Uncertainty-Aware Reliable Text Classification

Yibo Hu
The University of Texas at Dallas
Richardson, TX, USA
yibo.hu@utdallas.edu

Latifur Khan
The University of Texas at Dallas
Richardson, TX, USA
lkhan@utdallas.edu

ABSTRACT
Deep neural networks have significantly contributed to the success in predictive accuracy for classification tasks. However, they tend to make over-confident predictions in real-world settings, where domain shifting and out-of-distribution (OOD) examples exist. Most research on uncertainty estimation focuses on computer vision because it provides visual validation on uncertainty quality. However, few have been presented in the natural language process domain. Unlike Bayesian methods that indirectly infer uncertainty through weight uncertainties, current evidential uncertainty-based methods explicitly model the uncertainty of class probabilities through subjective opinions. They further consider inherent uncertainty in data with different root causes, vacuity (i.e., uncertainty due to a lack of evidence) and dissonance (i.e., uncertainty due to conflicting evidence). In our paper, we firstly apply evidentiary uncertainty in OOD detection for text classification tasks. We propose an inexpensive framework that adopts both auxiliary outliers and pseudo off-manifold samples to train the model with prior knowledge of a certain class, which has high vacuity for OOD samples. Extensive empirical experiments demonstrate that our model based on evidential uncertainty outperforms other counterparts for detecting OOD examples. Our approach can be easily deployed to traditional recurrent neural networks and fine-tuned pre-trained transformers.

CCS CONCEPTS
• Computing methodologies → Natural language processing: Neural networks.

KEYWORDS
out-of-distribution detection; uncertainty qualification; text classification

ACM Reference Format:
Yibo Hu and Latifur Khan. 2021. Uncertainty-Aware Reliable Text Classification. In Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD ’21), August 14–18, 2021, Virtual Event, Singapore. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3447548.3467382

1 INTRODUCTION
Deep neural networks have significantly contributed to the success of predictive accuracy for classification tasks in multiple domains. However, many applications require confidence in reliability. In real-world settings that contain out-of-distribution (OOD) samples, the model should know when it can not make a confident judgment rather than making an incorrect one. Studies show that traditional neural networks easily lead to over-confidence, i.e., a high-class probability in an incorrect class prediction [12, 14, 33]. Therefore, calibrated predictive uncertainty is crucial to avoid those risks.

In this paper, we are interested in qualifying uncertainty to solve OOD detection in text classification as it contains a wide range of Natural Language Processing (NLP) applications [5, 27]. Although fine-tuning pre-trained transformers [8] have achieved state-of-the-art accuracy on text classification tasks, they still suffer from the same over-confidence problem of traditional neural networks, making the prediction untrustful [16]. One partial explanation is over-parameterization [12]. Although transformers are pre-trained on a large corpus and get rich semantic information, it leads to over-confidence easily given limited labeled data during the fine-tuning stage [25]. Overall, compared to the Computer Vision (CV) domain, there is less work in qualifying uncertainty in the NLP domain. Among them, there are Bayesian and non-Bayesian methods.

Bayesian models qualify the model uncertainty by Bayesian neural networks (BNNs) [2, 28]. BNNs explicitly qualify model uncertainty by considering model parameters as distributions. Specifically, BNNs consider probabilistic uncertainty, i.e., aleatoric uncertainty and epistemic uncertainty [24]. Aleatoric only considers data uncertainty caused by statistical randomness. At the same time, epistemic refers to model uncertainty introduced by limited knowledge or ignorance in collected data. Monte Carlo Dropout [10] is a crucial technique to approximate variational Bayesian inference. It trains and evaluates a neural network with dropout [40] before each layer. BNNs have been explored for classification prediction or regression in CV applications. However, there has been less study in the NLP domain. Few work [33, 44, 48] empirically evaluate uncertainty estimation in text classification. Other attempts adopt MC Dropout in deep active learning [37, 38], sentiment analysis [1], or machine translation [53].

Non-Bayesian approaches use entropy [36] or softmax scores as a measure of uncertainty, which only considers aleatoric uncertainty [24]. OOD detection in text classification using GRU [7] or LSTM [18] has been studied in [15, 17]. Hendrycks et al. [16] empirically study pre-trained transformers’ performance on OOD detection. They point out transformers cannot clearly separate in-distribution (ID) and OOD examples. In addition, OOD detection has also been studied in dialogue systems [52] and document classification [13, 50]. Another line of non-Bayesian methods involves the calibration of probabilities. Temperature scaling [12] calibrates softmax probabilities by adding a scalar parameter to each class in a post-processing step. Thulasidasan et al. [43] explore the improvement of calibration and predictive uncertainty of models trained...
with mix-up [49] in the NLP domain. Kong et al. [25] use pseudo samples on and off the data manifold for calibration.

Besides probabilistic uncertainty and BNNs, evidential uncertainty is proposed based on belief/evidence theory and Subjective Logic (SL) [22, 23]. It considers different dimensions of uncertainty, such as vacuity (i.e., lack of evidence) or dissonance (i.e., uncertainty due to conflicting evidence). In the CV domain, Sensory et al. [34] propose evidential neural networks (ENN) to model the uncertainty of class probabilities based on SL explicitly. An ENN uses the predictions as subjective opinions and learns a function that collects evidence to form the opinions by a deterministic neural network from data. Several works [19, 34, 51] improve ENNs using regularization or generative models to ensure correct uncertainty estimation towards unseen examples in image classification. However, those methods for continuous feature space are not applicable to the discrete text.

To briefly demonstrate the motivation of our paper, we use a simple binary classification example in Table 1 and Figure 1 to answer the following questions:

- Why is it necessary to calibrate predictive uncertainty?
- What is the advantage of evidential uncertainty in OOD detection?
- How to design a regularization method to calibrate the predictive uncertainty?

Table 1, we assume that a classifier is only trained on the restaurant reviews dataset and has never seen examples from other domains. The probability denotes the prediction softmax probability. The evidence represents historical observations, denoted by Dirichlet distributions (no evidence when \( a = 1 \)). Before calibration, the classifier predicts Sentence 3, an obvious OOD example, as positive with high confidence. Thus it is necessary to calibrate predictive uncertainty to reduce over-confidence.

For a well-calibrated model, there are three common cases in predictions. Sentence 1 refers to correct confident classification, where we have enough evidence with no conflicts. Sentence 2 is vague and contains conflicting information like ‘bad’ and ‘acceptable’. The prediction will result in equal probability because each category contains conflicting information like ‘bad’ and ‘acceptable’. The model outputs the predictive uncertainty. We can easily get diverse text data as auxiliary outliers. However, generating adversarial examples via common gradient-based approaches is impossible in the NLP domain. Thus, we apply methods [11, 25, 41] to generate off-manifold adversarial examples from the embedding layer.

Our work provides the following key contributions: (i) We firstly apply evidential uncertainty to solve OOD detection tasks in the text classification. (ii) We propose an inexpensive framework that adopts both an auxiliary dataset of outliers and generated pseudo off-manifold samples to train a model with prior knowledge of a certain class, which has high vacuity for OOD samples. (iii) We validate our proposed method’s performance via extensive experiments of OOD detection and uncertainty estimation in text classification. Our approach significantly outperforms all the counterparts.

2 PRELIMINARIES

We briefly provide the background knowledge of evidential uncertainty and the advantage over probabilistic uncertainty.

2.1 Subjective Opinions in SL

A multinomial opinion in a given proposition \( x \) is represented by \( \omega_Y = (b_Y, u_Y, a_Y) \) where a domain is \( Y \equiv \{1, \ldots, K\} \), a random variable \( Y \) takes value in \( Y, K = |Y| \geq 2 \) and \( \omega_Y \) is given as \( \sum_{y \in Y} b_Y(y) + u_Y = 1 \). \( b_Y \) denotes belief mass function over \( Y \), \( u_Y \) denotes uncertainty mass representing vacuity of evidence. \( a_Y \) represents base rate distribution over \( Y \), with \( \sum_y a_Y(y) = 1 \). Then the projected probability distribution of a multinomial opinion is given by:

\[
p_Y(y) = b_Y(y) + a_Y(y)u_Y, \quad \forall y \in Y. \tag{1}
\]

Multinomial probability density over a domain of cardinality \( K \) is represented by the \( K \)-dimensional Dirichlet PDF where the special case with \( K = 2 \) is the Beta PDF as a binomial opinion. It denotes a domain of \( K \) mutually disjoint elements in \( Y \) and \( a_Y \) the strength vector over \( y \in Y \) and \( p_Y \) the probability distribution over \( Y \).

\[
\text{Dir}(p_Y; a_Y) = \frac{1}{B(a_Y)} \prod_{y \in Y} p_Y(y)^{a_Y(y) - 1}, \tag{2}
\]

Finally, in Figure 1 we also observe OOD examples and adversarial examples. Adversarial examples [4, 30, 42] refer to instances with small feature perturbations. A lot of studies [20, 21, 46] use adversarial examples to evaluate and improve neural networks’ robustness. We can use diverse outliers to calibrate the model to output high uncertainty in the OOD region [17]. Additionally, adversarial examples can be helpful to detect OOD examples close to ID regions. Thus, our approach adopts a mixture of an auxiliary dataset of outliers and close adversarial examples to calibrate the predictive uncertainty. We can easily get diverse text data as auxiliary outliers. However, generating adversarial examples via common gradient-based approaches is impossible in the NLP domain. Thus, we apply methods [11, 25, 41] to generate off-manifold adversarial examples from the embedding layer.
Table 1: Predictive uncertainty of sentiment analysis of restaurant reviews. The model without calibration demonstrates over-confidence. A well-calibrated classifier outputs the same expected probabilities for Case 2 and 3 that have different evidence.

| Calibrated? | Test Sentence                                                                 | Probability | Dirichlet | Uncertainty |
|------------|-------------------------------------------------------------------------------|-------------|-----------|-------------|
| No         | 1. 'This was the worst restaurant I have ever had the misfortune of eating at.' | \( p = [0.99, 0.1] \) | doesn’t apply | Over-confidence |
|            | 2. 'This restaurant is bad. Yet its food is acceptable considering the low price.' | \( p = [0.5, 0.5] \) | \( \alpha = [1, 50] \) | Low uncertainty |
|            | 3. 'Deep learning is data hungry.'                                               | \( p = [0.5, 0.5] \) | \( \alpha = [1, 1] \) | Lack of evidence |

![Figure 1: Visualization of the predictive uncertainty in Table 1. (a) Traditional NNs with softmax function before calibration demonstrates over-confidence. (b) A well-calibrated model shows high entropy in both conflicting and OOD regions. (c) and (d) shows evidential uncertainty that decompose the uncertainty in (b) based on different root causes. The pentagrams denote the three test cases in Table 1.](image)

where \( B(\alpha_Y) \) is a multivariate beta function as the normalizing constant, \( \alpha_Y(y) \geq 0 \), and \( p_Y(y) \neq 0 \) if \( \alpha_Y(y) < 1 \).

We term evidence as a measure of the amount of supporting observations collected from data in favor of a sample to be classified into a certain class. Let \( r_Y(y) \geq 0 \) be the evidence derived for the singleton \( y \in Y \). The total strength \( \alpha_Y(y) \) for the belief of each singleton \( y \in Y \) is given by:

\[
\alpha_Y(y) = r_Y(y) + \alpha_Y(y)W, \tag{3}
\]

where \( W \) is a non-informative weight representing the amount of uncertain evidence and \( \alpha_Y(y) \) is the base rate distribution. Given the Dirichlet PDF, the expected probability distribution over \( Y \) is:

\[
E_Y(y) = \frac{\alpha_Y(y)}{\sum_{y_i \in Y} \alpha_Y(y_i)} = \frac{r_Y(y) + \alpha_Y(y)W}{W + \sum_{y_i \in Y} r_Y(y_i)}, \forall y \in Y. \tag{4}
\]

The observed evidence in the Dirichlet PDF can be mapped to the multinomial opinions by:

\[
b_Y(y) = \frac{r_Y(y)}{S}, \quad u_Y = \frac{W}{S}, \tag{5}
\]

where \( S = \sum_{y_i \in Y} \alpha_Y(y_i) \). We set the base rate \( \alpha_Y(y) = \frac{1}{K} \) and the non-informative prior weight \( W = K \), and hence \( \alpha_Y(y) \cdot W = 1 \) for each \( y \in Y \), as these are default values considered in subjective logic.

2.2 Uncertainty Dimensions

Jøsang et al. [23] define multiple dimensions of a subjective opinion based on the formalism of SL. Vacuity refers to uncertainty caused by insufficient information to understand a given opinion.

It corresponds to uncertainty mass, \( u_Y \), of an opinion in SL as:

\[
\text{Vac}(\alpha_Y) = \frac{W}{S}. \tag{6}
\]

Dissonance denotes when there is an insufficient amount of evidence that can clearly support a particular belief. We observe high dissonance when the same amount of evidence is supporting multiple extremes of beliefs. Given a multinomial opinion with non-zero belief masses, the measure of dissonance can be obtained by:

\[
\text{Diss}(\alpha_Y) = \sum_{y_j \in \bar{Y}} \left( \frac{b_Y(y_j)}{\sum_{y_i \in \bar{Y}} b_Y(y_i)} \right) \left( \frac{\sum_{y_i \in \bar{Y}} b_Y(y_i) \text{Bal}(y_j, y_i)}{\sum_{y_i \in \bar{Y}} b_Y(y_i)} \right), \tag{7}
\]

where the relative mass balance between a pair of belief masses \( b_Y(y_j) \) and \( b_Y(y_i) \) is expressed by:

\[
\text{Bal}(y_j, y_i) = \begin{cases} 
\frac{|b_Y(y_j) - b_Y(y_i)|}{b_Y(y_j) + b_Y(y_i)}, & \text{if } b_Y(y_j) \neq b_Y(y_i) \\
0, & \text{otherwise}.
\end{cases} \tag{8}
\]

The above two uncertainty measures (i.e., vacuity and dissonance) can be interpreted using class-level evidence measures of subjective opinions. As in Table 1, given two classes (positive, and negative), we have three subjective opinions \( \{\alpha_1, \alpha_2, \alpha_3\} \), represented by the two-class evidence measures as: \( \alpha_1 = (1, 99) \) representing low uncertainty (entropy, dissonance and vacuity) which implies high confidence in a decision making context. \( \alpha_2 = (50, 50) \) indicating high inconclusiveness due to high conflicting evidence which gives high entropy and high dissonance, \( \alpha_3 = (1, 1) \) showing the case of high vacuity which is commonly observed in OOD samples. Therefore, vacuity can effectively distinguish OOD samples from boundary samples because it represents a lack of evidence.
3 APPROACH

3.1 Calibrating Evidential Neural Networks

ENNs [35] predict the evidence vector for the predicted Dirichlet distribution instead of softmax probability. Given a sample \( x_i \) with the input feature \( x_i \in \mathbb{R}^L \) and the ground-truth label \( y_i \), let \( f(x_i|\theta) \) represents the predicted evidence vector predicted by the classifier with parameters \( \Theta \). Then the corresponding Dirichlet distribution has parameters \( \alpha_i = f(x_i|\theta) + 1 \). The Dirichlet density \( \text{Dir}(p; \alpha) \) is the prior on the Multinomial distribution \( \text{Multi}(y_i|p_i) \). Then we optimize the following sum of squared loss for classification:

\[
\mathcal{L}(f(x_i|\theta), y_i) = \sum_{j=1}^{K} \frac{\left\| y_{ij} - p_{ij} \right\|^2}{B(\alpha_j)} + \sum_{j=1}^{K} \int \left( \frac{1}{p_{ij}} - 1 \right) \text{d}p_{ij} + \sum_{j=1}^{K} \left( y_{ij} - 2y_{ij} \right) E[p_{ij}] + E[p_{ij}].
\] (9)

Since Eq. (9) only relies on class labels of training samples, it does not directly measure the quality of the predicted Dirichlet distributions. The uncertainty estimates may not be accurate. Thus, we propose a regularization method that combines ENNs and language models to quantify evidential uncertainty in text classification tasks. Formally, given a set of samples \( \mathcal{D}_m = \{(x_1, y_1), \ldots, (x_N, y_N)\} \), where \( x_i \) refers to input embedding of sentences or documents and \( y_i \) is its label. Let \( P_{\text{in}}(x, y) \) and \( P_{\text{out}}(x, y) \) be the distributions of the OOD and ID samples respectively. Let \( g(\cdot) \) denote the function of the pre-trained feature extraction layers. Let \( h(\cdot) \) denote the task-specific layers. We use \( \Theta \) to represent the parameters of \( g \) and \( h \). Then we fine-tune our model by optimizing the following loss function over the parameters \( \Theta \):

\[
\min_{\Theta} \mathcal{F}(\Theta) = \mathbb{E}_{x,y-P_{\text{in}}(x,y)} \left[ \mathcal{L}(h \circ g(x|\Theta), y) \right] + \beta_{\text{in}} \cdot \mathbb{E}_{x-P_{\text{in}}(x)} \left[ \text{Vac}(h \circ g(x|\Theta)) \right] - \beta_{\text{out}} \cdot \mathbb{E}_{x-P_{\text{out}}(x)} \left[ \text{Vac}(h \circ g(x|\Theta)) \right].
\] (10)

The first item refers to the vanilla classification loss of ENN Eq. (9), which ensures a reasonable estimation of the ID samples’ class probabilities. The second item is to reduce the vacuity estimation on ID samples. The third item is to increase the vacuity estimation on OOD samples. \( \beta_{\text{in}} \) and \( \beta_{\text{out}} \) are the trade-off parameters. The goal of minimizing Eq. (10) is to achieve high classification accuracy, low vacuity output for ID samples, and high vacuity output for OOD samples. To ensure the model’s generalization to the whole data space, the choice of effective \( P_{\text{out}} \) is crucial. Although generative models have achieved success in the CV domain [19, 34], they do not apply to discrete text data. We adopt two methods that have achieved success in the NLP domain to get effective OOD regularization: (i) Using auxiliary OOD datasets; (ii) Generating off-manifold adversarial examples.

3.2 Utilizing Auxiliary Datasets

The auxiliary datasets disjointed from the test datasets can be used to calibrate the neural networks’ over-confidence for unseen samples. A critical finding in [17] is that the diversity of the auxiliary dataset is important. Hu et al. [19] report that the methods using diverse examples beat the methods that only use close adversarial examples [14, 34] in OOD detection in image classification. Our empirical observations also find that randomly generated sentences (we randomly sample words and concatenate them into fake sentences) do not improve the performance. One partial explanation is that these “sentences” do not contain useful semantic information. This is similar to the CV domain, where CNN models do not extract valuable features from random pixel image samples. Since it is easy to get a large corpus of diverse text data, utilizing a real dataset is inexpensive and straightforward. Let \( P_{\text{aux}}(x) \) be the distribution of the OOD auxiliary dataset, the regularization can be written as:

\[
\max_{\Theta} \mathbb{E}_{x-P_{\text{aux}}(x)} \left[ \text{Vac}(h \circ g(x|\Theta)) \right]
\] (11)

3.3 Utilizing Off-manifold samples

Kong et al. [25] encourage the model to output uniform distributions on pseudo off-manifold samples to alleviate the over-confidence in OOD regions. On the contrary, we apply off-manifold samples by enforcing the model to predict high vacuity:

\[
\max_{\Theta} \mathbb{E}_{x-P_{\text{off}}(x')} \left[ \text{Vac}(h \circ g(x'|\Theta)) \right]
\] (12)

where \( P_{\text{off}}(x') \) denotes the distributions of the adversarial examples. The off-manifold samples are generated from adding relatively large perturbations towards the outside of the data manifold. In our NLP tasks, the data manifold refers to the embedding space because the original text is not continuous. Formally, given a training ID sample (embedding) \( (x_i, y_i) \), we generate the off-manifold sample \( x'_i \) by:

\[
x'_i = \max_{x'_i \in S(\xi_i, \delta_{\text{off}})} \mathcal{L}(h \circ g(x'_i|\Theta), y_i)
\] (13)

where \( S(\xi_i, \delta_{\text{off}}) \) denotes an \( \ell_\infty \) sphere centered at \( x_i \) with a radius \( \delta_{\text{off}} \). The \( \delta_{\text{off}} \) is relatively large to ensure that the sphere \( \delta_{\text{off}} \) lies outside of the data manifold [11, 41]. Then we can get pseudo off-manifold samples from \( \delta_{\text{off}} \) along the adversarial direction, which is calculated from the gradient of the classification loss.

Off-manifold samples can improve the uncertainty estimation in close OOD regions. However, the generalization of adversarial samples relies on the diversity of the features of the training data. Hu et al. [19] report that models trained on CIFAR-10 can generate better adversarial examples for regularization than models trained on SVHN [32]. Because CIFAR-10 contains more diverse features than SVHN, a dataset of only street numbers. Our empirical observations find that off-manifold samples can help when combined with pre-trained transformers. However, it does not provide significant improvement in vanilla GRUs/ LSTMs. This is consistent with the empirical study [16] where pre-trained transformers outperform vanilla models in generalization towards OOD regions. The embeddings of pre-trained transformers contain rich features that benefit the generated adversarial examples. Thus following [25], we evaluate off-manifold regularization on BERT [8].

3.4 Mixture Regularization

Auxiliary datasets regularization provides an overall calibration improvement, while off-manifold regularization focuses more in the close OOD region. We replace the last item in Eq. (10), which represents the uncertainty regularization for OOD data to the mixture.
We follow the same benchmark in [17]. We use the same three
OOD examples. Our experiments consist of three parts: (i) We follow
4.1 Datasets
movie reviews used for sentiment classification. (ii) TREC [29] consists of highly polar
natural language inference. (ii) M30K [22] denotes Stanford Sentiment Treebank [39], a collection of
movie reviews used for sentiment classification. (iii) M30K refers
to the English portion of Multi-30K [9], a dataset of image descriptions. (iv) WMT16 denotes the English portion of the test set from
WMT16. (v) Yelp is a dataset of restaurant reviews.

4.2 Comparing Schemes
We compare several recent methods for qualifying uncertainty or
OOD detection in text categorization. (i) MSP refers to maximum softmax
probability, a baseline work of OOD detection [15]. (ii) DP refers to Monte Carlo Dropout [10], which applies dropout at
train and test time. We run ten it times and use the average MSP as the
uncertainty score. (iii) TS is a dataset of restaurant reviews.

Algorithm 1 Fine tuning our proposed mixed uncertainty model.
f denotes ENN h o g(·) with weights Θ, m is the batch size. d is the
dimension of features.
1: for each iteration do
2: Sample {(x(i), y(i))}m−1 i=1 ∼ D_in and {(x(i))}m i=1 ∼ D_out
3: Update ENN by descending the gradient
\[ \nabla_{\Theta} \sum_{m=1}^{m} [L(h(\theta), g(x(i))) + \beta_{in} Vac(h(\theta), g(x(i))) + \beta_{oe} Vacc(h(\theta), g(x(i))) + \beta_{off} \text{Vac}(h(\theta), g(x(i)))]. \]
4: Update ENN by ascending the gradient
\[ \nabla_{\Theta} \sum_{i=1}^{N} [\text{Vac}(h(\theta), g(x(i)))]. \]
5: Initialize x_i ← x_i + \delta_i with \delta_i ~ UNIF[−δ_{off}, δ_{off}]^d
6: Get the gradient of the classification loss
\[ \Delta_i = \text{sign}(\nabla_{x_i} L(f(x_i(\theta), y_i)). \]
7: Add perturbations towards off-manifold
\[ x_i' \sim \Pi_{x_i'} \sum_{i=1}^{N} \Delta_i]. \]
8: Update ENN by ascending the gradient
\[ \nabla_{\Theta} \sum_{i=1}^{N} \text{Vac}(h(\theta), g(x(i)))]. \]
9: end for

4.3 Metrics
We consider the following metrics in [15, 17]: The area under
the receiver operating characteristic curve (AUROC), the area under
the precision-recall curve (AUPR) and the False Alarm Rate at
90% Recall (FAR90). Higher AUROC indicates a higher probability
that a positive example has a higher score than a false example, which
means better accuracy. AUPR is similar to AUROC, but it
also considers the positive class’s base rate. Higher AUPR is better.
FAR90 measures the probability that a false example raises a false
alarm, assuming that 90% of all positive examples are detected. Lower FAR90 is better.
Table 2: The results of OOD detection using two-layer GRUs on multiple datasets. Our model (+OE) uses an auxiliary dataset for regularization.

| D_in  | 2D out  | FPR90 ↓ | AUROC ↑ | AUPR ↑  |
|-------|---------|---------|---------|--------|
|       | MSP     | DP      | OE      | Ours   | MSP     | DP      | OE      | Ours   | MSP     | DP      | OE      | Ours   |
| SNLI  | 38.2    | 27.4    | 21.6    | 12.5   | 13.2    | 87.6    | 91.4    | 92.7    | 95.1   | 93.7    | 71.3    | 78.0    | 81.4    | 86.3   | 71.9   |
| IMDB  | 45.0    | 36.0    | 27.8    | 19.2   | 9.2     | 79.9    | 85.1    | 88.0    | 93.6   | 96.0    | 42.4    | 50.8    | 54.5    | 74.4   | 76.3   |
| M30K  | 54.5    | 42.8    | 46.0    | 3.4    | 3.8     | 78.3    | 84.8    | 82.7    | 97.3   | 98.5    | 46.0    | 60.3    | 46.3    | 93.6   | 94.9   |
| WMT16 | 38.7    | 29.3    | 26.7    | 1.6    | 0.8     | 85.2    | 89.8    | 88.8    | 99.0   | 99.5    | 57.3    | 69.2    | 56.8    | 96.6   | 98.1   |
| Yelp  | 54.6    | 41.2    | 39.4    | 4.0    | 8.5     | 78.8    | 82.8    | 82.5    | 97.7   | 96.5    | 37.9    | 53.5    | 41.6    | 87.8   | 83.0   |
| Mean  | 44.4    | 35.4    | 32.3    | 8.1    | 7.1     | 81.96   | 86.62   | 86.94   | 96.34  | 96.8    | 50.98   | 60.72   | 36.12   | 87.74  | 84.94  |

For the GRU experiments, we use the source code of MSP and OE in [17]. We follow the same pre-processing steps and the base rate of D out to D en is 1:5 in each scenario. We implement ENN, DP, and our model based on the same two-layer GRUs. We pre-train the base classifier for five epochs and fine-tune five more epochs for OE and our model using WikiText-2. Except for DP, we pre-train it for ten epochs to ensure the same accuracy as others. We evaluate our model with auxiliary datasets regularization (+OE).

For the experiments on BERT, we follow the same setting in [25], which also contains the implementation of multiple baselines. We still set the base rate of D out to 1:5 to be consistent with the previous experiments. We construct sequence classifiers with one linear layer on top of the pooled output of a pre-trained uncased BERT base model. Then we fine-tune it with different models for ten epochs. We evaluate auxiliary datasets regularization (+OE), adversarial regularization (+AD), and the mixture method (MIX).

We fairly train all the baselines with their default parameters and report the average results. In the GRU experiments, we set β in = 0.1, β oe = 1, batch size = 128, learning_rate = 1e-4 in Adam optimizer of our model in all the experiments, which were fine-tuned considering the performance of both the OOD detection and ID classification accuracy. For the experiments on BERT, we set β oe = 1 in all +OE and MIX, learning_rate = 5e-5 in Adam optimizer in all experiments. But we use slightly different β in for each D train in, which is fine-tuned considering the accuracy and vacuity from the validation ID set. For more details, refer to Section 4.7 and our source code 1.

4.4 Out-of-Distribution Detection

In Table 2, our model on GRU significantly outperforms other approaches on SST and achieves the overall best results on TREC. Except on 20News, OE slightly outperforms ours. One partial explanation is that simple GRUs can not handle accuracy and uncertainty estimation simultaneously when handling longer texts. The average accuracy of all the models is only 73%, which indicates that the models have not learned the correct evidence.

Table 3 shows that pre-trained models still suffer from over-confidence. DP does not outperform MSP, which is consistent with [45] that MC Dropout only measures uncertainty in ID settings. TS still relies on softmax probability and tune its temperature parameter on the validation (ID) set. Thus TS does not generalize well in unseen data. Therefore, effective OOD detection models require regularization from OOD examples. OE using a diverse real auxiliary dataset beats MRC that adopts adversarial examples, except in the close OOD setting SST vs. IMDB. Our model (MIX) applies both regularizations and beats both of them.

Table 4 further analyzes the contribution of each regularization. Both +OE and +AD improve the performance of vanilla ENN. +OE outperforms the baseline OE. This indicates the effectiveness of evidential uncertainty when using the same regularization. While +OE provides an overall improvement, +AD is especially effective in distinguishing OOD examples. For example, in SST vs. IMDB, our model (MIX) applies both regularizations and beats both of them.

4.5 Predictive Uncertainty Distribution

We use boxplots to show the uncertainty distribution of different models deployed on BERT in Fig.3. Baselines use entropy as a measure of uncertainty. Our proposed model use vacuity (Vac) and the square root of dissonance (Dis) ranged from [0, 1]. We also show the output of our entropy (Ent). The top row shows the predictive uncertainty in D tra in and compares them to those for all the OOD datasets. We concatenate all the five OOD datasets as OOD examples in these experiments. The bottom row shows different models’

1https://github.com/snowood1/BERT-ENN
Table 3: The results of OOD detection using BERT on multiple datasets. Our model (MIX) applies both an auxiliary dataset and off-manifold adversarial samples for regularization.

| D_in | D_out | FPR90 ↓ | AUROC ↑ | AUPR ↑ |
|------|-------|---------|---------|--------|
|      |       | MSP | DP | TS | MRC | OE | Ours | MSP | DP | TS | MRC | OE | Ours | MSP | DP | TS | MRC | OE | Ours |
| SNLI | 23.34 | 9.00 | 79.50 | 85.20 | 83.90 | 99.40 | 98.70 | 96.10 | 81.50 | 75.50 | 75.30 | 99.70 | 99.40 | 100.00 | 85.10 | 80.90 | 87.80 | 97.60 | 100.00 | 100.00 |
| IMDB | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| M30K | 31.50 | 64.50 | 14.50 | 83.80 | 85.20 | 83.90 | 99.40 | 98.70 | 96.10 | 81.50 | 75.50 | 75.30 | 99.70 | 99.40 | 100.00 | 85.10 | 80.90 | 87.80 | 97.60 | 100.00 | 100.00 |
| WMT16 | 31.50 | 64.50 | 14.50 | 83.80 | 85.20 | 83.90 | 99.40 | 98.70 | 96.10 | 81.50 | 75.50 | 75.30 | 99.70 | 99.40 | 100.00 | 85.10 | 80.90 | 87.80 | 97.60 | 100.00 | 100.00 |
| Yelp | 31.50 | 64.50 | 14.50 | 83.80 | 85.20 | 83.90 | 99.40 | 98.70 | 96.10 | 81.50 | 75.50 | 75.30 | 99.70 | 99.40 | 100.00 | 85.10 | 80.90 | 87.80 | 97.60 | 100.00 | 100.00 |

Table 4: The ablation study of different regularization’s effects on BERT-ENNs. We show vanilla ENNs, with auxiliary outliers (+OE), with off-manifold examples (+AD), and with the mixture of both methods (MIX). We also list the best counterpart OE.

| D_in | D_out | FPR90 ↓ | AUROC ↑ | AUPR ↑ |
|------|-------|---------|---------|--------|
|      |       | OE | ENN | +OE | +AD | MIX | OE | ENN | +OE | +AD | MIX | OE | ENN | +OE | +AD | MIX |
| SNLI | 23.34 | 9.00 | 79.50 | 85.20 | 83.90 | 99.40 | 98.70 | 96.10 | 81.50 | 75.50 | 75.30 | 99.70 | 99.40 | 100.00 | 85.10 | 80.90 | 87.80 | 97.60 | 100.00 | 100.00 |
| IMDB | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| M30K | 31.50 | 64.50 | 14.50 | 83.80 | 85.20 | 83.90 | 99.40 | 98.70 | 96.10 | 81.50 | 75.50 | 75.30 | 99.70 | 99.40 | 100.00 | 85.10 | 80.90 | 87.80 | 97.60 | 100.00 | 100.00 |
| WMT16 | 31.50 | 64.50 | 14.50 | 83.80 | 85.20 | 83.90 | 99.40 | 98.70 | 96.10 | 81.50 | 75.50 | 75.30 | 99.70 | 99.40 | 100.00 | 85.10 | 80.90 | 87.80 | 97.60 | 100.00 | 100.00 |
| Yelp | 31.50 | 64.50 | 14.50 | 83.80 | 85.20 | 83.90 | 99.40 | 98.70 | 96.10 | 81.50 | 75.50 | 75.30 | 99.70 | 99.40 | 100.00 | 85.10 | 80.90 | 87.80 | 97.60 | 100.00 | 100.00 |

4.6 Parameter Study

The most important parameters are $\delta_{off} = 0.01$ and $\beta_{oe} = 1$. $\delta_{off}$ influences the performance of adversarial regularization greatly. We find that $\delta_{off} = 0.01$ achieves the best performance across all of our experiments. Figure 4 shows the FPR90 of our model using off-manifold regularization (+AD) in the scenario SST ($D_{in}$ vs. IMDB ($D_{out}$). We observe the same performance in all the other scenarios. When $\delta_{off}$ is too small, the generated samples might be too close to the manifold and may harm the confidence of the ID region. Too much perturbation leads to ineffective samples for regularization.
Figure 3: Top row: The boxplots of predictive uncertainty of different models on different datasets $D_{test}^{in}$ vs. examples from all the four OOD datasets $D_{test}^{out}$. Bottom row: The boxplots of predictive uncertainty of successful and failed predictions in different datasets $D_{test}^{in}$. Our model uses entropy (Ent), vacuity (Vac), dissonance (Dis) as a measure of uncertainty, while other models use entropy.

We also compare the effect of the weights of different regularization terms in the mixture formula. We find that +OE provides an overall improvement in calibration, and we simply set $\beta_{oe} = 1$. We try different $\beta_{ad} = 1$ or 0.1 to better distinguish close OOD examples. $\beta_{in}$ is tuned via the validation ID set within three possible values 0, 0.01 and 0.1. Since the first item in Eq. (10) already assigns considerable confidence in training samples during the classification process, it also reduces ID samples’ vacuity. Large $\beta_{in}$ may also affect the accuracy. Therefore we only use a small $\beta_{in}$ to scale the vacuity of ID examples slightly. The summary of different weights can be seen in Table 5.

5 RELATED WORK

Our study is related to uncertainty qualification [2, 10, 35], OOD detection [15, 17] and confidence calibration [12, 25, 43]. We have discussed the NLP applications of these fields in the Introduction.

Other baselines not included in our experiments include Deep Ensemble [26], which average the softmax outputs of five models with different initialization. A recent empirical study [33] proves that Deep Ensemble performs better than Dropout and Temperature Scaling under dataset shift of NLP tasks using LSTM [18]. However, fine-tuning multiple pre-trained transformer models is computationally expensive. Besides, the advantage of our considered baseline OE over this method has been reported in [31]. Therefore we do not consider this method as a baseline in our paper. Another line of work, Stochastic Variational Bayesian Inference [2, 28, 47] can be applied to CNN models but hard to be applied in other architectures such as LSTMs [33]. Hu et al. [19], Sensoy et al. [35] also prove the advantage of ENNs over multiple Stochastic Variational Bayesian Inference methods.

6 CONCLUSION

Qualifying uncertainty is essential for reliable classification, but less work has been studied in the NLP domain. We firstly apply evidential uncertainty based on SL to solve OOD detection in the text classification. We combine ENNs and language models to measure vacuity and dissonance. Our proposed model uses auxiliary datasets of outliers and off-manifold samples to train a model with prior knowledge of a certain class, which has high vacuity for OOD samples. Extensive experiments show that our approach significantly outperforms all the counterparts.
