Benchmarking Active Learning Techniques: Insights from Multi-Domain Fake News Detection

Sergio Flesca[1,2,†], Marco Gagliardi[1,2,†], Danilo Maurmo[1,2,†], Francesco Scala[3,*,†] and Eugenio Vocaturo[2,1,†]

[1]DIMES - Department of Computer Engineering, Electronic Modeling, and Systems Engineering, University of Calabria, Rende, Italy

[2]CNR-NANOTEC Institute of Nanotechnology of the National Research Council, Rende, Italy

[3]CNR-ICAR Institute of High Performance Computing and Networking of the National Research Council, Rende, Italy

Abstract

The rapid spread of fake news poses a major societal challenge, requiring efficient and generalizable detection methods. Active Learning offers a viable solution by reducing annotation costs while enhancing model performance. This study benchmarks multiple Active Learning strategies for fake news detection across two distinct domains: political discourse (Politifact) and entertainment news (GossipCop). We evaluate uncertainty-based methods (Entropy Sampling, Least Confidence) alongside more advanced techniques (Core-Set, K-Means, BADGE, BALD), assessing their effectiveness, efficiency and sustainability. Our findings highlight Entropy Sampling as the most accurate approach, particularly in the political domain, while K-Means emerges as the most computationally efficient. Additionally, we analyze the environmental impact of Active Learning-based training, underscoring its role in optimizing both performance and resource consumption. These insights contribute to the development of scalable and energy-efficient misinformation detection systems.

Keywords

Active learning, Cross-domain, Multi-domain, fake news detection, Natural Language Processing, NLP

1. Introduction

The rapid development of the World Wide Web since the mid'90s has significantly transformed the way the people communicate. Online social media platforms like X (old Twitter) and Facebook improve real-time information spread. Social media has become the primary platform for online interaction and information exchange thanks to their ease of use, low cost and fast dissemination[1]. However, the internet has also become a place for fake news sharing such as misleading information, fake reviews, deceptive advertisements, rumors and false political statements. As a result, fake news has emerged as a major issue for both industry and academia, as it is widely used to mislead and manipulate online users with biased or false information. Fake news can have dangerous effects on both individuals and society [2]. First, it can mislead people leading them to accept false beliefs. A well-known study by Pennycook et al.[3] demonstrates

SEBD 2025: 33rd Symposium on Advanced Database Systems, June 16-19, 2025, Ischia, Italy

*Corresponding author.

†These authors contributed equally.

sergio.flesca@unical.it (S. Flesca); marco.gagliardi@dimes.unical.it (M. Gagliardi); danilo.maurmo@dimes.unical.it (D. Maurmo); francesco.scala@icar.cnr.it (F. Scala); eugenio.vocaturo@cnr.it (E. Vocaturo)

0000-0002-4164-940X (S. Flesca); 0009-0009-7839-1682 (M. Gagliardi); 0009-0000-0367-2337 (D. Maurmo); 0009-0007-5224-0910 (F. Scala); 0000-0001-7457-7118 (E. Vocaturo)

that repeated exposure to fake news increases its perceived truthfulness. This is a phenomenon known as the *implied truth effect*. Even individuals with higher media literacy are susceptible to this cognitive bias because familiarity encourages a sense of reliability. Second, fake news can alter how people perceive and react to true ones. When individuals encounter debunked fake news, they may develop an increased scepticism toward all news, including legitimate reports. Lewandowsky et al.[4] describe the *backfire effect*: the attempts to correct false beliefs can inadvertently reinforce them. Third, the widespread dissemination of fake news can mine the credibility of the entire news ecosystem. A study by Lazer et al.[5] highlights how the large-scale spread of disinformation affects democracies by encouraging suspiciousness in key institutions, including the media, healthcare and government. Thus, fake news has become a serious social problem that cannot be ignored.

Contribution In this paper we aim to analyze and compare various active learning techniques for the detection of fake news, evaluating their performance to identify the most effective approaches across some domains. Some classification-based and emissions-based metrics are employed to ensure the best evaluation possible. In particular, the inclusion of energy consumption and carbon emission as evaluation criteria underscores the importance of sustainability.

Organization The rest of this paper is structured as follows: Section 2 provides an overview of existing approaches and highlights the role of active learning in fake news detection. Section 3 outlines our methodology and experimental setup, active learning techniques, and evaluation metrics used. Section 4 presents our experimental study, offering a comparative analysis of the methods across different datasets, detailing them and evaluation metrics. Finally, the conclusion section 5 summarizes key findings, discusses the main contributions and limitations of our work, and suggests directions for future research.

2. Literature review

Developing a text classifier for a new problem requires access to training data and their corresponding labels. Labeling is typically performed by human annotators and the common approach involves labeling as many text documents as possible, training a classifier and acquiring additional data and labels if the performance is ineffective. However, randomly selecting documents to extend the dataset can be inefficient because the newly added documents may not add valuable information for classification. Active learning aims to optimize this process by selecting the most “difficult" unlabeled documents and querying annotators for their labels [6, 7, 8, 9]. This strategy has the potential to significantly reduce the effort required for developing a new classification system [10].

2.1. Active Learning applied on cross-domain Fake News Detection

Some recent studies explored active learning with various methodologies to improve the efficiency and accuracy of misinformation detection [11, 12]. The study by Bhattacharjee [13] shifts the focus to news veracity detection, proposing a collaborative human-machine learning framework. The approach employs a deep-shallow fusion model, where insights from a deep learning classifier are combined with shallow feature-based models leveraging linguistic and content-specific features. This allows the model to adapt to various domains or news genres, improving its generalizability. The results indicate that this fusion-based method leads to a

25% improvement in performance, with the model requiring significantly fewer annotated samples. The combination of shallow and deep features, along with dynamic feature weighting, proves to be a powerful strategy for detecting fake news. The use of prioritized active learning ensures that only the most “difficult" samples are selected for human annotation, further reducing the labeling cost. While many fake news detection frameworks rely exclusively on deep learning, this work advocates for combining deep and shallow models to capture both detailed contextual information and broader linguistic patterns. This fusion approach ensures a more comprehensive analysis of news content.

Farinneya et al. [14] explores the challenge of detecting rumors on social media, focusing particularly on Twitter data. The goal is to leverage the limited domain-specific labeled data and utilize information from other domains by combining Active Learning with Transfer Learning. Various active learning strategies are applied, including Least Confidence [6] and Query by Committee [15]. The results demonstrate that the method effectively achieves performance comparable to fully supervised models while using only 42% of the labeled dataset, with faster convergence in terms of F1 score. In terms of methodology, this work uniquely distinguishes itself by its ability to apply knowledge from different domains to enhance rumor detection. This is crucial as rumors in one context may share characteristics with those in other domains. Additionally, the use of *TweetBERT* for textual data representation is notable, as it is specifically fine-tuned for social media content, contributing to the solid performance observed.

The study by Sahan et al. [10] adopts a more traditional approach to fake news detection, focusing on text-based classification using Active Learning (AL). The authors explore how standard AL techniques can reduce the labeling burden in text-based fake news detection tasks. Unlike the domain adaptation discussed earlier, this study focuses on selecting the most informative text samples for human labeling based on uncertainty. This study serves as a useful benchmark, showing how traditional AL methods can play a valuable role in text-based disinformation detection, especially when compared to more complex techniques like domain adaptation or geometric deep learning.

Lee et al.[16] take a different approach by addressing the issue of domain adaptation in disinformation detection, particularly for emerging topics such as the COVID-19 pandemic or the war in Ukraine. In these cases, existing models trained on older or unrelated domains may perform poorly due to differences in data distribution between the source and target domains. To address this challenge, is proposed an energy-based domain adaptation framework that integrates active learning. The key innovation is the combination of energy-based models with AL to transfer knowledge from domains with abundant labeled data to those with limited labeled data. The model minimizes the energy gap between the source and target domains to align their representations, ensuring that knowledge from source domains can be effectively applied to emerging domains. AL is used to selectively label the most uncertain samples in the target domain, thereby improving model performance with minimal labeled data. Experiments show that this method improves accuracy by 10% in domain adaptation tasks compared to baseline models, highlighting the framework’s effectiveness in adapting to new and evolving disinformation contexts. This work is notable for its focus on cross-domain learning, an often-overlooked area in disinformation detection research, but one that is crucial for addressing misinformation on new and unfamiliar topics.

Kato et al. [17] examines domain bias in supervised fake news detection. The study highlights

how models trained on domain-specific datasets struggle with generalization and proposes a strategy to mitigate this bias using paired datasets. Using FakeNewsAMT, a dataset with paired real and fake news on the same topic, the authors analyze domain bias through deep learning models, particularly BERT [18]. They experiment with noun phrase masking to mitigate bias but observe no improvement in accuracy, indicating that named entities are not the primary source of domain bias. Instead, they identify significant lexical overlap between paired real and fake news, which helps models generalize better across domains. Comparing models trained on paired wrt unpaired datasets, they observe that paired data improve cross-domain detection accuracy significantly in most cases. This suggests that dataset structure, specifically pairing real and fake news with similar lexical patterns, plays a key role in mitigating domain bias. This study contributes to domain adaptation research by showing that dataset design itself can enhance model generalization, offering an alternative to adversarial training and domain-invariant feature extraction. Future research could explore similar dataset structures in different misinformation contexts, such as scientific or health-related fake news.

Barnabò et al.[19] explore the application of graph neural networks for misinformation detection. This study highlights the shift from traditional text classification methods to geometric deep learning, focusing on the dissemination patterns of news articles across social networks. The authors argue that fake news detection should not rely solely on news content but also consider how information spreads through networks. The main contribution of this study is the development of Deep Error Sampling, a novel AL strategy that selects samples based on prediction errors combined with uncertainty sampling. The advantage of using GNNs in this context is their ability to model relationships between users and the articles they share, thus capturing the social dynamics of disinformation propagation. By applying AL to GNNs, the authors reduce the annotation burden for human fact-checkers, who play a crucial role in verifying the authenticity of news articles. Results show that Deep Error Sampling outperforms traditional AL methods, reducing labeling costs by up to 25% while achieving a 2% improvement in the area under the curve. This approach highlights the potential of graph-based representations for fake news detection, especially when combined with AL to enhance efficiency. While previous studies primarily focused on text-based models, this work demonstrates that understanding the propagation behavior of disinformation is equally important, offering a new perspective on how news is classified as true or false.

The study by Folino et al.[20] introduces an approach addressing the challenge of resource efficiency in fake news detection. The authors propose a semi-supervised method that combines active learning (AL) with a BERT model. By integrating AL with a BERT model, the approach seeks to reduce computational costs and human effort. This is particularly important in scenarios with significant constraints on memory, time and energy, such as small organizations or non-profits lacking resources for large-scale model training. Compared to other approaches, the key contribution of this work lies in its resource efficient design. While many studies focus on maximizing model performance, this work emphasizes the importance of balancing accuracy with computational demands, especially when deploying pretrained language models in real-world applications [21, 22].

All the studies trivially tackle the challenge of fake news detection using active learning, but the diversity of methodologies underlines that there is no unique solution. The choice of methodology depends on data characteristics, resource availability and the nature of fake

news. Approaches can differ significantly: some optimize computational efficiency while others focus on network dynamics or domain adaptation. A study integrates transfer learning with active learning, another employs feature representation, another one utilizes heterogeneous graph neural networks to capture relationships among diverse entities while another addresses multilingual fake news detection via a multi-model neural ensemble. These strategies showcase distinct strengths and offer complementary solutions.

3. Methodology

3.1. Active Learning techniques

Uncertainly sampling: Least confidence and Entropy-based algorithms The simplest and commonly used query framework is *uncertainty sampling* [23]. In this approach, an *active learner* selects the instances for which it has the highest uncertainty in label assignment.

- **Least Confidence** [6]: let p be the probability of the most likely class for a data instance x. Then the least confidence score assigned to x is simply computed as $1 - p$:

$$x^{LC}{=}\arg\min_x P(y^{|x;\theta)} \tag{1}$$

  where $y^{=\arg\max_y P(y|x;\theta)}$ represents the most likely class label;
- **Entropy** [24]: measures the overall uncertainty across all classes. A high entropy value indicates the model is unsure about the correct class. For a data instance x, if there are C classes and p_i is the probability of the i-th class, the entropy is calculated as:

$$x^{ENT}{=}\arg\max_x -\sum_i P(y_i|x;\theta)\log P(y_i|x;\theta) \tag{2}$$

  where y_i ranges over all possible labels. Entropy is an information-theoretic measure that quantifies the uncertainty of a distribution and is often used as an impurity indicator.

CoreSet sampling The core-set selection method theorized by [25] identifies a subset of data points such that training a model on this subset achieves competitive performance on the entire dataset. In Active Learning, only a sub-sampled pool of labeled data is available, with a query budget b and a learning algorithm A_s that outputs parameters w given a labeled set s. The pool-based optimization problem is defined as:

$$\min_{s^1:|s^1|\leq b} E_{x,y\sim p_z}\left[l(x,y;A_{s^0\cup s^1})\right] \tag{3}$$

Batch Active Learning extends this by selecting a large batch of points for labeling in each iteration. The AL loss is upper-bounded as:

$$E_{x,y\sim pz}\left[l(x,y;A_s)\right] \leq \underbrace{\left|E_{x,y\sim pz}\left[l(x,y;A_s)\right] - \frac{1}{n}\sum_{i\in[n]} l(x_i,y_i;A_s)\right|}_{\text{Generalization Error}} + \underbrace{\left(\frac{1}{|S|}\sum_{j\in S} l(x_j,y_j;A_s)\right)}_{\text{Training Error}} + \underbrace{\left(\frac{1}{n}\sum_{i\in[n]} l(x_i,y_i;A_s) - \frac{1}{|S|}\sum_{j\in S} l(x_j,y_j;A_s)\right)}_{\text{Core-Set Loss}} \tag{4}$$

The population risk of the model trained on s is influenced by Training Error, Generalization Error, and Core-Set Loss. Given that CNNs exhibit low Training Error and Generalization Error is provably bounded, the key challenge is minimizing Core-Set Loss:

$$\min_{s^1:|s^1|\leq b}\left|\frac{1}{n}\sum_{i\in[n]} l(x_i,y_i;A_{s^0\cup s^1}) - \frac{1}{|s^0+s^1|}\sum_{j\in s} l(x_j,y_j;A_{s^0\cup s^1})\right| \tag{5}$$

K-means sampling K-means sampling in active learning introduced in [26] addresses the limitations of traditional uncertainty sampling. While uncertainty sampling selects the most uncertain instance near the decision boundary, it does not consider the overall data distribution, which is crucial for batch selection. K-means sampling overcomes this by first identifying a set of uncertain instances within the margin and then clustering them based on feature similarity. From each cluster, the most central instance (the medoid) is selected for labeling, ensuring that the chosen samples are both informative and representative of different regions of the feature space. This method offers several advantages. It prevents the classifier from focusing too narrowly on specific areas by preserving the density distribution of the dataset, avoiding redundancy, and promoting diversity in the selected samples. This is particularly beneficial in high-dimensional domains like text classification, where redundant training samples can slow down model convergence. Additionally, representative sampling contributes to a more balanced reduction of the hypothesis space, improving generalization.

BALD: Bayesian Active Learning by Disagreements BALD[27] is an Active Learning model applied to Gaussian Processes for Classification. It expresses information gain through predictive entropies, identifying the set of points among the unlabeled examples that maximize the expected decrease in Shannon entropy[24]. In this framework, a set of points is considered, consisting of some labeled examples and others with unknown labels D, along with the existence of a group of latent parameters θ that manage the dependency between input data and their labels $p(y \mid x, \theta)$. The central objective is to reduce the number of possible hypotheses as

quickly as possible, minimizing uncertainty over the model parameters through entropy. The points in D' are chosen from D to maximize the expected entropy decrease, meaning they are points where the parameters are more certain but, at the same time, on which there is the greatest disagreement:

$$\arg\min_{D} H[\theta \mid D'] = -\int p(\theta \mid D') \ln p(\theta \mid D') d\theta \tag{6}$$

The selection criterion consists of choosing data that maximize the disagreement of the parameters between the current model and its subsequent updates[28]. Solving equation 6 is an NP-Hard problem; therefore, a greedy approximation strategy is applied, allowing work on individual elements x rather than the entire set of unlabeled samples D'. The new objective is then to find individual points x that maximize the expected entropy decrease:

$$\arg\max_{D} H[\theta \mid D] - E_{y \sim p(y|xD)}[H[\theta \mid x, y, D]] \tag{7}$$

BADGE: Batch Active Learning by Diverse Gradient Embeddings BADGE[29] selects diverse sets of points with high magnitude when represented in a hallucinated gradient space, meaning a space containing false or misleading information. This creates a strategy that incorporates predictive uncertainty with sample diversity in the selected batches without the need for hand-tuned hyperparameters.

The algorithm starts with a set M of examples chosen uniformly at random from U, for which the labels are to be determined. The core steps of BADGE, performed iteratively for t steps, consist of gradient embedding and sampling computation. Specifically, for each x belonging to the pool U, the label preferred by the current model is computed, along with the gradient of the loss on x and the computed label, considering the parameters of the last layer of the network. Points are then selected from the obtained gradient embedding vectors using *k-means++* initialization. Finally, the labels of these examples are queried, the model is retrained, and the cycle is repeated.

3.2. Pipeline

The active learning process follows an iterative approach, beginning with a small labeled dataset. In each iteration, new examples are selected and added to the training set, gradually expanding the dataset until it reaches the desired size. To ensure robustness in performance evaluation, each active learning method is tested across multiple independent experiments. The pipeline relies on *BERT* for both text embedding and classification.

3.3. Evaluation metrics

We evaluated the performance of the models with Accuracy, Precision, Recall and F1, furthermore we assessed the environmental and computational efficiency of our models using *execution time* (measured in seconds), *energy consumption* (measured in kWh) and *carbon emissions* (measured in kg), calculated with the *CodeCarbon* [30] library. This permits the quantification of the

sustainability of the training process by estimating the energy usage and the resulting carbon footprint. A need for an index that summarize prediction precision and sustainability is satisfied by the *Efficiency metric*, calculated this way:

$$\text{Efficiency Index} = \frac{Accuracy}{Energy}. \tag{8}$$

These metrics provide a more punctual view of the model performances, balancing effectiveness with ecological impact, computational efficiency and classification accuracy.

4. Experimental evaluation

4.1. Test bed

Datasets To achieve the objective of comparing active learning methodologies across domains, this study leverages two benchmark datasets that represent distinct domains of misinformation. These datasets are diverse in their content and structure, making them ideal for evaluating the adaptability and robustness of different active learning techniques. Below, we provide a detailed description of each of them:

- **Politifact** [31]: is a dataset which includes political information. It comprises 600 labeled instances annotated as true and false. The dataset includes claims made by politicians, public figures and media. Politifact's domain-specific nature makes it an excellent resource for evaluating active learning in handling nuanced and context-heavy political discourse;
- **GossipCop** [32]: focuses on misinformation in the entertainment and celebrity news domain. The dataset is derived from *FakeNewsNet dataset*[33] and contains 4356 instances (not considering duplicates) with balanced representations of both true and false claims. The specificity of this dataset allows for testing how well active learning models perform in detecting misinformation in highly targeted domains.

By utilizing these datasets, this study evaluates the performance of active learning techniques across domains each characterized by challenges and patterns of misinformation.

Hyperparameter and experimental setup The initial labeled dataset consists of 100 instances, with 20 new examples added per iteration until a total of 400 labeled instances is reached. Given this setup, the process completes in 15 iterations. Each method is evaluated through 7 independent experiments. The *BERT* model is trained using a batch size of 16, the ADAM [34] optimizer, a learning rate of 0.002, and 10 training epochs.

4.2. Test results

Comparative analysis The Figure 1 shows the evaluation metrics for the Politifact dataset. The x-axis represents the number of labeled examples while the y-axis represents the value of the corresponding metric. Analyzing the accuracy, it is observed that Entropy Sampling is generally the best strategy, achieving over 90% accuracy with around 250 examples. BADGE is better of Entropy sampling only in the first steps of iteration. For the precision metric, Entropy Sampling and CoreSet are the best strategies, holding above 0.90 at the end of the graph. Looking at Recall, BADGE dominates in most places but Entropy Sampling is equally competitive. At the

end analyzing the F1-score, Entropy Sampling is the most balanced strategy with values above 0.90. BADGE and CoreSet are following with a steady growth. The overall analysis shows that Entropy Sampling and Badge are the best performing strategies in all metrics, while BALD is the least effective.

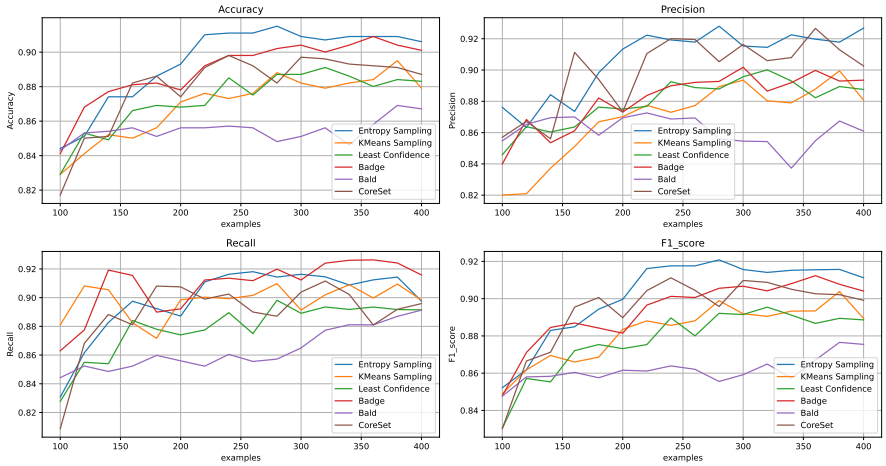

Figure 1: Metrics of Politifact

The Figure 2 shows the evaluation metrics for the GossipCop dataset. The x-axis represents the number of labeled examples while the y-axis represents the value of the corresponding metric. The BALD strategy was trained only for 10 iterations as the training time was excessive. Analyzing the accuracy, it is observed that Entropy Sampling is generally the best strategy, achieving over 74% accuracy with around 350 examples. All model are better of Entropy sampling in the first steps of iteration. For the precision metric, Entropy Sampling is also the best strategies, holding above 0.75. Looking at Recall, BADGE prevails in most cases despite being trained for fewer iterations. In the end analyzing the F1-score, almost all models achieve a similar result with values above 0.76. BADGE and CoreSet are following with a steady growth. The overall analysis shows that Entropy Sampling is the best performing strategy in all metrics, while BALD is the least effective.

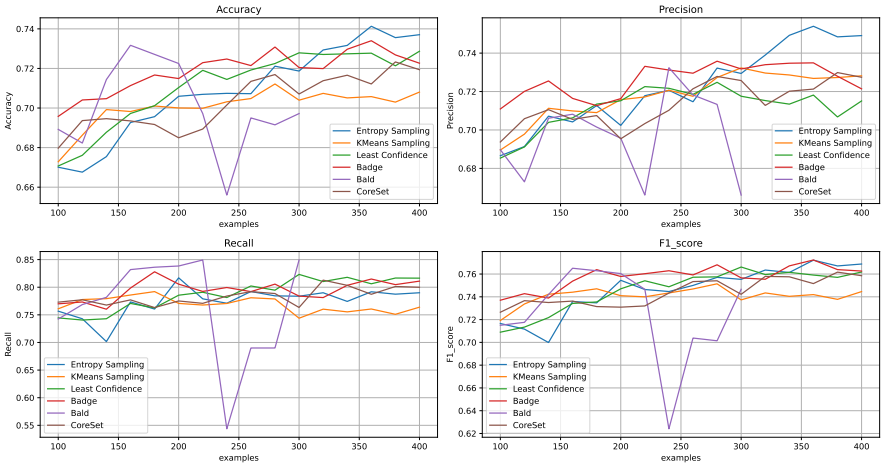

Figure 2: Metrics of GossipCop

Dataset	Technique	Time (seconds)	Energy (Wh)	Emission (g_{CO_2})	Efficiency Index
Politifact	Least confidence	**3993.46**	129.8	54.3	0,687
	Entropy sampling	3999.68	126.0	**28.5**	**0,723**
	CoreSet	4124.58	**125.6**	43.6	0,715
	K-means	4044.94	132.7	31.1	0,675
	BALD	6076.47	189.5	62.6	0,460
	BADGE	4135.18	132.0	55.3	0,689
Gossipcop	Least confidence	6283.23	196.7	76.2	0,371
	Entropy sampling	6068.44	194.1	62.4	0,382
	CoreSet	**5342.77**	**181.5**	52.8	**0,398**
	K-means	6070.90	196.9	**36.9**	0,362
	BALD	15577.84	482.3	206.4	0,152
	BADGE	7594.02	241.1	57.2	0,304

Table 1
Comparative table of the consumptions

In the table 1, all the indices used for this analysis are reported for each dataset. For the Politifact dataset the Entropy sampling method is still the best method both in terms of emissions generated and in terms of efficiency index, it can be seen how the performances in terms of execution time and energy consumed are very close to the best obtained, this confirms that for the Politifact dataset the Entropy Sampling strategy is the best. In the Gossipcop dataset the best strategy in terms of efficiency is Core-Set which achieves the best performances for each index except for emissions generated where the K-means technique is the leader. Unlike the performances for the Politifact dataset where the entropy sampling strategy was the best in the efficiency aspect, with this dataset we find that entropy sampling remains the best in terms of accuracy and other metrics but it is not the best in efficiency as k-means has higher values for each index.

Dataset Impact The choice of initial instances is a very important element to take into consideration when applying Active learning techniques, since they determine both the performance and the results. During the experiments for this study it was noted that the accuracy and other initial metrics had a high variance, this was due to the choice of the first 100 instances, if the instances were able to adequately cover the distributions of the dataset, better performance was achieved compared to other experiments where the performance was lower and more instances were used to reach them:

- **Politifact**: this dataset consists of 600 instances on political statements labeled as true and false. The best AL technique for this dataset was Entropy Sampling that with 280 instances achieved very high performance. To validate the importance of using this technique, the same model used for AL techniques was trained with the entire model divided into training set (480 instances), validation set (60 instances) and test set (60 instances). As can be seen in the table 2, the results achieved by the full model are lower in all metrics, which confirms the importance of using AL techniques that, by reducing the number of training instances, also reduce the execution time, emissions and energy consumption;
- **GossipCop**: this dataset is composed of 4365 instances deals with gossip and entertainment news. In this dataset, unlike the previous one, we have two active learning techniques to consider because with Entropy sampling better performances are achieved while with K-means good performances are obtained but with greater efficiency. To validate the techniques, these two techniques were compared with a model trained with all the instances, composed of 3492 training instances, 436 validation instances and 437 test instances. In this case the results achieved by using AL techniques are slightly lower for the k-means technique and almost identical for the entropy sampling technique. The results obtained are excellent considering the number of training instances. As the use of this technique will significantly reduce both the training time and emissions and energy.

Dataset	Strategy	Nº of instances	Accuracy	Recall	Precision	F1-score
Politifact	None (whole dataset)	480	89.0%	0.87	0.91	0.89
	Entropy Sampling	280	**91.5%**	**0.93**	**0.91**	**0.92**
GossipCop	None (whole dataset)	3492	**75.0%**	**0.75**	**0.80**	**0.78**
	Entropy Sampling	280	74.0%	0.75	0.79	0.77
	K-means	280	71.0%	0.71	0.77	0.75

Table 2
Validation AL techniques on both datasets

Insights From the study of the results obtained, key information can be obtained on the effectiveness and efficiency of AL techniques applied to the classification of fake news:

- The choice of initial instances in the training set influence the final performance of the model. A more balanced training dataset that adequately represents the characteristics of the entire dataset allows to obtain better results with a lower number of iterations;

- The results highlight the importance of the choice of the AL strategy on the final performance. The Entropy sampling technique demonstrated better performance with a reduced number of instances compared to other techniques, demonstrating its ability to select more informative instances. However, for the GossipCop dataset k-means technique achieved a higher overall efficiency, balancing performance and consumption;
- From the results obtained, it is noted that after a certain number of iterations, the addition of new instances within the training set does not lead to significant improvements, this highlights how AL techniques are important in the initial stages of training by reducing the number of instances to be labeled.

5. Discussion and conclusion

The aim of this study was to evaluate and compare different AL techniques for Fake News detection. In order to obtain a robust comparative analysis, two datasets representative of different domains were selected. The experimental analysis highlighted significant divergences between the AL techniques tested. In the Politifact dataset, the Entropy Sampling method resulted both the most effective and the most efficient, reaching an accuracy higher than 90% with about 250 labeled instances and an Efficiency Index of 0,7234. The BADGE strategy obtained good performances in the early stages of training, but lost effectiveness during training without increasing performances. For the GossipCop dataset, Entropy Sampling was confirmed as the most robust method in terms of accuracy, exceeding 74% with 350 labeled examples, while in terms of efficiency the K-means method obtained the best performances in almost all indicators. The results obtained confirm the value of AL in improving the computational efficiency of fake news detection models, reducing the number of input data needed for a good classification. Overall, the Entropy Sampling technique was the most effective for both datasets, but its advantage was more marked in the political domain than in the entertainment one. Furthermore, the sustainability analysis highlighted how alternative strategies, such as K-Means, can offer a better compromise between accuracy and computational efficiency.

Limitations Although several limitations should be noted, this study offers insightful information on the particular topic. First of all, the study's two datasets might not accurately reflect the wide variety of fields and situations in which these methods could be implemented. Therefore, caution is needed when applying our findings to other fields. Additionally, our review does not cover all possible AL methods. While we focus on a selection of the most popular approaches, other methods may yield different or complementary results. At the end, this paper focuses exclusively on the BERT model. The exclusion of some sophisticated models like Roberta [35], Sentence Transformer [36] or other Large Language Models is due to the lack of computational resources.

Future Works Future works could address these limitations by expanding the datasets to cover a larger range of domains, exploring additional Active Learning strategies and integrating a wider array of classification algorithms to assess their efficacy in similar tasks, or could be explored the integration of the neural approach for detecting balance-aware polarized communities, as proposed by Gullo et al. [37], with active learning strategies to enhance fake news detection. This integration could lead to more efficient identification of misinformation by focusing on influential nodes within these communities. Furthermore, we will focus on

developing a framework for fake news detection that is both cross-domain and multimodal. Current approaches primarily analyze textual data within a single domain, but misinformation often spans multiple domains and media types. A promising direction is the integration of graph-based structures to model relationships between different content modalities, such as text, images, and metadata. Graph neural networks and knowledge graphs can help identify hidden connections, enhancing detection capabilities. Incorporating Active Learning into this framework will further optimize data annotation, improving efficiency and scalability. This approach aims to provide a more holistic and adaptable system for misinformation detection, addressing the growing complexity of fake news dissemination.

Acknowledgments

SF, MG, and DM are partly funded by the PRIN MIRFAK project (H53D23008120001). FS is partially supported by research project FAIR (PE00000013) Spoke 9 - Green-aware AI, under the NRRP (National Recovery and Resilience Plan) MUR program funded by the NextGenerationEU. The aforementioned founder had no role in data collection and analysis, decision to publish, or preparation of the manuscript.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] X. Zhang, A. A. Ghorbani, An overview of online fake news: Characterization, detection, and discussion, Information Processing & Management 57 (2020) 102025.

[2] K. Shu, S. Wang, H. Liu, Beyond news contents: The role of social context for fake news detection, in: Proceedings of the twelfth ACM international conference on web search and data mining, 2019, pp. 312–320.

[3] G. Pennycook, A. Bear, E. T. Collins, D. G. Rand, The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings, Management science 66 (2020) 4944–4957.

[4] S. Lewandowsky, U. K. Ecker, C. M. Seifert, N. Schwarz, J. Cook, Misinformation and its correction: Continued influence and successful debiasing, Psychological science in the public interest 13 (2012) 106–131.

[5] L. et al., The science of fake news, Science 359 (2018) 1094–1096.

[6] B. Settles, Active Learning Literature Survey, Technical Report, University of Wisconsin-Madison Department of Computer Sciences, 2009.

[7] S. Flesca, D. Mandaglio, F. Scala, A. Tagarelli, Learning to active learn by gradient variation based on instance importance, in: 2022 26th International Conference on Pattern Recognition (ICPR), 2022, pp. 2224–2230. doi:10.1109/ICPR56361.2022.9956039.

[8] S. Flesca, D. Mandaglio, F. Scala, A. Tagarelli, A meta-active learning approach exploiting instance importance based on learning gradient variation, in: Proceedings of the 31st

Symposium of Advanced Database Systems, Galzingano Terme, Italy, July 2nd to 5th, 2023, volume 3478 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023, pp. 535–544.

[9] S. Flesca, D. Mandaglio, F. Scala, A. Tagarelli, A meta-active learning approach exploiting instance importance, Expert Systems with Applications 247 (2024) 123320. doi:https://doi.org/10.1016/j.eswa.2024.123320.

[10] M. Sahan, V. Smidl, R. Marik, Active learning for text classification and fake news detection, in: Int. Symposium on Computer Science and Intelligent Controls (ISCSIC), IEEE, 2021, pp. 87–94.

[11] L. Martirano, P. Zicari, M. Guarascio, S. F. Pisani, C. Comito, You can spread but you cannot hide: Discovering accurate multi-modal deep fusion models for fake news detection, in: Proc. of the 13th International Conference on Complex Networks and their Applications (CNA), 2025, pp. 108–111.

[12] L. L. Cava, D. Costa, A. Tagarelli, Is contrasting all you need? contrastive learning for the detection and attribution of ai-generated text, in: U. Endriss, F. S. Melo, K. Bach, A. J. B. Diz, J. M. Alonso-Moral, S. Barro, F. Heintz (Eds.), ECAI 2024 - 27th European Conference on Artificial Intelligence, 19-24 October 2024, Santiago de Compostela, Spain - Including 13th Conference on Prestigious Applications of Intelligent Systems (PAIS 2024), volume 392 of *Frontiers in Artificial Intelligence and Applications*, IOS Press, 2024, pp. 3179–3186. URL: https://doi.org/10.3233/FAIA240862. doi:10.3233/FAIA240862.

[13] S. D. Bhattacharjee, A. Talukder, B. V. Balantrapu, Active learning based news veracity detection with feature weighting and deep-shallow fusion, in: 2017 IEEE International Conference on Big Data (Big Data), IEEE, 2017, pp. 556–565.

[14] P. Farinneya, et al., Active learning for rumor identification on social media, in: Findings of the association for computational linguistics: EMNLP, 2021, pp. 4556–65.

[15] H. S. Seung, M. Opper, H. Sompolinsky, Query by committee, in: Proc. of the Fifth Annual Workshop on Computational Learning Theory, COLT '92, ACM, New York, NY, USA, 1992, p. 287–294.

[16] K. Lee, G. Mou, S. Sievert, Energy-based domain adaption with active learning for emerging misinformation detection, in: IEEE Int. Conference on Big Data (Big Data), 2022, pp. 2305–08.

[17] S. Kato, L. Yang, D. Ikeda, Domain bias in fake news datasets consisting of fake and real news pairs, in: 12th Int. Congress on Advanced Applied Informatics (IIAI-AAI), 2022, pp. 101–106.

[18] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: North American Chapter of the Association for Computational Linguistics, 2019. URL: https://api.semanticscholar.org/CorpusID:52967399.

[19] G. Barnabò, et al., Deep active learning for misinformation detection using geometric deep learning, Online Social Networks and Media 33 (2023) 100244.

[20] F. Folino, G. Folino, M. Guarascio, L. Pontieri, P. Zicari, Towards data-and compute-efficient fake-news detection: An approach combining active learning and pre-trained language models, SN Computer Science 5 (2024) 470.

[21] A. Avignone, A. Fiori, S. Chiusano, G. Rizzo, Generation of textual/video descriptions for technological products based on structured data, in: 2023 IEEE 17th International Conference on Application of Information and Communication Technologies (AICT), 2023,

pp. 1–7. doi:10.1109/AICT59525.2023.10313177.
[22] M. Cevallos, M. De Biase, E. Vocaturo, E. Zumpano, Fake news detection on covid 19 tweets via supervised learning approach, in: 2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2022, pp. 2765–2772. doi:10.1109/BIBM55620.2022.9994918.
[23] D. D. Lewis, W. A. Gale, A sequential algorithm for training text classifiers, in: B. W. Croft, C. J. van Rijsbergen (Eds.), SIGIR '94, Springer London, London, 1994, pp. 3–12.
[24] C. E. Shannon, A mathematical theory of communication, The Bell System Technical Journal 27 (1948) 379–423. doi:10.1002/j.1538-7305.1948.tb01338.x.
[25] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: C. Burges, L. Bottou, M. Welling, Z. Ghahramani, K. Weinberger (Eds.), Advances in Neural Information Processing Systems, volume 26, Curran Associates, Inc., 2013. URL: https://proceedings.neurips.cc/paper_files/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf.
[26] Z. Xu, K. Yu, V. Tresp, X. Xu, J. Wang, Representative sampling for text classification using support vector machines, in: Advances in Information Retrieval: 25th European Conference on IR Research, ECIR 2003, Pisa, Italy, April 14–16, 2003. Proceedings 25, Springer, 2003, pp. 393–407.
[27] N. Houlsby, F. Huszár, Z. Ghahramani, M. Lengyel, Bayesian active learning for classification and preference learning, 2011. URL: https://arxiv.org/abs/1112.5745. arXiv:1112.5745.
[28] X. Cao, I. W. Tsang, Bayesian active learning by disagreements: A geometric perspective, 2021. URL: https://arxiv.org/abs/2105.02543. arXiv:2105.02543.
[29] J. T. Ash, C. Zhang, A. Krishnamurthy, J. Langford, A. Agarwal, Deep batch active learning by diverse, uncertain gradient lower bounds, 2020. URL: https://arxiv.org/abs/1906.03671. arXiv:1906.03671.
[30] B. Courty, et al., mlco2/codecarbon: v2.4.1, 2024. URL: https://doi.org/10.5281/zenodo.11171501. doi:10.5281/zenodo.11171501.
[31] N. Vo, K. Lee, Where are the facts? searching for fact-checked information to alleviate the spread of fake news, arXiv preprint arXiv:2010.03159 (2020).
[32] K. Shu, D. Mahudeswaran, S. Wang, D. Lee, H. Liu, Fakenewsnet: A data repository with news content, social context and spatialtemporal information for studying fake news on social media, 2019. URL: https://arxiv.org/abs/1809.01286. arXiv:1809.01286.
[33] K. Shu, G. Zheng, Y. Li, S. Mukherjee, A. H. Awadallah, S. Ruston, H. Liu, Leveraging multi-source weak social supervision for early detection of fake news, arXiv preprint arXiv:2004.01732 (2020).
[34] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, 2014. URL: http://arxiv.org/abs/1412.6980, cite arxiv:1412.6980Comment: Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015.
[35] Y. L. et al., Roberta: A robustly optimized bert pretraining approach, 2019. URL: https://arxiv.org/abs/1907.11692. arXiv:1907.11692.
[36] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, 2019. URL: https://arxiv.org/abs/1908.10084. arXiv:1908.10084.
[37] F. Gullo, D. Mandaglio, A. Tagarelli, Neural discovery of balance-aware polarized communi-

ties, Mach. Learn. 113 (2024) 6611–6644. URL: https://doi.org/10.1007/s10994-024-06581-4. doi:10.1007/s10994-024-06581-4.