Knowledge-Driven Bayesian Uncertainty Quantification for Reliable Fake News Detection

Julia Puczyńska[a,b], **Youcef Djenouri**[c,d], **Michał Bizoń**[a], **Tomasz Michalak**[e,f] **and Piotr Sankowski**[e,f,g]

[a]IDEAS NCBR Sp. z o.o., 69 Chmielna Street, 00-801 Warsaw, Poland
[b]IPPT PAN, Pawińskiego 5B, 02-106 Warsaw, Poland
[c]University of South-Eastern Norway (USN), Post office box 4, 3199 Borre, Norway
[d]Norwegian Research Center (NORCE), Oslo, Norway
[e]Institute of Informatics, Warsaw University, Banacha 2, 02-097, Warsaw, Poland
[f]IDEAS Research Institute, 27 Królewska, 00-060 Warsaw, Poland
[g]MIM Solutions, 47 Świeradowska, 02-662 Warsaw, Poland

Abstract. The pervasive dissemination of fake news presents significant challenges to societal well-being and informed decision-making, necessitating robust detection mechanisms with calibrated uncertainty measures. This paper proposes a novel hybrid framework for fake news detection, integrating uncertainty quantification with a domain-specific Knowledge Base approach. The BANED knowledge base models word-level probabilistic significance, leveraging statistical support metrics to assess prediction uncertainty. By incorporating these metrics into a Bayesian framework, our method provides well-calibrated predictive distributions, offering enhanced interpretability and robustness in the presence of ambiguous or conflicting news data. The proposed approach is evaluated on the FakeNewsNet and ISOT Fake News datasets, demonstrating competitive accuracy and superior reliability compared to state-of-the-art Bayesian inference techniques. Combining word-level probabilistic significance with Monte Carlo Dropout decreases mean calibration error and narrows the interquartile range of predictions. Full code and supplementary materials of BANED might be found at https://github.com/micbizon/BANED.

1 Introduction

Detecting fake news[1] is increasingly critical due to its detrimental impact on democratic institutions, public health, and safety [32, 28]. The recent surge in automated content generation — amplified by large language models — has only intensified this challenge, making traditional verification methods like manual fact-checking inadequate for the scale and speed required. To address this, there is a growing need for automatic detection systems. However, without explainability and accountability, automatically marking and limiting visibility of online content infringes on free speech protections. For such systems to be trustworthy in high-stakes contexts, their predictions must be accompanied by rigorous *uncertainty quantification* ($\mathcal{UQ}$).

$\mathcal{UQ}$ enables models to express confidence in their outputs, supporting statistical inference and improving decision-making under uncertainty [1]. It is crucial for enhancing the reliability and robustness of fake news detection models [29, 37, 10]. By quantifying uncertainty, detection systems can provide more nuanced and transparent decisions, boosting user trust and enabling better risk management [24]. This approach not only strengthens the credibility of automated detection but also facilitates more informed and cautious decision-making in the context of fake news.

Among the leading approaches to $\mathcal{UQ}$ are methods based on Bayesian inference, which offer principled ways to manage uncertainty and address the limitations of overconfident or brittle deep learning models [30]. These methods may prove to be the solution to many of current challenges of deep learning [25], such as the problem of unreliable fake news detection models. In this work, we evaluate the effectiveness of techniques like Monte Carlo Dropout (MC) and introduce our own, dropout-based approach. MC Dropout is a practical method to approximate Bayesian inference, that leverages the principles of random sampling and statistical modeling, at the same time providing scalability and applicability to basically any network structure [39, 4, 38]. In contexts like fake news detection, MC simulations can help quantify the uncertainty inherent in content classification, allowing for more robust and informed decision-making. They offer a nuanced understanding of the risks and probabilities associated with different outcomes, making them invaluable for addressing complex, uncertain systems [8].

In each forward pass, Monte Carlo (MC) Dropout randomly disables a subset of neurons (See Figure 1 for more details). This returns a distribution of results that can be subject to statistical reasoning. By including dropout layers in a network's architecture and running data through the model multiple times it simulates a wide range of possible outcomes, providing a comprehensive picture of potential scenarios and their likelihoods.

However, MC Dropout has notable limitations. As an approximation to Bayesian inference, its estimates depend heavily on dropout rates and may suffer from poor calibration—where predicted probabilities diverge from actual likelihoods. Furthermore, its reliance on stochastic behavior in internal layers makes the process opaque and difficult to explain. This lack of interpretability limits its potential for building trust and ensuring accountability in automated decision-making.

[1] We recognize "disinformation" is a broader and more appropriate term in this context than "fake news" [15, 14]. We use it here for simplicity and in hopes to reach a wider audience, since it is much more popular.

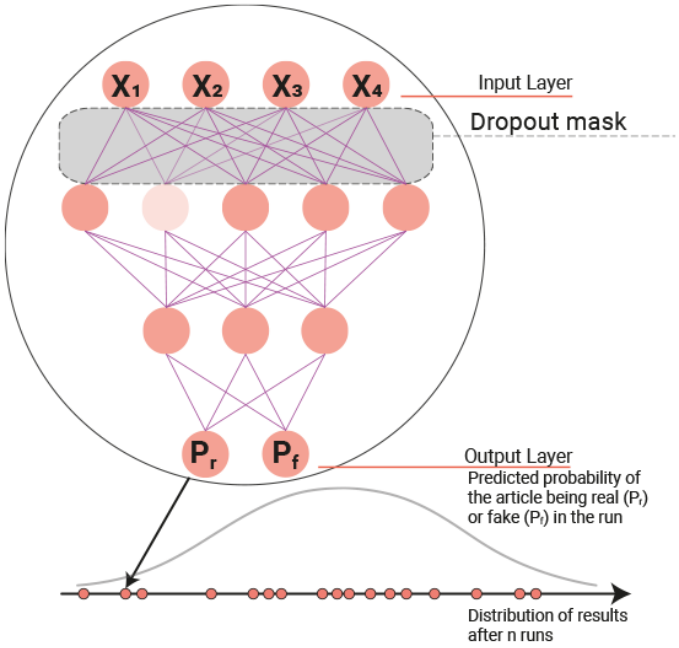

Figure 1. Monte Carlo Dropout

To address these shortcomings, we propose a novel approach to $\mathcal{UQ}$ in fake news detection by combining MC Dropout with word-level probabilistic significance. Probabilistic significance identifies word patterns that may be crucial for discerning deceptive content [35, 17, 6], offering a richer representation of contextual cues and potentially reducing overconfidence in model predictions [21, 26]. In our approach, these significant patterns are used to construct a *knowledge base*, which informs the dropout process. We hypothesize that the extra knowledge contained in this knowledge base will complement the statistical nature of MC. A hybrid approach leveraging both Monte Carlo and the knowledge base can thus lead to more robust and reliable uncertainty quantification by addressing the individual limitations of Monte Carlo and providing a comprehensive understanding of the data.

Typically, uncertainty quantification is applied to problems that require a full scale of predictions, e.g., for molecular property prediction we need to know the probability of the molecule being non-reactive from 0% to 100% [30]. In fake news detection we may intend to separate some of the data to weed out news that have a high probability of being true, in order to analyse only the news that may be fake. To this end, we introduce *adapted calibration error* that shows how well a chosen method is doing in the scope that may be of interest for classification.

In summary, our contributions are as follows:

1. We develop Knowledge Base[2] Dropout (KB Dropout) and Knowledge Base Monte Carlo Dropout (KBMC Dropout) that combine knowledge base built on word-level probabilistic significance with Monte Carlo Dropout specifically for uncertainty quantification in fake news detection. We show that our hybrid approach guides the MC process in fake news detection more effectively through identification of trends within the datasets, which can be further explored along with other $\mathcal{UQ}$ methods.
2. We modify existing evaluation metric, i.e., calibration, specifically the Mean Absolute Calibration Error, so it is better suited for the purposes of certain aspects of fake news classification.
3. To facilitate broader applicability of our approach, we develop an open-source knowledge base BANED (knowledge Base for fAke NEws Detection) by extracting a collection of frequent patterns from both fake and real news data from FakeNewsNet.
4. Our evaluation demonstrates the effectiveness of combining word-level probabilistic significance with MC. In particular, combined methods can achieve two times lower mean calibration error and narrower scope of 50% lowest standard deviations of predictions (0.12 compared to 0.14). Knowledge Base Dropout approach by itself reaches values of sharpness much lower than MC. These results suggest that applying word-level probabilistic significance-based dropout to $\mathcal{UQ}$ may be beneficial in order to balance calibration error and sharpness and adjust existing $\mathcal{UQ}$ methods for the purposes of fake news detection.

It is important to note that our word-level significance should be supported by other markers and explainable models to create a just fake news detection system.

2 Related Work

Uncertainty Quantification The rapid progress in machine learning and deep learning has greatly benefited various fields, but the unpredictability of these models often limits real-world deployment. While accuracy and precision measure model performance on test data, $\mathcal{UQ}$ assesses how much trust can be placed in individual predictions, especially on new data [16, 11, 41]. Uncertainty arises from data variability (aleatoric uncertainty) [19, 40] or limitations (epistemic uncertainty) [34, 3]. Since ML models rely on data quality, $\mathcal{UQ}$ helps gauge confidence for each prediction, aiding decision-making in areas like urban mobility [27], drug discovery [18], text classification [38], and fake news detection [2]. Despite the complexity of deep learning models, effective $\mathcal{UQ}$ tools remain limited. Existing approaches primarily include ensemble methods [36], Bayesian neural networks [13], and conformal prediction [4].

Fake News Detection Detecting fake news presents a significant challenge due to both the linguistic and visual nature of disinformation [23]. This complexity necessitates the use of multimodal models, particularly to capture nuanced elements like sarcasm or subtle image-text inconsistencies [20]. Disinformation is deliberately crafted to resemble legitimate news, further complicating detection efforts. As a result, many detection systems rely on complex, often black-box models, which raise critical concerns around social bias, censorship, fairness, and safety [22, 5, 9, 33]. Errors in detection can have serious consequences; either allowing harmful misinformation to spread or unjustly censoring valid content, potentially leading to public distrust and backlash. Addressing these issues is crucial to advancing reliable, transparent, and socially responsible fake news detection systems.

Discussion We aim to develop a $\mathcal{UQ}$ method adaptable to various deep learning models, capable of evolving with the changing fake news landscape. Our goal is to make it simple and accessible for fact-checkers, leading us to enhance Monte Carlo Dropout with an informed version—Knowledge Base Dropout (KB Dropout).

[2] While not symbolic or logic-based in the classical AI sense, our use of the term "knowledge base" is inspired by structured, empirically derived information (in this case, statistically supported linguistic patterns), chosen to emphasize that our word-level frequencies are extracted from real-world data and meant to encode domain knowledge in a reusable way.

3 Methodology

The KB Dropout approach emulates human evaluation by leveraging prior knowledge to assess the reliability of information. KB is constructed to capture word-level prediction uncertainty and contextual significance using probabilistic metrics. The trained fake news classifier is tested iteratively, where words are probabilistically dropped during each test pass based on their uncertainty and significance, as derived from the KB.

3.1 Knowledge Base Creation

Let $\mathcal{V} = \{w_1, w_2, \ldots, w_m\}$ denote the vocabulary of the dataset. Each document t_i undergoes preprocessing to produce a set of words D_i. The complete dataset is represented as $\mathcal{D} = \{D_1, D_2, \ldots, D_n\}$.

Word-Level Probability The probability of a word $w \in \mathcal{V}$ appearing in the dataset is computed as:

$$P(w) = \frac{\sum_{i=1}^{n} \mathbb{I}(w \in D_i)}{\sum_{i=1}^{n} |D_i|} \tag{1}$$

where $\mathbb{I}(w \in D_i)$ is the indicator function, equal to 1 if w appears in D_i, and 0 otherwise. $|D_i|$ is the total number of words in document D_i.

Prediction Uncertainty For each word $w \in \mathcal{V}$, its uncertainty score is defined as:

$$U(w) = 1 - |P_{\text{real}}(w) - P_{\text{fake}}(w)| \tag{2}$$

where $P_{\text{real}}(w)$ and $P_{\text{fake}}(w)$ are computed as follows:

$$P_{\text{real}}(w) = \frac{\text{Count}_{\text{real}}(w)}{N_{\text{real}}}, \tag{3}$$

and,

$$P_{\text{fake}}(w) = \frac{\text{Count}_{\text{fake}}(w)}{N_{\text{fake}}} \tag{4}$$

$\text{Count}_{\text{real}}(w)$ and $\text{Count}_{\text{fake}}(w)$ represent the total occurrences of w in real and fake news datasets, respectively. $N_{\text{real}} = \sum_{w \in \mathcal{V}} \text{Count}_{\text{real}}(w)$ is the total number of words in all real news documents. $N_{\text{fake}} = \sum_{w \in \mathcal{V}} \text{Count}_{\text{fake}}(w)$ is the total number of words in all fake news documents.

Contextual Co-occurrence To model the contextual impact of word pairs, we compute the joint probability of two words $w_i, w_j \in \mathcal{V}$ co-occurring in the dataset:

$$P_{i,j} = \frac{\text{CoCount}(w_i, w_j)}{N_{\text{co}}} \tag{5}$$

where,

$$CoCount(w_i, w_j) = \sum_{k=1}^{n} \mathbb{I}(w_i \in D_k \wedge w_j \in D_k) \tag{6}$$

$CoCount(w_i, w_j)$ is the number of documents where w_i and w_j co-occur. $N_{\text{co}} = \sum_{w_p, w_q \in \mathcal{V}} \text{CoCount}(w_p, w_q)$ is the total number of word-pair co-occurrences across all documents.

3.2 Knowledge Base Monte Carlo Dropout

Dropout is a technique that can help mitigate overfitting during training as well as quantify uncertainty. Conventional dropout is typically applied to hidden layers and has a constant rate. Our proposed Knowledge-Base approach leverages prior knowledge to inform the dropout decisions. Figure 2 sketches the three stages of our framework: calculating word-level probabilistic significance, retrieval and uncertainty quantification.

3.2.1 Dropout Algorithm

Let us consider an input sequence $x = x_1, x_2, ..., x_s$, where x_i represents an individual word. In the proposed directional dropout approach, we apply a selective, data-driven dropout mechanism guided by word frequency information from a vocabulary $\mathcal{V}$; if the word x_i from the input sequence x is found in vocabulary $\mathcal{V}$, we draw a random number r_i between 0 and 1. If r_i is smaller than the frequency of x_i in the dataset, x_i is dropped from the input sequence. In effect, this creates a dropout scheme that is inversely proportional to the word's informativeness: common words are more likely to be dropped, while rare or distinctive words are retained. Because the time complexity of this task is $O(n * m)$ (compared to linear MC Dropout), we tested it with subsets of the knowledge base, including only the words with a frequency higher than a certain threshold (from 0.1 to 0.5).

3.2.2 Uncertainty Quantification Calculation

We then move on to quantifying the model's uncertainty. To do that we introduce dropout layers into the model, which turns off certain neurons in the network with a certain probability. Then we run the model multiple times to create a distribution of predictions for test data. We calculate the $\mathcal{UQ}$ score for each item. The $\mathcal{UQ}$ score for an individual item from the test set I_i is calculated by averaging the predictions $P_{i,j}$ for this item from each run of the model:

$$\mathcal{UQ}_{I_i} = \frac{1}{k} \sum_{j=1}^{k} P_{i,j}, \tag{7}$$

where k responds to how many times we have run the model, either with regular dropout layers or with our informed dropout mask.
In standard Monte Carlo Dropout (MC), dropout layers are inserted into the model with a fixed rate, randomly deactivating neurons during inference to simulate variability. However, this approach is agnostic to the semantic content of the input and the contextual relevance of individual features. With KB Dropout we drop out neurons from the input layer, i.e., words from the analysed article. These words are dropped out with the probability equal to the probability of the corresponding patterns from the knowledge base (both words and sets of words). We evaluate a combined KBMC Dropout approach, where KB and MC approaches are used together with varied dropout (α) and minimum support (μ) values.

Sharpness and scope The sharpness score reflects the decisiveness of the method,where the lower the score, the more stable the results. It can be defined as:

$$\kappa_{avg} = \frac{1}{n} \sum_{i=1}^{n} \frac{\sum_{j=1}^{k} |P_{i,j} - \bar{P}_i|}{k}, \tag{8}$$

where n is the number of samples, k responds to how many times we have run the model, $P_{i,j}$ is the prediction of the model for the i-th

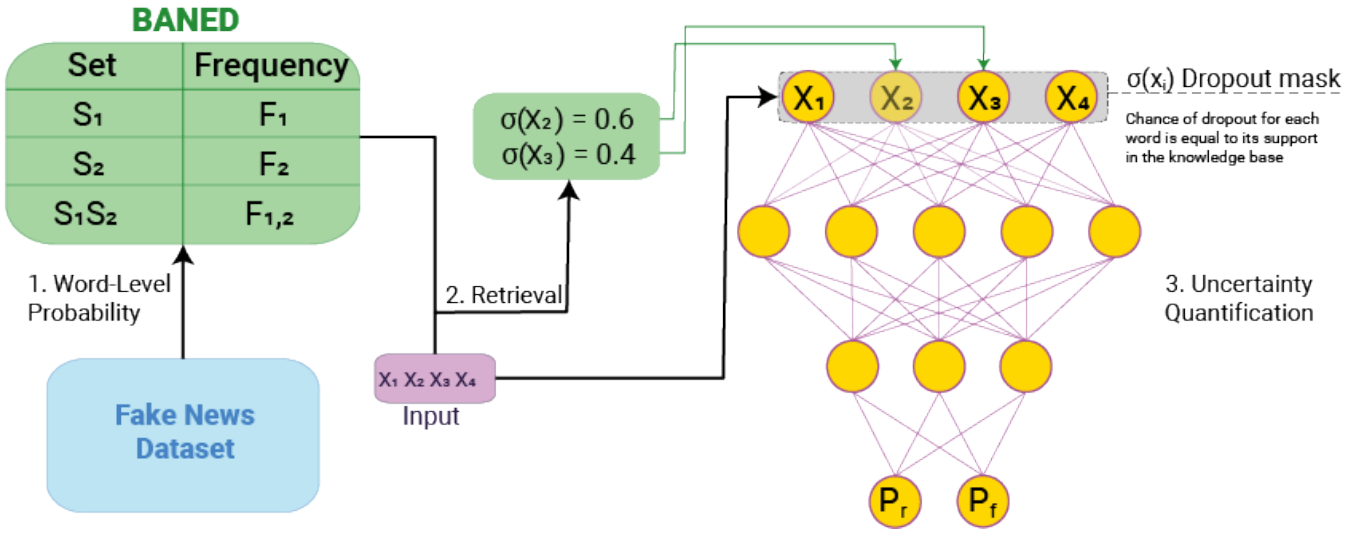

Figure 2. KB Dropout Principle

sample in the j-th run and $\bar{P}_i$ is the mean average of predictions for the i-th sample.

Additionally, we report an additional measure for sharpness, which is the scope encompassing 50% of predictions with the lowest standard deviation, providing insight into the most confident model outputs:

$$\kappa_{50\%} = \frac{1}{n}\sum_{i=1}^{n}\sigma_{i,d}, \tag{9}$$

where n is the number of samples and $\sigma_{i,d}$ is the maximum of the lowest 50% of standard deviation values of predictions for the i-th element.

Extended Calibration Error for Fake News Detection $\mathcal{UQ}$ is important in plenty of different areas and all of these have their own quirks. Probabilistic forecasts that deal with continuous variables require appropriate tools for evaluation of different $\mathcal{UQ}$ methods. The evaluated aspects are usually sharpness, which is connected to the variation of predictions for specific items, and calibration, which corresponds to the difference between assessed probabilities of predictions and actual frequency of labels.

Calibration can be assessed through the mean absolute calibration error.

Definition 1. *The mean absolute calibration error for uncertainty quantification is defined as:*

$$\gamma = \sum_{m=1}^{M}\frac{|B_m|}{n}\left|\frac{\sum_{i\in B_m}\mathbf{1}_{\{y_i=\hat{y}_i\}} - \sum_{i\in B_m}\hat{p}_i}{|B_m|}\right|, \tag{10}$$

where M is the number of bins, B_m is the set of samples in the m^{th} bin, n is the total number of samples, y_i is the true label for sample i, $\hat{y}_i$ is the predicted label for sample i, and $\hat{p}_i$ is the predicted probability for sample i.

However, for fake news detection, a model's prediction is often most useful above a certain threshold. This is because, in the decision-making process of, e.g., a social media platform administrator, only probability values higher than 0.5 allow for certain actions, such as content deletion or account banning. Therefore, in this paper, we introduce an adapted version of calibration that can help evaluate the model's predictions for the values of greatest interest.

Definition 2. *We define the adapted mean absolute calibration error designed for fake news detection as:*

$$\gamma_\theta = \sum_{m=1}^{M^\theta}\frac{|B_m^\theta|}{n^\theta}\left|\frac{\sum_{i\in B_m^\theta}\mathbf{1}_{\{y_i=\hat{y}_i\}} - \sum_{i\in B_m^\theta}\hat{p}_i}{|B_m^\theta|}\right|, \tag{11}$$

where θ is the uncertainty threshold ranged from 0 to 1, M^θ is the number of bins where their predicted probability exceed θ, B_m^θ is the set of samples in the m^{th} bin where its predicted probability exceeds θ, and n^θ is the total number of samples where their predicted probability exceed θ.

4 Theoretical Analysis

Proposition 1. *The adaptive dropout rate $p(w)$ for a word w based on its probability $P(w)$ from the knowledge base is given by:*

$$p(w) = \frac{P(w)}{\sum_{v\in\mathcal{V}}P(v)}, \tag{12}$$

where $\mathcal{V}$ is the vocabulary set containing all the words in the dataset.

Proof. The probability $P(w)$ represents the frequency or occurrence of word w in the context of the dataset. To normalize the dropout probability $p(w)$, we need to ensure that the sum of the dropout probabilities across all words in the vocabulary is 1. EQ. 12 ensures that:

$$\sum_{w\in\mathcal{V}}p(w) = \sum_{w\in\mathcal{V}}\frac{P(w)}{\sum_{v\in\mathcal{V}}P(v)} = 1. \tag{13}$$

Hence, $p(w)$ is a valid probability distribution over the vocabulary $\mathcal{V}$. □

Proposition 2. *The $\mathcal{UQ}$ score, calculated by averaging predictions over multiple runs with adaptive dropout, is a consistent, empirical approximation of the model's uncertainty for a large number of runs k. Its reliability improves with an increasing k, under the assumption that the dropout-induced variability reflects epistemic uncertainty.*

Proof. Let $P_{i,j}$ be the prediction for item I_i in the j-th run of the model, where the input vector $\boldsymbol{X}$ is modified by an adaptive dropout mask $\boldsymbol{M}_j$ such that:

$$M_{i,j} = \begin{cases} 1 & \text{with probability } 1 - p(w_i), \\ 0 & \text{with probability } p(w_i). \end{cases}$$

The modified input vector for the j-th run is:

$$\boldsymbol{X}_j = \boldsymbol{X} \odot \boldsymbol{M}_j, \tag{14}$$

where $\odot$ denotes element-wise multiplication. The prediction for the j-th run is $P_{i,j} = f(\boldsymbol{X}_j)$, where $f(\cdot)$ represents the model's prediction function. By replacing this value of $P_{i,j}$ in EQ. 7, and by using the law of large numbers [12], as k approaches infinity, the sample mean converges to the expected value:

$$\lim_{k \to \infty} \frac{1}{k} \sum_{j=1}^{k} P_{i,j} = \mathbb{E}[P_{i,j}]. \tag{15}$$

Thus, for a sufficiently large number of runs k, UQ_{I_i} is a consistent estimate of the true model uncertainty. □

5 Performance Evaluation

Datasets and Experimental Settings We evaluate the designed method using the FakeNewsNet (FNN) [31] and ISOT Fake News datasets[3]. The FNN dataset includes a total of 23,190 articles, and the ISOT Fake News Dataset contains 44,899 articles mostly from 2016 and 2017. To ensure consistency and reduce noise, we applied a multi-step preprocessing pipeline to the title and body of each news article. Contractions (e.g., "don't", "it's") were expanded using the contractions library to improve tokenization and semantic understanding. Non-alphabetic characters were removed, keeping only English letters and spaces to eliminate numbers, punctuation, and special characters. We used NLTK's word_tokenize and stopword list to tokenize each text and remove common English stopwords (e.g., "the", "and", "is"), keeping only informative words. The remaining words were stemmed using the Porter Stemmer and converted to lowercase. Duplicates were removed by converting the word list to a set. For missing titles or text (marked as "nan"), we filled the title with the first sentence of the text or vice versa. The processed data was saved as a CSV, preserving the original structure with normalized, tokenized fields. To facilitate our evaluation, we randomly split the dataset into 80% training and 20% test sets. This ensures the model learns from a substantial portion of the data while the test set provides an unbiased benchmark to assess generalization and robustness in detecting fake news.

All experiments were conducted on a Nvidia DGX server with one A100 80GB GPU, an AMD EPYC CPU (2×2.25 GHz cores with two threads each), and 128 GB RAM. We used a fully connected neural network in Keras for binary classification of fake vs. real news. The model includes: a flat input layer, initial dropout based on knowledge base support, a dense layer with 64 ReLU units, Monte Carlo dropout, another dense layer with 32 ReLU units, and a final dropout layer before a sigmoid-activated output neuron that predicts the probability of an article being fake. The model was compiled using the `Adam` optimizer and trained with `binary cross-entropy` loss, a standard choice for binary classification problems. The performance metric was `accuracy`, training ended after five epochs when accuracy reached around 99%. The simplicity of the model was introduced in order to simulate the direst of setups (which is often the cast in fact-checking reality), so we could compare the $\mathcal{UQ}$ methods and not the classification model itself. We evaluate the chosen methods using well-known $\mathcal{UQ}$ metrics, such as calibration (γ) and sharpness (κ), and visualize their results [7].

[3] https://www.kaggle.com/datasets/csmalarkodi/isot-fake-news-dataset

Comparison with State-of-the-art Existing Bayesian inference-based methods of quantifying uncertainty that are both scalable and independent from the underlying model include MC Dropout, bootstrapping and Deep Ensembles. Though the other two outperform MC [30], we decided to include it due to its smaller computational requirements, compared to Deep Ensembles.

To evaluate the effectiveness of the chosen uncertainty quantification methods, we matched the computed uncertainty scores with the ground truth labels. We calculated their mean absolute calibration error, mean sharpness value, represented by standard deviations of predictions, the scope of 50% lowest standard deviation values and the adapted calibration error with 50, 75 and 90% thresholds.

5.1 *FakeNewsNet Dataset*

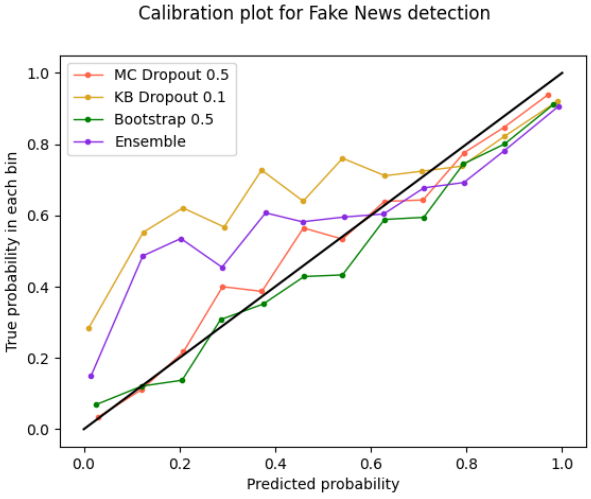

Figure 3. Calibration for base methods with FNN Dataset

As seen in Figure 3, KB Dropout and ensemble are overly uncertain for values 0.0 - 0.6 of predicted probability, but all of the base methods performance increases when reaching higher values. The methods combined with ensemble seem to be worst calibrated for lower values of predictions compared to other KB-related technique (as shown in Figure 4). The average calibration is best for bootstrapping joined with knowledge base, however from the 0.75 threshold MC and KBMC Dropout clearly outperform all the other methods (as shown in Table 2).

Figures 6 and 7 visualize the distribution of predictions. High concentration of values close to 0 showcases a repeatability of predictions. KB and ensemble have the lowest scope of 50% standard deviations, and the methods that include MC Dropout have the best average sharpness, indicating more confidence in the predictions. A trade-off between calibration and sharpness is typical in $\mathcal{UQ}$ methods, where improved calibration may come at the cost of increased prediction uncertainty. Thus, in figure 5, methods with lowest calibration error have the highest sharpness and vice versa. Ensemble

Table 1. Comparison between Monte Carlo Dropout, Knowledge Base Dropout and Knowledge Base Monte Carlo. Chosen metrics are represented by γ for calibration and κ for sharpness. The μ and α values were chosen based on the ablation study.

Data	$\mathcal{UQ}$	γ_{avg}	$\gamma_{0.5}$	$\gamma_{0.75}$	$\gamma_{0.9}$	κ_{avg}	$\kappa_{50\%}$
FNN	KB	0.333	0.247	0.147	0.063	**0.069**	**0.003**
	MC	0.135	0.096	**0.025**	**0.001**	0.115	0.077
	Bp	**0.087**	**0.078**	0.103	0.028	0.169	0.086
	En	0.113	0.103	0.098	0.050	0.130	**0.010**
	$KB_{0.4}MC_{0.4}$	0.165	0.126	**0.050**	**0.001**	**0.101**	0.063
	$KB_{0.1}MC_{0.5}$	0.213	0.162	0.066	**0.001**	0.126	0.100
	$Bp_{0.5}KB_{0.1}$	**0.083**	**0.077**	0.106	0.033	0.182	0.182
	$KB_{0.1}En$	0.092	0.104	0.110	0.027	0.187	0.187
	$KB_{0.2}En$	0.090	0.080	0.096	0.027	0.160	0.160
ISOT	KB	**0.118**	**0.038**	**0.083**	0.221	0.076	1.818e-06
	MC	0.356	0.316	0.174	0.029	0.039	4.420e-04
	Bp	0.291	0.256	0.173	0.032	0.100	0.009
	En	0.299	0.400	0.323	0.475	**0.023**	**1.096e-07**
	$KB_{0.4}MC_{0.4}$	0.300	0.249	0.176	0.060	0.105	2.065e-04
	$KB_{0.1}MC_{0.5}$	0.229	0.195	0.159	**0.014**	0.150	0.005
	$Bp_{0.5}KB_{0.1}$	0.240	0.179	0.159	0.029	0.143	0.021
	$KB_{0.1}En$	0.280	0.298	0.213	0.163	0.130	4.245e-05
	$KB_{0.2}En$	0.281	0.305	0.222	0.186	0.135	7.993e-06

Table 2. Comparison of execution time (in miliseconds) of the chosen $\mathcal{UQ}$ methods for both datasets

Dataset	$KB_{0.1}$	$MC_{0.5}$	$KB_{0.4}MC_{0.4}$	$KB_{0.1}MC_{0.5}$	$Bp_{0.5}$	En	$Bp_{0.5}KB_{0.1}$	$KB_{0.1}En$	$KB_{0.2}En$
FakeNewsNet	1091,619	731,814	720,333	1005,603	739,770	-	985,706	-	-
ISOT	2694,679	2111,435	2170,936	2455,887	2162,841	-	2483,342	-	-

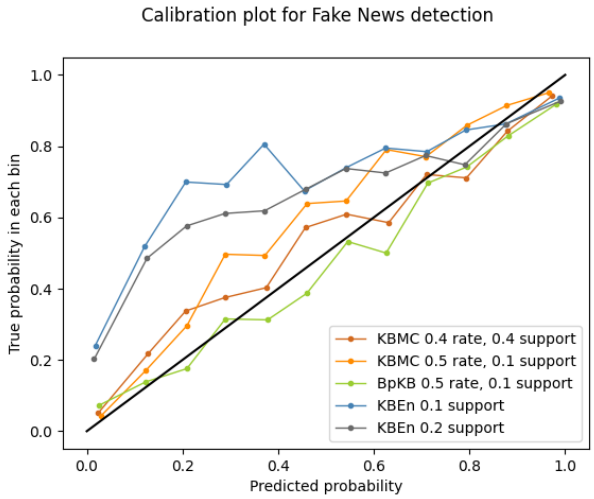

Figure 4. Calibration for combined methods with FNN Dataset

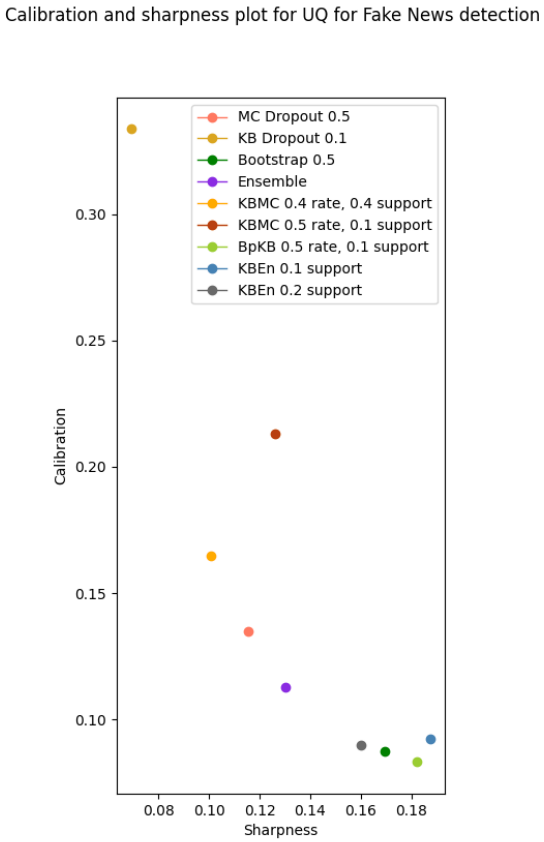

Figure 5. Comparison of all of the chosen methods on the basis of the average calibration and sharpness scores

yields the best results in general, with the lowest sum of sharpness and calibration, but MC Dropout is very close behind it, with even better results with higher thresholds. The differences between the other methods are not that obvious, as most of them are better than others depending on the metric.

5.2 ISOT Fake News Dataset

We decided to use another dataset in order to verify our results; especially those of MC Dropout, proven to be worse than bootstrap and deep ensembles in other studies. We trained the models on 80% of the data and tested it on the other 20% - however, as we mentioned, this dataset's proportion between real and fake news is 1:1. Knowledge Base Dropout has the best calibration results of all the methods, including thresholds 0.5 and 0.75 - at 0.9 it is actually second worst. KBMC 0.4 0.4 has the lowest calibration error at 0.9 and second lowest on average, BpKB is second best at all thresholds. Ensemble, however, is best in terms of sharpness. When comparing all methods, KB Dropout has the lowest sum of both metrics and the lowest square root of sum of squares of each metric's value. Second best are either ensemble, which has the lowest sum, or KBMC 0.4 0.4, with the lowest square root of the sum of squares of its metric's values. In conclusion, the KB Dropout approach allows for adjustment of re-

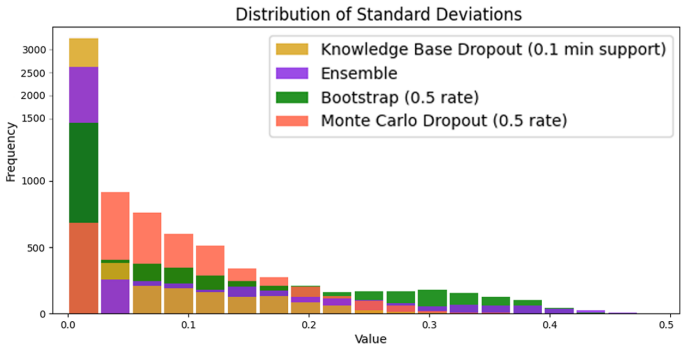

Figure 6. Standard deviations of predictions for base methods (bars overlap)

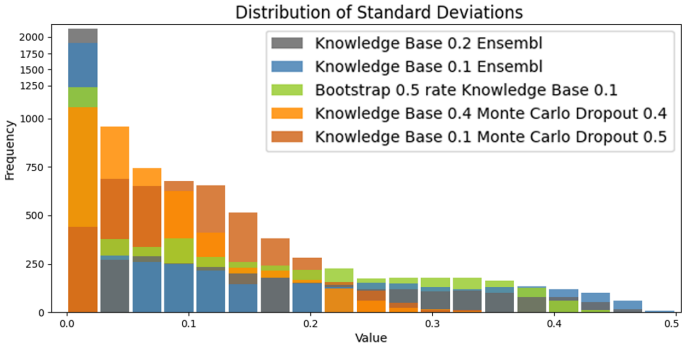

Figure 7. Standard deviations of predictions for combined methods (bars overlap)

quired sharpness and calibration of the model, with non-random, explainable predictions. The computational cost of running the Knowledge Base can be reduced with higher μ values. It is possible that dropout of patterns from the database helps separate real news from the ambiguous and fake ones due to the model being able to make a prediction based solely on other, more important clues—it would be beneficial to investigate what these might be. For now these word patterns seem to be crucial for the model's certainty when dealing with more ambiguous news, since it seems to be too uncertain towards some articles with the KB Dropout. The increased elimination of word patterns heavily supports sharpness or confidence of predictions. This is probably because the reduced chance of dropout leads to similar runs through the model.

Computational costs An important question arises: aren't the methods we developed too computationally expensive? We measured inference time per article, and among the tested approaches, the results are fairly consistent. MC Dropout 0.5, Bootstrapping, and KBMC Dropout 0.4 0.4, using the FNN-based classifier, showed the best performance cost-wise, with runtimes of approximately 730 ms per article. The entire process is significantly longer when we use the ISOT dataset, which is due to its articles being much longer. Ensemble methods were not included in this analysis, due to their different proportion between training and inference time compared to other $\mathcal{UQ}$ methods (ensemble being massively training-heavy).

Ablation Study for FakeNewsNet In these experiments we varied the dropout rate α and minimum support μ values, using the FakeNewsNet-trained models and testing data. We applied α values

Table 3. Effect of dropout rate α and minimum support μ values on the Monte Carlo Dropout, Knowledge Base Monte Carlo and the two of them

α	μ	γ_{avg}	$\gamma_{0.5}$	$\gamma_{0.75}$	$\gamma_{0.9}$	κ_{avg}	$\kappa_{0.5}$
0.0	0.1	0.333	0.247	0.147	0.063	0.069	0.003
	0.2	0.386	0.272	0.154	0.060	0.052	0.001
	0.3	0.381	0.244	0.149	0.085	0.041	1.461e-04
	0.4	0.441	0.262	0.150	0.097	0.020	1.062e-05
0.2	0.1	0.098	0.104	0.118	0.063	0.160	0.053
	0.2	0.089	0.080	0.104	0.036	0.152	0.049
	0.3	0.097	0.090	0.105	0.036	0.149	0.042
	0.4	0.100	0.091	0.107	0.037	0.148	0.039
0.3	0.1	0.078	0.079	0.107	0.032	0.193	0.114
	0.2	0.067	0.059	0.095	0.021	0.185	0.111
	0.3	0.074	0.062	0.095	0.020	0.178	0.101
	0.4	0.076	0.064	0.096	0.019	0.177	0.098
0.4	0.1	0.063	0.056	0.094	0.022	0.197	0.135
	0.2	0.050	0.043	0.077	0.016	0.188	0.131
	0.3	0.052	0.044	0.074	0.014	0.184	0.126
	0.4	0.053	0.045	0.075	0.014	0.182	0.122
0.5	0.1	0.039	0.020	0.051	0.016	0.194	0.156
	0.2	0.076	0.048	0.023	0.012	0.187	0.156
	0.3	0.079	0.050	0.018	0.009	0.185	0.153
	0.4	0.076	0.049	0.020	0.009	0.184	0.149

of 0.2, 0.3, 0.4, and 0.5, and μ values of 0.1, 0.2, 0.3 and 0.4. The results, summarized in Table 3, indicate that as α in MC increases, the mean absolute calibration error generally decreases and the sharpness slightly increases, indicating once again a trade-off between calibration and sharpness. For μ for Knowledge Base Monte Carlo it seems to be the opposite, with higher values leading to more conservative and less calibrated results. The combined approach yields balanced results, with low calibration error and moderate sharpness. It is worth noting that in this approach sharpness is consistently increasing with μ. Calibration error on the other hand reaches its lowest values for μ=0.2 with 0.2-0.4 dropout rates.

Fake news are going to be increasingly automatically generated in bulk with language models. Therefore, methods of fake news detection and associated with them $\mathcal{UQ}$ methods need to be specially adjusted, in order to reduce manual workload and to hopefully stay ahead of the dangers of disinformation. Introducing explainability through informed computational methods, such as the KB approach, can introduce accountability to this crucial, deeply misunderstood and often mistrusted area of ML applications.

6 Conclusion

This paper introduces a novel disinformation detection method that combines word-level probabilistic dropout with uncertainty quantification through a technique called Knowledge Base (KB) Dropout. Using the Apriori algorithm, frequent word patterns are stored in a knowledge base, which helps assess model uncertainty. KB Dropout uses word support from these patterns to produce calibrated predictions, improving both detection accuracy and uncertainty estimation. Results show that KB and KBMC Dropout outperform traditional Monte Carlo methods in sharpness and calibration. The model's confident predictions after dropping high-support words suggest it captures content beyond general news language. While the proposed method is model-agnostic, current results are based on a simple fully connected layer network; future work will explore its integration with more advanced architectures such as transformers and evidential deep learning. Additionally, further validation will be conducted using diverse and synthetic datasets, with efforts directed toward refining the BANED knowledge base and optimizing the pattern mining process to support broader and more robust multimodal news analysis.

References

[1] M. Abdar, F. Pourpanah, S. Hussain, D. Rezazadegan, L. Liu, M. Ghavamzadeh, P. Fieguth, X. Cao, A. Khosravi, U. R. Acharya, et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information fusion*, 76:243–297, 2021.

[2] N. Ayoobi, S. Shahriar, and A. Mukherjee. Seeing through ai's lens: Enhancing human skepticism towards llm-generated fake news. In *Proceedings of the 35th ACM Conference on Hypertext and Social Media*, pages 1–11, 2024.

[3] V. Bengs, E. Hüllermeier, and W. Waegeman. On second-order scoring rules for epistemic uncertainty quantification. In *International Conference on Machine Learning*, pages 2078–2091. PMLR, 2023.

[4] D. Bethell, S. Gerasimou, and R. Calinescu. Robust uncertainty quantification using conformalised monte carlo prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 20939–20948, 2024.

[5] Q. Chang, X. Li, and Z. Duan. Graph global attention network with memory: A deep learning approach for fake news detection. *Neural Networks*, 172:106115, 2024.

[6] W. Chen, B. Zhang, and M. Lu. Uncertainty quantification for multilabel text classification. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(6):e1384, 2020.

[7] Y. Chung, I. Char, H. Guo, J. Schneider, and W. Neiswanger. Uncertainty toolbox: an open-source library for assessing, visualizing, and improving uncertainty quantification. *arXiv preprint arXiv:2109.10254*, 2021.

[8] S. D. Das, A. Basak, and S. Dutta. A heuristic-driven uncertainty based ensemble framework for fake news detection in tweets and news articles. *Neurocomputing*, 491:607–620, 2022.

[9] Y. Djenouri, A. N. Belbachir, T. Michalak, and G. Srivastava. A federated convolution transformer for fake news detection. *IEEE Transactions on Big Data*, 2023.

[10] Y. Djenouri, A. Belhadi, G. Srivastava, and J. C.-W. Lin. Advanced pattern-mining system for fake news analysis. *IEEE Transactions on Computational Social Systems*, 10(6):2949–2958, 2023.

[11] R. Duan, B. Caffo, H. X. Bai, H. I. Sair, and C. Jones. Evidential uncertainty quantification: A variance-based perspective. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2132–2141, 2024.

[12] P. Erd. On a new law of large numbers. *J. Anal. Muth*, 22:103–l, 1970.

[13] G. Franchi, O. Laurent, M. Leguéry, A. Bursuc, A. Pilzer, and A. Yao. Make me a bnn: A simple strategy for estimating bayesian uncertainty from pre-trained models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12194–12204, 2024.

[14] Á. F. Gambín, A. Yazidi, A. Vasilakos, H. Haugerud, and Y. Djenouri. Deepfakes: current and future trends. *Artificial Intelligence Review*, 57 (3):64, 2024.

[15] A. Gelfert. Fake news: A definition. *Informal logic*, 38(1):84–117, 2018.

[16] K. Huang, Y. Jin, E. Candes, and J. Leskovec. Uncertainty quantification over graph with conformalized graph neural networks. *Advances in Neural Information Processing Systems*, 36, 2024.

[17] C. Kamath. On the role of data mining techniques in uncertainty quantification. *International Journal for Uncertainty Quantification*, 2(1), 2012.

[18] L. Klarner, T. G. Rudner, M. Reutlinger, T. Schindler, G. M. Morris, C. Deane, and Y. W. Teh. Drug discovery under covariate shift with domain-informed prior distributions over functions. In *International Conference on Machine Learning*, pages 17176–17197. PMLR, 2023.

[19] H. Li, J. Song, L. Gao, X. Zhu, and H. Shen. Prototype-based aleatoric uncertainty quantification for cross-modal retrieval. *Advances in Neural Information Processing Systems*, 36, 2024.

[20] Q. Li, M. Gao, G. Zhang, W. Zhai, J. Chen, and G. Jeon. Towards multimodal disinformation detection by vision-language knowledge interaction. *Information Fusion*, 102:102037, 2024.

[21] Y. Li, K. Lee, N. Kordzadeh, and R. Guo. What boosts fake news dissemination on social media? a causal inference view. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 234–246. Springer, 2023.

[22] A. M. Luvembe, W. Li, S. Li, F. Liu, and X. Wu. Caf-odnn: Complementary attention fusion with optimized deep neural network for multimodal fake news detection. *Information Processing & Management*, 61 (3):103653, 2024.

[23] Z. Ma, M. Luo, H. Guo, Z. Zeng, Y. Hao, and X. Zhao. Event-radar: Event-driven multi-view learning for multimodal fake news detection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5809–5821, 2024.

[24] Y. Orlovskiy, C. Thibault, A. Imouza, J.-F. Godbout, R. Rabbany, and K. Pelrine. Uncertainty resolution in misinformation detection. *arXiv preprint arXiv:2401.01197*, 2024.

[25] T. Papamarkou, M. Skoularidou, K. Palla, L. Aitchison, J. Arbel, D. Dunson, M. Filippone, V. Fortuin, P. Hennig, A. Hubin, et al. Position paper: Bayesian deep learning in the age of large-scale ai. *arXiv preprint arXiv:2402.00809*, 2024.

[26] S. Pranave, S. K. Uppada, A. Vishnu Priya, and B. SivaSelvan. Frequent pattern mining approach for fake news detection. In *International Conference on Deep Learning, Artificial Intelligence and Robotics*, pages 103–118. Springer, 2021.

[27] W. Qian, D. Zhang, Y. Zhao, K. Zheng, and J. James. Uncertainty quantification for traffic forecasting: A unified approach. In *2023 IEEE 39th International Conference on Data Engineering (ICDE)*, pages 992–1004. IEEE, 2023.

[28] G. Raman, B. AlShebli, M. Waniek, T. Rahwan, and J. C.-H. Peng. How weaponizing disinformation can bring down a city's power grid. *PloS one*, 15(8):e0236517, 2020.

[29] M. Rivera, J.-F. Godbout, R. Rabbany, and K. Pelrine. Combining confidence elicitation and sample-based methods for uncertainty quantification in misinformation mitigation. *arXiv preprint arXiv:2401.08694*, 2024.

[30] G. Scalia, C. A. Grambow, B. Pernici, Y.-P. Li, and W. H. Green. Evaluating scalable uncertainty estimation methods for deep learning-based molecular property prediction. *Journal of chemical information and modeling*, 60(6):2697–2717, 2020.

[31] K. Shu, D. Mahudeswaran, S. Wang, D. Lee, and H. Liu. Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big data*, 8(3): 171–188, 2020.

[32] K. Shu, G. Zheng, Y. Li, S. Mukherjee, A. H. Awadallah, S. Ruston, and H. Liu. Leveraging multi-source weak social supervision for early detection of fake news. *arXiv preprint arXiv:2004.01732*, 2020.

[33] C.-O. Truică, E.-S. Apostol, and P. Karras. Danes: Deep neural network ensemble architecture for social and textual context-aware fake news detection. *Knowledge-Based Systems*, 294:111715, 2024.

[34] H. Wang and Q. Ji. Epistemic uncertainty quantification for pre-trained neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11052–11061, 2024.

[35] X. Wang, A. Hosseininasab, P. Colunga, S. Kadıoğlu, and W.-J. van Hoeve. Seq2pat: Sequence-to-pattern generation for constraint-based sequential pattern mining. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 12665–12671, 2022.

[36] L. Wu and S. A. Williamson. Posterior uncertainty quantification in neural networks using data augmentation. In *International Conference on Artificial Intelligence and Statistics*, pages 3376–3384. PMLR, 2024.

[37] Y. Wu, B. Shi, J. Chen, Y. Liu, B. Dong, Q. Zheng, and H. Wei. Rethinking sentiment analysis under uncertainty. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 2775–2784, 2023.

[38] D. Zhang, M. Sensoy, M. Makrehchi, B. Taneva-Popova, L. Gui, and Y. He. Uncertainty quantification for text classification. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3426–3429, 2023.

[39] J. Zhang. Modern monte carlo methods for efficient uncertainty quantification and propagation: A survey. *Wiley Interdisciplinary Reviews: Computational Statistics*, 13(5):e1539, 2021.

[40] W. Zhang, Z. M. Ma, S. Das, T.-W. L. Weng, A. Megretski, L. Daniel, and L. M. Nguyen. One step closer to unbiased aleatoric uncertainty estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 16857–16864, 2024.

[41] Z. Zou, X. Meng, A. F. Psaros, and G. E. Karniadakis. Neuraluq: A comprehensive library for uncertainty quantification in neural differential equations and operators. *SIAM Review*, 66(1):161–190, 2024.