pbrandt}@utdallas.edu,
josorio1@arizona.edu, vito.dorazio@mail.wvu.edu
✉ Corresponding Author: salsarra@ksu.edu.sa

Abstract

Active learning (AL) reduces annotation costs by selecting the most informative samples for labeling. However, traditional AL methods rely on a single heuristic, limiting data exploration and annotation efficiency. This paper introduces two ensemble-based AL methods: **Ensemble Union**, which combines multiple heuristics to improve dataset exploration, and **Ensemble Intersection**, which applies majority voting for robust sample selection. We evaluate these approaches on the United Nations Parallel Corpus in both English and Spanish using domain-specific models such as ConfliBERT. Our results show that ensemble-based AL strategies outperform individual heuristics, achieving classification performance comparable to full dataset training while using significantly fewer labeled examples. Although focused on political texts, the proposed methods are applicable to broader NLP annotation tasks where labeling costs are high.

1 Introduction

Efficient and high-quality annotation of political texts is essential for advancing natural language processing (NLP) applications in areas such as conflict analysis, electoral preferences, political speech, election prediction, misinformation and bias detection, and policy research (Linegar et al., 2023). The manual labeling of political documents requires expert domain knowledge and remains time-consuming, costly, and often lacks replicability and explainability (Baumgartner et al., 1998). While some researchers advocate the creation of synthetic data (Halterman, 2023) and the use of generative artificial intelligence for NLP tasks (Heseltine and Clemm von Hohenberg, 2024) as cost-effective ways to leverage human annotations, such approaches raise questions about data validity and ethics that are highly sensitive and consequential for political science research. Fortunately, Active Learning is a promising solution to select the most informative samples for labeling, mitigating annotation costs (Settles, 2009). Despite its success across various domains, traditional AL methods typically rely on a single acquisition heuristic, which can limit their effectiveness in exploring complex datasets such as political texts.

While active learning demonstrates strong performance in reducing labeling efforts (Rahman et al., 2024; Abdeljaber et al., 2025), its reliance on a single acquisition heuristic can lead to suboptimal exploration of the data space. Different heuristics prioritize distinct aspects of uncertainty, representativeness, or diversity. This indicates that no single method is universally optimal across datasets or tasks (Beluch et al., 2018; Gal et al., 2017). In politically oriented corpora, where domain-specific terms and linguistic expressions can vary widely depending on the context, ideology, and regional factors, this limitation is even more pronounced. There then is a growing need for active learning methods capable of leveraging few examples of human judgment to better balance exploration to improve annotation efficiency and model performance in political text classification.

To address these limitations, two novel ensemble-based active learning methods designed to combine the strengths of multiple acquisition heuristics are introduced. The first method, *Ensemble Union*, aggregates the top-ranked samples from several individual strategies to enhance dataset exploration within each annotation round. The second method, *Ensemble Intersection*, applies a majority voting scheme to identify the most consistently informative samples across different acquisition functions. Using complementary perspectives of multiple heuristics, these ensemble approaches improve annotation efficiency and model performance. The proposed approaches offer researchers increased

Proceedings of Recent Advances in Natural Language Processing,pages 57–66
Varna, Sep 8–10, 2025
https://doi.org/10.26615/978-954-452-098-4-007

leverage of a few human annotations over single acquisition heuristic, which is advantageous in complex domains like political text classification.

The paper is organized as follows. Section 2 reviews related work on active learning, ensemble methods, and political text classification. Section 3 describes the datasets used in our experiments. Section 4 details the active learning approaches, including our proposed ensemble methods. Section 5 outlines the experimental setup, and Section 6 shows the results and analysis. Finally, Section 7 concludes the paper with directions for future work.

2 Related Work

Active learning (AL) has been extensively studied as a strategy to reduce annotation costs by selecting the most informative examples for labeling. Traditional AL approaches often rely on uncertainty-based acquisition functions. AL methods such as least confidence sampling, margin sampling, and entropy-based selection are used to prioritize samples where the model exhibits the highest uncertainty (Lewis and Gale, 1994; Settles, 2009). These methods have demonstrated success across various domains, including text classification, image recognition, and medical data analysis. However, their reliance on a single heuristic limits their ability to explore the broader structure of the data distribution, motivating the investigation of alternative or complementary strategies (Shelmanov et al., 2021).

Recent AL research has explored combining multiple acquisition strategies to overcome the limitations of relying on a single heuristic. Ensemble approaches in active learning aim to leverage the complementary strengths of different selection methods to improve data exploration and model robustness (Beluch et al., 2018; Siddhant and Lipton, 2020). For instance, uncertainty-based techniques can be combined with diversity sampling to ensure that the selected instances are both informative and representative of the overall data distribution (Liu et al., 2019a). Such ensemble methods have shown promise in areas like image classification and natural language processing, but their application to political text classification remains underexplored.

Political text classification presents unique challenges compared to traditional text domains. Data classification or extraction tasks in political documents often exhibit nuanced language, implicit biases, and context-dependent meanings that can vary significantly across regions, cultures, and ideological perspectives (Vosoughi et al., 2018; Brandt and Sianan, 2025; Hamborg et al., 2019). These complexities make annotation particularly difficult and expensive, as domain expertise is frequently required to correctly interpret subtle variations in language and intent. Recent work has explored political bias detection, conflict event extraction, and framing analysis, but the scarcity of annotated datasets and the high variability within political texts continue to pose major obstacles for developing robust NLP models. Synthetic data approaches have recently been proposed (Halterman, 2023), but this raises issues about the verisimilitude of the annotations to the actual texts of interest.

Given the challenges of political text classification and the scarcity of annotated data, effective active learning strategies are essential to improve model performance at lower annotation costs. While ensemble strategies have been explored in active learning, prior work focused predominantly on combining model outputs or uncertainty scores in domains such as image classification and general natural language processing tasks (Beluch et al., 2018; Siddhant and Lipton, 2020). Our approach ensembles acquisition functions directly to enhance the sample selection process itself, improving annotation efficiency without relying on model-level disagreement. The effectiveness of this method is shown in the domain of political text classification, an area where active learning remains underutilized despite the high cost and complexity of annotation.

3 Datasets

Our experiments use the United Nations Parallel Corpus (UNPC) curated by Osorio et al. (2024, 2025). This database comes from a collection of 86,307 official United Nations Security Council resolutions issued between 1990 and 2014 and comprising more than 11 million sentences fully aligned in the six official UN languages created by native speakers (Ziemski et al., 2016). The curated data are a random sample of 11,160 aligned sentences in both English (EN) and Spanish (SP) on key topics such as human rights, protection of civilians, and terrorism. The corpus provides a highly relevant resource for political conflict and violence, making it suitable for evaluating active learning strategies in domain-specific natural language processing tasks. The data annotation process is based on a rigorous annotation protocol involving fluently bilingual human coders to classify the sentences into five classes: non-relevant, verbal cooperation, verbal conflict, material cooperation, and material conflict based on the CAMEO ontology (Gerner et al., 2002). Table 1 presents a set of annotation examples in both English and Spanish according to the different categories considered in the study.

	Cooperation	**Conflict**
Verbal	The delegates agree to proceed *Los delegados acuerdan proseguir*	Students accuse the government *Estudiantes acusan al gobierno*
Material	The UN delivered humanitarian aid *La ONU entregó ayuda humanitaria*	Insurgents attacked the market *Los insurgentes atacaron el mercado*

Not relevant
All the children went to school *Todos los niños fueron a la escuela*

Table 1: Annotation Example

The dataset was partitioned into training and test sets, preserving the class distribution across the labels. Stratified sampling ensures proportional representation of each class across subsets so the resulting distributions are nearly identical, maintaining balance and enabling reliable model evaluation. Despite the stratification, the overall class distribution remains imbalanced, with “Not Relevant” (class 0) comprising over 53% of the data.

The English and Spanish UNPC sentences were each split evenly (50/50) into training and testing sets, with 5,581 sentences each. Choosing a 50/50 approach rather than a more conventional training-heavy split makes the classification task more challenging. For active learning experiments, the initial labeled set consisted of 1% of the training data, with 10 additional samples labeled per round. We conducted 110 active learning rounds across all experiments, including those involving the Ensemble Intersection approach. In contrast, the Ensemble Union method required only 22 rounds to complete, resulting in approximately 20.6% of the training data being labeled by the end of the process.

4 Methods

This section describes the active learning framework adopted in our experiments, the acquisition functions evaluated, the proposed ensemble-based selection methods, and the model training configurations used throughout the study.

Our active learning framework follows a pool-based sampling strategy. Initially, a small subset of labeled examples is used to train a model, while the remaining instances form an unlabeled pool. In each round of active learning, the model predicts over the unlabeled pool, and a subset of samples is selected based on an acquisition function. These selected samples are then labeled and added to the training set, and the model is retrained from scratch with the expanded labeled set. This process repeats iteratively until a predefined annotation budget is exhausted. By carefully selecting the most informative examples in each round, active learning aims to achieve high model performance while minimizing the total number of labeled instances required.

4.1 Acquisition Functions

In active learning, acquisition functions determine which unlabeled samples should be used for annotation in each round. These functions aim to identify the instances that are expected to maximize model improvement if labeled. There are various acquisition strategies focusing on different notions of informativeness, such as uncertainty, margin between predictions, or diversity within the dataset. In this work, we evaluate several widely-used acquisition functions described below.

Top Confidence Sampling: Top Confidence Sampling selects instances where the model exhibits the highest predicted confidence for a single class label. The intuition behind this strategy is that examples with very confident predictions are likely to be redundant or already well-understood by the model, whereas those with lower confidence could be more informative. In practice, active learning frameworks invert this logic and prioritize instances with the lowest confidence scores for labeling (Lewis and Gale, 1994). This method has been widely used as a simple yet effective baseline for uncertainty-based sampling.

Maximum Entropy Sampling: Maximum Entropy Sampling selects instances for which the model’s predicted class distribution has the highest entropy. In this approach, entropy measures the uncertainty associated with the prediction: higher entropy indicates that the model is less confident and more uncertain about the correct label. By prioritizing high-entropy examples, the active learning process focuses on samples that are expected to provide the greatest informational gain when labeled. This method has been widely adopted in active learning research due to its simplicity and effectiveness (Settles, 2009).

Margin Sampling: selects instances where the model's top two class probability scores are closest together. The margin between the highest and second-highest predicted probabilities serves as an uncertainty measure: a smaller margin indicates greater uncertainty, suggesting that the model is unsure which label to assign. By targeting examples with the smallest margins, this strategy aims to find samples near decision boundaries, where additional labeled data could most effectively refine the model. Margin Sampling has been widely studied as a competitive uncertainty-based selection method in active learning (Scheffer et al., 2001).

Monte Carlo Dropout Sampling: estimates model uncertainty by applying dropout at inference time and performing multiple stochastic forward passes through the model. The variance in the predicted probabilities across these passes serves as a measure of uncertainty: higher variance indicates greater uncertainty about the correct label. By selecting instances with the highest predictive variance, this method aims to identify examples that the model finds most ambiguous. Monte Carlo Dropout Sampling has been widely adopted for active learning with deep neural networks, particularly for leveraging Bayesian uncertainty estimates (Gal and Ghahramani, 2016).

Core-set Sampling: selects a subset of unlabeled instances that best represent the overall data distribution. It formulates active learning as a coverage problem, aiming to minimize the distance between the labeled and unlabeled data points in feature space. By prioritizing diverse and representative samples, Core-set Sampling helps ensure that the model generalizes better across the dataset. This method has been particularly influential for active learning with deep neural networks, where embedding-based diversity becomes crucial (Sener and Savarese, 2018).

4.2 Proposed Ensemble Methods

While individual acquisition functions capture different notions of uncertainty or diversity, relying on a single active learning strategy may still limit data exploration during learning process. To address this, we propose two ensemble-based methods that combine the outputs of multiple acquisition functions. The first method, *Ensemble Union*, aggregates top-ranked samples from different heuristics to enhance exploration within each annotation round. The second method, *Ensemble Intersection*, selects samples based on majority voting across heuristics, prioritizing instances consistently identified as informative. These ensemble approaches aim to balance exploration and exploitation more effectively, leading to improved annotation efficiency and model performance.

Ensemble Union. This method combines the top-ranked samples selected by multiple acquisition functions in each active learning round. Each acquisition function independently ranks the unlabeled instances by its selection criterion. From each acquisition function, the top m samples are selected, resulting in a candidate pool containing up to $n \times m$ samples, where n is the number of acquisition functions utilized. Duplicate instances are selected only once if chosen by multiple strategies.

Each sample is assigned a positional score based on its rank within its respective acquisition list. For a sample i ranked in list L, the positional score is:

$$\text{score}_L(i) = N - \text{rank}_L(i)$$

where N is the number of selected samples per strategy (typically $N = 50$). To aggregate scores across strategies, the final score for each unique sample is defined as the maximum positional score it achieves across all acquisition functions:

$$\text{FinalScore}(i) = \max_{L \in S}(\text{score}_L(i)),$$

where S is the set of all acquisition strategies. From all unique samples, the top K samples based final scores are selected for annotation in each round.

By selecting multiple samples from different heuristics and aggregating them, Ensemble Union promotes broader exploration of the dataset, leading to faster coverage of informative regions. Since more samples are labeled per round, the total number of active learning rounds r is reduced proportionally according to $r = \frac{t}{n}$ where t is the number of rounds a single acquisition function would require to reach the target labeling budget. And n is the number of acquisition functions. Figure 1 illustrates the sample aggregation and duplicate removal process used in Ensemble Union. Here, the sample with index 6 from Strategy A, index 3 from Strategy B, and indexes 1, 10, and 6 from Strategy C are excluded due to duplication, as they were already selected by other acquisition functions.

An additional advantage of the proposed approach is improved efficiency in execution time. Achieving the target budget of 54 data points using a single active learning strategy would require

6 rounds, assuming a selection of 9 samples per round. In contrast, the Ensemble Union method reaches the same target in only 3 rounds, with each round acquiring 18 samples. This reduction in the number of iterations highlights the efficiency gains offered by the ensemble approach.

In our approach, each strategy contributes a full batch of instances per round. This ensures sufficient coverage, even in cases where multiple strategies select the same samples. The Ensemble Union method also maintains balanced representation across acquisition functions, with data points evenly distributed when selections differ.

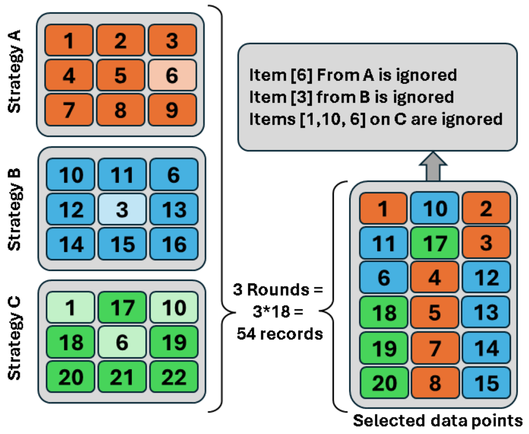

Figure 1: Ensemble Union selection and removal.

Ensemble Intersection. This method aggregates selections from multiple acquisition functions based on majority voting and positional scoring in these steps:

1. **Training:** Fine-tune the model using the current labeled dataset.

2. **Querying:** Each acquisition function independently selects its top k samples from the unlabeled pool, producing ranked lists $L_1, L_2, \ldots, L_n$.

3. **Aggregation and Scoring:** The selected samples are aggregated and scored based on two components:

   - **Frequency Score** $f(x)$ counts how many acquisition functions selected a sample x:

   $$f(x) = \sum_{i=1}^{n} \mathbb{1}\{x \in L_i\}$$

   - **Position Score** $p(x)$ assigns higher weights to samples appearing earlier in ranked lists:

   $$p(x) = \sum_{i=1}^{n} \sum_{j=0}^{k-1} \mathbb{1}\{x = L_i[j]\} \times (k - j)$$

   Samples are ranked first by descending frequency score $f(x)$; ties are broken by descending position score $p(x)$. The top b samples according to this combined ranking are selected for labeling.

4. **Updating:** The newly selected samples are added to the labeled set and removed from the unlabeled pool.

This aggregation approach prioritizes samples that are consistently deemed informative across multiple acquisition functions while also rewarding those ranked highly within individual methods. Figure 2 illustrates the scoring and selection process for Ensemble Intersection. In this example, the selected data points are prioritized from left to right, top to bottom, based on a combination of their frequency across strategies and their relative position within each ranking. Records with indices 4, 1, 10, and 20 appear in two or more strategies, indicating higher consensus and, by extension, stronger indicative value. As a result, these records are prioritized during selection. This approach is illustrated in Strategy C, where record index 18—despite being ranked earlier than records such as 4 and 20—was not selected, highlighting that consensus among strategies can outweigh individual ranking in a single list. The weight table is shown in the figure, along with sample weight calculations for selected records. For instance, the weight of the record with index 4 is computed by summing its contributions across all its occurrences: it receives a score of 6 from Strategy A, 6 from Strategy B, and 5 from Strategy C, resulting in a total weight of 17. As a result, records that are both commonly selected and highly ranked across strategies are prioritized. Unlike the ensemble union method, the ensemble intersection requires the same number of iterations as an individual active learning strategy because it does not increase the number of selected samples per round. Instead, it prioritizes and selects only those data points that are jointly recognized by multiple strategies and rank highly within each. As illustrated in the figure, to reach a target of 54 samples, the method must run for 9 rounds, with each round contributing 6 prioritized records. This ensures that the selected samples are agreed upon by multiple strategies and reflect higher consensus importance across them.

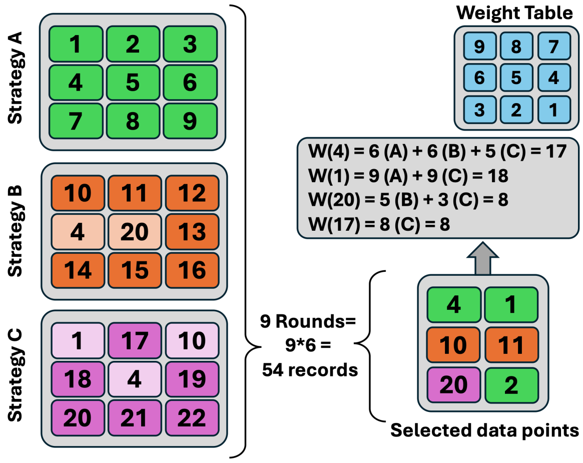

Figure 2: Ensemble Intersection scoring and selection.

4.3 Model Training Setup

To evaluate the generalizability of our active learning methods, we fine-tune a diverse set of pre-trained transformer models. For the English UNPC dataset, we use BERT-base-uncased (Devlin et al., 2018), DistilBERT-base-uncased (Sanh et al., 2019), and RoBERTa-base (Liu et al., 2019b), alongside ConfliBERT (Hu et al., 2022) for domain-specific political text. We include two categories of ConfliBERT models: *ConfliBERT-Cont.*, which are initialized from BERT and further pre-trained on conflict-related texts, and *ConfliBERT-Scr.*, which are trained from scratch using the same domain corpus. For Spanish datasets (UNPC-Spanish), we fine-tune BETO, a Spanish pre-trained BERT model (Cañete et al., 2020).

The selection includes compact models (DistilBERT), robust models trained with larger corpora (RoBERTa), and domain-adapted models (ConfliBERT), ensuring that our ensemble strategies are tested across varied architectures and pretraining regimes. This diversity allows us to verify that the proposed methods improve performance consistently across different model types.

All models are fine-tuned from pre-trained checkpoints without additional pretraining. Fine-tuning is performed using the AdamW optimizer with a learning rate of 1×10^{-5}, $\epsilon = 1 \times 10^{-8}$, and a batch size of 32. Models are trained for a maximum of 10 epochs per annotation round, with early stopping based on validation loss if no improvement is observed for 3 consecutive epochs. A dropout rate of 0.9 is used where applicable to facilitate uncertainty estimation through Monte Carlo Dropout Sampling. During each active learning round, the model is retrained from scratch using the expanded labeled set. Hyperparameters such as learning rate, batch size, and training procedure were kept consistent across all acquisition strategies and ensemble methods to ensure fair comparisons.

5 Experimental Setup

We evaluate all active learning strategies and ensemble methods under consistent experimental conditions. The initial labeled sets, unlabeled pools, and train/test splits follow the configurations described in Section 3. Active learning proceeds iteratively by querying new samples, labeling them, and retraining the model at each round (Schröder et al., 2023).

In addition to the active learning methods, we establish two baselines. First, a *Random Sampling Baseline*, where 10 randomly selected unlabeled samples are added to the labeled set per round without using acquisition heuristics. Second, a *Full Dataset Training baseline*, where each model is trained on the entire available labeled training set to measure the upper bound of F1 performance.

Standard acquisition functions select 10 new samples per round, while ensemble methods such as Ensemble Union and Ensemble Intersection aggregate selections from multiple strategies, labeling larger batches to accelerate dataset coverage. The total annotation budget was set to approximately 20% of the training set size for each dataset.

Performance is evaluated every 5 rounds on the held-out test set, using F1-score as the primary evaluation metric. Our methodological approach runs each experiment across three random seeds, and reports the averaged results.

6 Results and Analysis

The study evaluates seven active learning strategies, including five standard acquisition functions and two proposed ensemble methods on the UNPC data in both English and Spanish using various BERT-based models. To establish performance baselines, we also trained each model on the full dataset and implemented a random sampling strategy. As reported in Table 2, the evaluation compares the best F1-scores achieved by each method under comparable annotation sets, providing a comprehensive assessment of active learning effectiveness.

Dataset	Model	Benchmark		Active Learning Strategies						
		Full Dataset	Random	Top Conf.	Max Ent.	MC Drop.	Margin	Coreset	Ens.	Ens. Union
UNPC English	ConfliBERT Cont. Cased	77.93	73.88	75.33	75.40	74.88	75.47	73.65	**76.47**	76.44
	ConfliBERT Cont. Uncased	78.71	71.13	74.96	73.02	73.09	74.20	71.83	74.32	**75.29**
	ConfliBERT Scr. Cased	75.88	72.88	75.46	75.02	74.99	75.61	73.97	**76.40**	76.39
	ConfliBERT Scr. Uncased	78.80	72.99	75.33	75.23	75.20	75.43	74.50	75.50	**76.44**
	BERT Cased	78.85	71.65	74.33	73.56	73.32	75.03	73.90	**75.66**	75.55
	BERT Uncased	78.91	72.29	74.01	74.20	74.01	74.01	73.56	74.99	**75.80**
	RoBERTa-Base	77.23	73.98	76.47	75.71	75.60	74.93	76.40	75.92	**76.80**
	DistilBERT-Base Uncased	68.86	60.03	61.42	63.00	64.56	64.61	62.85	63.59	**66.94**
UNPC Spanish	BETO Cased	77.24	74.09	75.78	76.22	75.55	75.93	74.78	**76.25**	76.17
	BETO Uncased	75.46	72.78	75	75.09	75.22	74.35	74.65	74.45	**75.27**

Note: results in bold font indicate top performing active learning strategy for a given model.

Table 2: Summary of results across datasets, models, and active learning strategies. (Best F1 Scores)

Overall Performance Trends.

Across all datasets and models, the proposed ensemble methods—Ensemble Union and Ensemble Intersection—consistently outperform individual active learning strategies as well as the random sampling baseline. Notably, Ensemble Union frequently achieves the highest F1-scores. This performance gain stems from aggregating multiple acquisition heuristics, which allows the ensemble to capture diverse and complementary aspects of uncertainty and representativeness. By integrating these strengths, the ensemble methods enable more robust, informative, and diverse sample selection, leading to improved model generalization. As a result, models trained on a fraction of the data selected through these methods often match or exceed the performance of models trained on the full dataset, underscoring the practical value of ensemble-based active learning in reducing labeling costs. These findings suggest that ensembling offers a simple yet powerful mechanism to boost active learning effectiveness without introducing complex modeling overhead.

Model-Level Insights. Model selection is key for active learning outcomes. Models pre-trained on domain-specific political text, particularly ConfliBERT, consistently outperform general-purpose models. This performance boost underscores the advantage of using domain-specific language models for specialized classification tasks. It seems that selecting the right combination of active learning strategy and the appropriate language model is the key to achieve high levels of performance. In contrast, DistilBERT—designed as a compact and faster alternative to BERT—consistently shows lower F1-scores compared to larger transformer models. DistilBERT's reduced parameter count and compressed training objectives, while beneficial for efficiency, appear to trade off representational capacity necessary for high-accuracy political text classification. This suggest that model capacity and pretraining domain alignment both critically influence active learning effectiveness.

Language-Level Insights. Results in Table 2 reveal some performance differences between English and Spanish datasets across models and active learning strategies. Models fine-tuned on Spanish texts (BETO) generally achieved slightly lower F1-scores when compared to models trained on English datasets. These differences may reflect variations in the model's pretraining corpus size, language modeling resources, or domain-specific adaptation between English and Spanish. Nevertheless, ensemble-based active learning methods consistently improved model performance across both languages, demonstrating the leverage of the proposed ensemble strategies.

Ensemble Union Efficiency. Beyond improvements in F1 performance, the Ensemble Union method demonstrated a significant practical advantage: it requires fewer active learning rounds compared to traditional acquisition strategies. Aggregating multiple sampling heuristics and selecting a larger batch of samples per round, Ensemble Union accelerated exploration of the unlabeled dataset, reducing the number of iterations needed to reach the target annotation budget. This efficiency lowers annotation overhead and speeds model development cycles. Even with fewer total rounds, Ensemble Union achieved performance comparable to or exceeding that of methods that operated with smaller batch sizes across more rounds. These findings show that ensemble-based active learning improves sample efficiency, leading to both higher performance and reduced human labeling effort.

AL Performance and Sample Size. In essence, active learning strategies aim at minimizing the amount of human effort, resources, and attention necessary to achieve high levels of performance. As Figure 3 shows, all active learning strategies improve their performance as the size of the annotation set increases. In line with the results discussed above, the Ensemble Intersection and Ensemble Union show the best performance. With the exception of DistilBERT–that consistently reports the lowest F1 scores–all other panels in the plot indicate that the Ensemble strategies proposed in this study achieve high levels of performance with about 5% of annotations. Such remarkably high levels of performance with such a small sample size represent considerable savings for researchers in terms of human effort and financial resource.

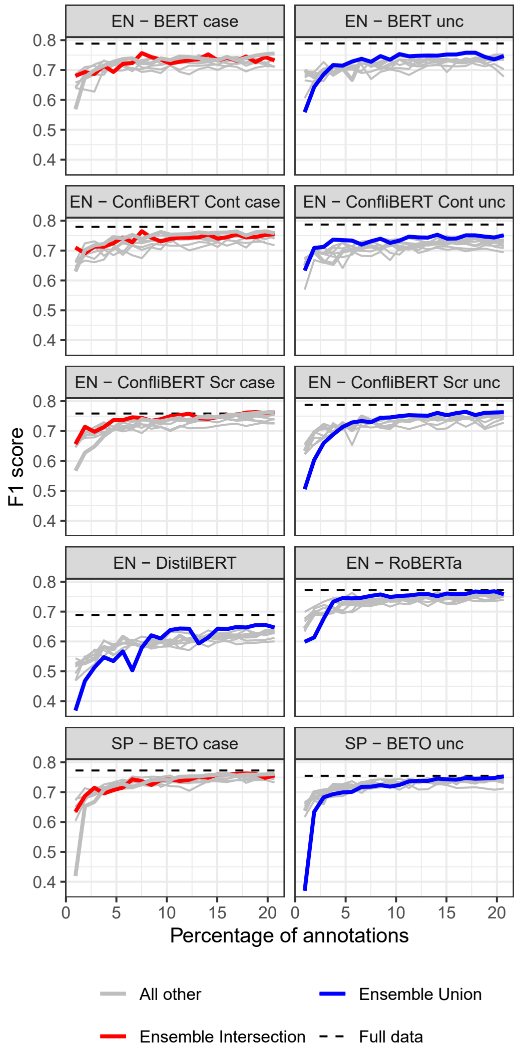

Figure 3: Active Learning Performance by Annotation Percentage in English (EN) and Spanish (SP) texts. *Note: Red and blue lines indicate the best performing active learning strategies per model.*

7 Conclusion

This paper introduced two ensemble-based active learning strategies—Ensemble Union and Ensemble Intersection—to improve data exploration and annotation efficiency in political text classification tasks. We evaluated these methods across multiple models and two datasets, covering English and Spanish political texts. Our results demonstrate that ensemble active learning consistently outperforms individual acquisition strategies and random baselines, achieving performance close to or exceeding models trained on full datasets while using significantly fewer labeled examples.

Ensemble Union, in particular, offered practical advantages by accelerating dataset coverage and reducing the number of active learning rounds needed to reach target annotation budgets. Ensemble Intersection showed robust performance across both large and small datasets, demonstrating stability in sample selection even under constrained labeling scenarios. The findings suggest that ensemble-based active learning provides a viable and efficient framework for annotation-intensive natural language processing tasks, particularly in politically-oriented and multilingual domains.

Future work includes the development of Dynamic Active Learning approaches, where reinforcement learning agents adaptively select sampling strategies based on observed model and data characteristics. Additionally, extending evaluation to cybersecurity and biomedical datasets, and further exploring multilingual settings such as Arabic texts, are promising directions for demonstrating the generalizability and scalability of active learning methods across diverse domains.

Acknowledgments

This research was supported by NSF award 2311142; used Delta at NCSA / University of Illinois through allocation CIS220162 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by NSF grants 2138259, 2138286, 2138307, 2137603, and 2138296. This material is based upon High Performance Computing (HPC) resources supported by the University of Arizona TRIF, UITS, and Research, Innovation, and Impact (RII). Sultan Alsarra would like to extend his appreciation to the Deanship of Scientific Research at King Saud University for funding his work through Waed Program (W25-99).

References

Luay Abdeljaber, Naif Alatrush, Sultan Alsarra, Mahrusa Billah, Afraa Alshammari, and Latifur Khan. 2025. Active learning for medical text classification: Insights from ophthalmology and broader medical domains. In *2025 IEEE 13th International Conference on Healthcare Informatics (ICHI)*, pages 216–227.

Frank Baumgartner, Bryan Jones, and Michael MacLeod. 1998. Lessons from the Trenches: Ensuring Quality, Reliability, and Usability in the Creation of a New Data Source. *The Political Methodologist*, 8(2):1–10.

William H Beluch, Tim Genewein, Bernd Bauer, and Joachim M Buhmann. 2018. The power of ensembles for active learning in image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9368–9377.

Patrick T. Brandt and Marcus Sianan. 2025. Measurement of event data from text. *Frontiers in Political Science*, Volume 6 - 2024.

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2020. Spanish pre-trained bert model and evaluation data. In *PML4DC at ICLR 2020*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR.

Yarin Gal, Riashat Islam, and Zoubin Ghahramani. 2017. Deep bayesian active learning with image data. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1183–1192.

Deborah J Gerner, Philip A Schrodt, Omür Yilmaz, and Rajaa Abu-Jabr. 2002. Conflict and mediation event observations (cameo): A new event data framework for the analysis of foreign policy interactions. *International Studies Association, New Orleans*.

Andrew Halterman. 2023. Synthetically generated text for supervised text analysis. *Political Analysis*, pages 1–14.

Felix Hamborg, Karsten Donnay, and Bela Gipp. 2019. Automated identification of media bias in news articles: an interdisciplinary literature review. *International Journal on Digital Libraries*, 20(4):403–429.

Michael Heseltine and Bernhard Clemm von Hohenberg. 2024. Large language models as a substitute for human experts in annotating political text. *Research & Politics*, 11(1):20531680241236239. Publisher: SAGE Publications Ltd.

Yibo Hu, MohammadSaleh Hosseini, Erick Skorupa Parolin, Javier Osorio, Latifur Khan, Patrick Brandt, and Vito D'Orazio. 2022. ConfliBERT: A pre-trained language model for political conflict and violence. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5469–5482, Seattle, United States. Association for Computational Linguistics.

David D Lewis and William A Gale. 1994. A sequential algorithm for training text classifiers. *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 3–12.

Mitchell Linegar, Rafal Kocielnik, and R. Michael Alvarez. 2023. Large language models and political science. *Frontiers in Political Science*, 5. Publisher: Frontiers.

Shiyu Liu, Weitong Huang, Xipeng Liu, and Xiaodan Zhu. 2019a. Learning to sample: An active learning framework. In *Proceedings of the 2019 IEEE International Conference on Data Mining (ICDM)*, pages 687–696. IEEE.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Javier Osorio, Sultan Alsarra, Amber Converse, Afraa Alshammari, Dagmar Heintze, Latifur Khan, Naif Alatrush, Patrick T. Brandt, Vito D'Orazio, Niamat Zawad, and Mahrusa Billah. 2024. Keep it Local: Comparing Domain-Specific LLMs in Native and Machine Translated Text using Parallel Corpora on Political Conflict. In *2024 2nd International Conference on Foundation and Large Language Models (FLLM)*, pages 542–552.

Javier Osorio, Afraa Alshammari, Naif Alatrush, Dagmar Heintze, Amber Converse, Sultan Alsarra, Latifur Khan, Patrick T. Brandt, and Vito D'Orazio. 2025. The devil is in the details: Assessing the effects of machine-translation on llm performance in domain-specific texts. In *Proceedings of Machine Translation Summit XX, Volume 1*, pages 315–332. European Association for Machine Translation.

Fariha Rahman, Sadaf Halim, Anoop Singhal, and Latifur Khan. 2024. Alert: Active learning enhanced robust threat-mapper for efficient annotation of cyber threat intelligence reports. In *Data and Applications Security and Privacy XXXVIII (DBSec 2024)*, San Jose, CA, USA. Springer International Publishing.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Tobias Scheffer, Christian Decomain, and Stefan Wrobel. 2001. Active hidden markov models for information extraction. In *International symposium on intelligent data analysis*, pages 309–318. Springer.

Christopher Schröder, Lydia Müller, Andreas Niekler, and Martin Potthast. 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*, pages 84–95, Dubrovnik, Croatia. Association for Computational Linguistics.

Ozan Sener and Silvio Savarese. 2018. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations (ICLR)*.

B Settles. 2009. Active learning literature survey (computer sciences technical report 1648) university of wisconsin-madison. *Madison, WI, USA: Jan*.

Artem Shelmanov, Dmitri Puzyrev, Lyubov Kupriyanova, Denis Belyakov, Daniil Larionov, Nikita Khromov, Olga Kozlova, Ekaterina Artemova, Dmitry V. Dylov, and Alexander Panchenko. 2021. Active learning for sequence tagging with deep pre-trained models and Bayesian uncertainty estimates. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1698–1712, Online. Association for Computational Linguistics.

Aditya Siddhant and Zachary C Lipton. 2020. Bayesian active learning for natural language processing: a survey. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 609–623.

Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science*, 359(6380):1146–1151.

Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The United Nations parallel corpus v1.0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3530–3534, Portorož, Slovenia. European Language Resources Association (ELRA).