Bayesian BERT for Trustful Hate Speech Detection

Kristian Miok [1 2] Blaž Škrlj [3] Daniela Zaharie [1] Marko Robnik Šikonja [2]

Abstract

Hate speech is an important problem in the management of user-generated content. In order to remove offensive content or ban misbehaving users, content moderators need reliable hate speech detectors. Recently, deep neural networks based on transformer architecture, such as (multilingual) BERT model, achieve superior performance in many natural language classification tasks, including hate speech detection. So far, these methods have not been able to quantify their output in terms of reliability. We propose a Bayesian method using Monte Carlo Dropout within the attention layers of the transformer models to provide well-calibrated reliability estimates. We evaluate the introduced approach on hate speech detection problems in several languages. Our approach not only improves the classification performance of the state-of-the-art multilingual BERT model but the computed reliability scores also significantly reduce the workload in inspection of offending cases and in reannotation campaigns.

1. Introduction

With the rise of the social network popularity, hate speech phenomena has significantly increased (Davidson et al., 2017). Hate speech not only harms both minority groups and the whole society but it can lead to actual crimes (Bleich, 2011). Hence, automated hate speech detection mechanisms are urgently needed. On the other hand, falsely accusing people of hate speech is also a problem. Many content providers rely on human moderators to reliably decide if a given context is offensive or not but this is a mundane and stressful job which can even cause post-traumatic stress disorders[1]. There are many attempts to automate detection of hate speech in the social media using machine learning, but existing models lack quantification of reliability for their decisions. In the last few years, recurrent neural networks (RNNs) were the most popular choice in text classification. The LSTM networks, the most successful RNN architecture, were already successfully adapted for assessment of predictive reliability in hate speech classification (Miok et al., 2019b). Recently, neural network architecture with attention layers, called transformer architecture (Vaswani et al., 2017), shows even better performance on almost all language processing tasks. Using transformer networks for the task of masked language modelling produced breakthrough pretrained models such as BERT (Devlin et al., 2018). Hence, the attention mechanism seems to be at the forefront of natural language understanding with potentially huge impact on language applications. We aim to investigate the behavior of the attention mechanism concerning the reliability of predictions. We focus on the hate speech recognition task.

In hate speech detection, reliable predictions are needed to remove harmful contents and possibly ban malicious users without harming the freedom of speech (Miok et al., 2019b). Standard neural networks are inadequate for assessment of predictive uncertainty, and the Bayesian framework is the principled approach to doing so. However, classical Bayesian inference techniques do not scale well in neural networks with high dimensional parameter space (Izmailov et al., 2019). Various methods were proposed in order to overcome this problem (Myshkov & Julier, 2016). One of the most efficient method is called Monte Carlo Dropout (MCD) (Gal & Ghahramani, 2016). Its idea is to use dropout in neural networks as a regularization technique (Srivastava et al., 2014) and interpret it as a method that mimics Bayesian approach.

We propose a model that combines the attention mechanism in transformer networks with the MCD based Bayesian inference in order to estimate reliability of hate speech predictions. Our main contributions are estimating prediction uncertainty of the attention network (AN) and BERT model and testing the proposed reliability methods within the multilingual hate speech detection tasks.

[*] Equal contribution [1] Department of Computer Science, West University of Timisoara, Timisoara, Romania [2] Faculty of Computer and Information Science, University of Ljubljana, Ljubljana, Slovenia [3] Jožef Stefan International Postgraduate School, Jožef Stefan Institute, Ljubljana, Slovenia. Correspondence to: Kristian Miok <kristian.miok@e-uvt.ro>.

Presented at the ICML 2020 Workshop on Uncertainty and Robustness in Deep Learning.

[1] https://www.bbc.com/news/technology-51245616

The paper consists of four more sections. In Section 2, we propose the methodology for uncertainty assessment using attention layers and MCD. Section 3 presents the data sets and the evaluation scenario. The obtained results are presented in Section 4, followed by the conclusions in Section 5.

2. Bayesian Attention Networks

The BERT model (Devlin et al., 2018) is the representative of transformer networks and has achieved state-of-the art results in many NLP tasks, including text classification (Xu et al., 2020; Gururangan et al., 2019; Chang et al., 2019). In this work, we introduce Monte Carlo Dropout to transformer networks and BERT with the intention to construct their Bayesian variants. Analysis of different amounts of dropout, different BERT variants modifications, and their hyperparameters would require huge computational resources, e.g., training a single BERT model on four TPUs requires more than a month time. Due to limited computational resources, we explore these issues in a limited setting, first on only the encoder part of the BERT architecture, called Attention Network (AN), and then on the entire pretrained BERT model.

In the following subsections, we first formally define the Attention Network architecture, and then make it Bayesian by introducing MCD. We describe how we can introduce MCD principle into the already pretrained BERT model.

2.1. Attention Networks

The basic architecture of Attention Network follows the architecture of transformer networks (Vaswani et al., 2017) and is shown in Figure 1. The architecture is similar to the encoder part of the transformer architecture. The difference is in the output part where a single output head was added to perform either binary classification, using the sigmoid activation function. By applying only the encoder part of transformer architecture, orders of magnitude less parameters are needed to learn a particular classification task, e.g., in this work, we used at maximum 3 million parameters. The architecture can contain many attention heads, where a single attention heads is computed as:

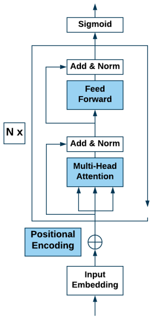


Figure 1. A scheme of Attention Networks. In layers colored blue we introduce the dropout.

$$o_h = \mathrm{softmax}(\frac{\boldsymbol{Q} \cdot \boldsymbol{K}^T}{\sqrt{d_k}}) \cdot \boldsymbol{V},$$

The attention matrices are commonly known as the query $\boldsymbol{Q}$, the key $\boldsymbol{K}$, and the value matrix $\boldsymbol{V}$. The o_h represents the output. The attention function can be described as mapping a query and a set of key-value pairs to an output, where the query, keys, values, and output are all vectors. The output is computed as a weighted sum of the values, where the weight assigned to each value is computed by a compatibility function of the query with the corresponding key. The d_k represents the dimensionality of the keys. The positional encoding, as discussed in (Vaswani et al., 2017), represents a matrix that encodes individual positions in a matrix of same dimensionality as the one holding the information on sequences (input embedding).

2.2. Monte Carlo Dropout for Attention Networks

MCD was recently used within various models and architectures in order to obtain the prediction uncertainty and improve the classification results (Miok, 2018; Miok et al., 2019c;a). Transformer networks were not analyzed yet. In our proposal, called Monte Carlo Dropout Bayesian Attention Networks (BAN or MCD AN) contrary to the original dropout setting, the dropout layers are active also during the prediction phase. In this way, predictions are not constant and are sampled from the *learned* distribution, thereby forming an ensemble of predictions. The obtained distribution can be, for example, inspected for higher moment properties and can offer additional information on the certainty of a given prediction. During the prediction phase, all layers except the dropout layers are deactivated. Forward pass on such partially activated architecture is repeated for a fixed number of samples, which can be combined to obtain the final probability, or further inspected as a distribution underlying the probability.

2.3. Monte Carlo Dropout for BERT

Monte Carlo dropout was used for the BERT predictions in the same way as for BAN. MCD can provide multiple predictions during the test time completely free, as long as the dropout was used during the training time (Gal, 2016). Training neural network with dropout distributes the information contained in the neurons throughout the network. Hence, during the prediction, such trained neural network will be robust; using the dropout principle a new prediction is possible in each forward pass, and sufficiently large set of such predictions can be used to estimate the reliability. BERT model is trained with 10% of dropout in all of the layers and thus allows for multiple predictions using the described principle. We call this model MCD BERT and it naturally provides reliability estimates. A possible limitation of this approach is that during training a single dropout rate of 10% is used, while other dropout probabilities might be more suitable for reliability estimation.

3. Evaluation Setting

We evaluate the proposed novelties concerning two main aspects: the calibration of returned probabilities, prediction performance and prediction uncertainty estimation. We first describe the hate speech data sets used, followed by the implementation details. In the last two subsections, we present evaluation measures for prediction performance and calibration.

3.1. Hate Speech Datasets

In order to test the proposed methodology in the multilingual context, we applied our models to three different data sets.

1. **English** data set[2] originates in a study regarding hate speech detection and the problem of offensive language (Davidson et al., 2017). Our data set consists of 5000 tweets. We took 1430 tweets labeled as hate speech and randomly sampled 3670 tweets from the collection of remaining 23353 tweets.

2. **Croatian** data set was collected by the Styria company within the EU Horizon 2020 project EMBEDDIA[3]. The text was extracted from the database of user comments, from the vecernji.hr[4] news portal. The original data set consists 9,646,634 comments described with 11 attributes from which we selected 8422 comments, one half of which were labelled as hate speech by human moderators, the other half was randomly chosen from the non-problematic comments.

3. **Slovenian** data set is a result of the Slovene national project FRENK[5]. The text data set used in the experiment was the combination of two different studies made on Facebook comments on the LGBT homophobia and anti-migrants published in the (Ljubešić et al., 2019). For the final data set we select 2182 hate and 2182 non-hate speech comments.

3.2. Prediction models

We used three types of prediction models: MCD LSTM networks (Miok et al., 2019b), MCD Bayesian Attention Networks (MCD AN) and MCD BERT. As the input to MCD LSTM we used pretrained word embeddings, sentence encoder for English (Cer et al., 2018) and fastText[6] for Slovenian and Croatian. For MCD AN we used simple tokenizer[7]. For the MCD BERT we used BERT's tokeizer. The summary is collected in Table 2.

3.3. Implementation details

We implemented the proposed MCD ANs in PyTorch[8]. The main hyperparameters of the architecture are the number of attention heads and the number of attention layers. The proposed adaptive classification threshold is computed after each validation set evaluation, i.e. every time we compute the performance on the validation set.‘

Other parameters are set as follows. We use the Adamax optimizer (Kingma & Ba, 2014), a variant of Adam based on infinity norm. Binary cross-entropy loss guides the training. In order to automatically stop training, we use the stopping step of 10 – if after 10 optimization steps the performance on the validation set is not improved, the training stops.

We explored the following hyperparameter tuning space: the validation percentage (size of validation set) was varied between 5% and 10%. The rationale for testing different percentages of validation set sizes is that the data considered is small, hence considering too high validation percentages could omit the classifier from viewing crucial instances and thus reduce its final performance. Given enough data, however, the percentage should be as high as possible. Number of epochs was either 30 or 100, number of hidden layers and attention heads was 1 or 2. Maximum padding of the input sequences was either 48, 32 or 64. Learning rate was either 0.001 or 0.0005 and the adaptive threshold was either enabled or disabled.

MCD LSTM networks consist of an embedding layer, LSTM layer, and a fully connected layer within the Word2Vec and ELMo embeddings. In order to obtain best architectures for the LSTM and MCD LSTM models, various number of units, batch size, dropout rates and so on were fine-tuned. For BERT implementation the BERT base was implemented for both English and multilingual versions using Hugging Face code [9].

[2] https://github.com/t-davidson/hate-speech-and-offensive-language

[3] http://embeddia.eu

[4] https://www.vecernji.hr

[5] “FRENK - Raziskave Elektronske Nespodobne Komunikacije” (engl. “Research on Electronic Inappropriate Communication”)

[6] https://fasttext.cc

[7] https://keras.io/preprocessing/text/

[8] https://gitlab.com/skblaz/bayesianattention

[9] https://huggingface.co/transformers/model_doc/bert.html

Table 1. Comparison of predictive models using sentence embeddings. We present average classification accuracy, precision, recall and F_1 score (and standard deviations), computed using 5-fold cross-validation. All the results are expressed in percentages and the best results for each language is typeset in bold.

	English Tweets				Croatian Comments				Slovenian Comments			
Model	**Accuracy**	**Precision**	**Recall**	**F1**	**Accuracy**	**Precision**	**Recall**	**F1**	**Accuracy**	**Precision**	**Recall**	**F1**
MCD LSTM	81.0 [1.2]	81.5 [1.8]	82.5 [2.7]	81.9 [1.3]	63.7 [1.0]	68.5 [1.2]	40.8 [4.0]	51.0 [3.3]	55.3 [0.69]	53.5 [4.27]	57.0 [9.55]	43.13 [0.8]
MCD AN	83.3 [1.7]	80.5 [3.47]	82.8 [3.9]	81.6 [3.4]	61.4 [2.0]	58.6 [9.3]	30.2 [11.0]	38.1 [8.6]	57.4 [1.7]	49.3 [3.7]	27.9 [7.8]	35.1 [6.3]
BERT	90.9 [0.7]	89.3 [1.4]	90.8 [1.3]	90.0 [0.7]	70.8 [1.0]	67.1 [1.9]	56.2 [2.0]	61.2 [1.5]	66.4 [5.0]	66.1 [6.8]	70.7 [5.5]	67.8 [2.5]
MCD BERT	**91.4 [0.7]**	**90.4 [1.5]**	**90.4 [0.6]**	**90.4 [0.8]**	**71.5 [1.2]**	**67.4 [2.3]**	**59.1 [3.6]**	**62.9 [1.7]**	**68.4 [1.9]**	**68.2 [0.8]**	**69.2 [2.9]**	**68.6 [1.6]**

Table 2. Characteristics of the used datasets: number of the tweets/comments and the embedding architecture used for each of the datasets.

Dataset	**Size**	**MCD LSTM**	**MCD AN**
English	5000	Sentence	Tokenizer
Croatian	8422	Fasttext	Tokenizer
Slovenian	4364	Fasttext	Tokenizer

4. Results

Results consists of three parts: calibration results, prediction performance, and visualization of uncertainty.

4.1. Prediction Performance

The results that compare 4 different models are presented in the Table 1. It can be observed that MCD BERT provide the best results for the all of the 3 data sets. As the MCD BERT is slightly better that BERT we can conclude for the tweets for which BERT is on borderline, multiple predictions can influence decision in the right direction.

With intention to statistically test if MCD BERT could indicate problematic predictions we investigate 1000 test tweets splitting them on the *confused* and *certain*. As the BERT generally provide very extreme predictions the criteria was: the test tweet is confusing if the variance computed from 1000 predictions is greater then 0.1 otherwise it was classified as certain. In the Table 3 two by two contingency results are presented for each of the three languages data sets. The Chi-square test for the English MCD BERT results was found to be very significant with p-value= 2.2e-16. The Chi-square test for BERT model results was found to be less significant with p-value = 1.384e-11. For CRO BERT the Chi-square test was not significant with p-value= 1 so it is clear that we can not classify tweets based just on the probability. On the other hand, for the CRO MCD BERT the criteria based on the variance > 0.1 provide better spit so the Chi-square test become significant with p-value= 8.348e-16. The p-values for the SLO BERT and SLO MCD BERT are 0.0037 and 0.0002 respectively. Also, based on the ratios between mistake and no mistake it can be observed that number of true mistakes in the confused group is high for MCD BERT.

Table 3. Two-by-two contingency table for Certain/Confused vs Mistake yes/no.

Language	**Mistake**	**BERT**		**MCD BERT**	
		Certain	**Confused**	**Certain**	**Confused**
EN	**No**	880	31	891	24
	Yes	71	18	62	23
Ratio		0.08	0.58	0.06	0.95
CRO	**No**	1176	35	1053	152
	Yes	461	14	336	139
Ratio		0.39	0.4	0.31	0.91
SLO	**No**	576	28	537	55
	Yes	241	27	229	51
Ratio		0.42	0.96	0.42	0.92

From those results it can be concluded that the MCD BERT provides better understanding of the how much we can trust our predictions compared to the simple BERT.

5. Conclusions

In practical setting, automatic detection of hate speech not only requires high precision but also prediction uncertainty estimates. In times when social networks suffer from high amount of offensive messages, wrong classifications can damage the minorities, lower the level of democratic debate but also damage the freedom of speech. In technological terms, natural language approaches are witnessing a switch from recurrent neural networks with pretrained word embeddings to large pretrained transformer models, BERT being the best example of this. We introduce the Monte Carlo dropout into attention layers of transformer neural networks as a tool for prediction uncertainty estimation. We demonstrate the methodology on the hate speech detection task.

The results of our empirical evaluation show that MCD can improve BERT results regarding both the prediction performance and uncertainty estimation. For all three languages hate speech datasets, the MCD enhanced BERT and mBERT preformed best. Further, we show that MCD BERT reliability scores provide information on the trusted and dubious predictions. This information can significantly reduce the amount of work in reannotation of questionable cases.

References

Bleich, E. The rise of hate speech and hate crime laws in liberal democracies. *Journal of Ethnic and Migration Studies*, 37(6):917–934, 2011.

Cer, D., Yang, Y., Kong, S.-y., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., et al. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*, 2018.

Chang, W.-C., Yu, H.-F., Zhong, K., Yang, Y., and Dhillon, I. X-bert: extreme multi-label text classification with bert. *arXiv preprint arXiv:1905.02331*, 2019.

Davidson, T., Warmsley, D., Macy, M., and Weber, I. Automated hate speech detection and the problem of offensive language. In *Eleventh international aaai conference on web and social media*, 2017.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Gal, Y. Uncertainty in deep learning. *University of Cambridge*, 1:3, 2016.

Gal, Y. and Ghahramani, Z. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, pp. 1050–1059, 2016.

Gururangan, S., Dang, T., Card, D., and Smith, N. A. Variational pretraining for semi-supervised text classification. *arXiv preprint arXiv:1906.02242*, 2019.

Izmailov, P., Maddox, W. J., Kirichenko, P., Garipov, T., Vetrov, D., and Wilson, A. G. Subspace inference for bayesian deep learning. *arXiv preprint arXiv:1907.07504*, 2019.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Ljubešić, N., Fišer, D., and Erjavec, T. The frenk datasets of socially unacceptable discourse in slovene and english. In *International Conference on Text, Speech, and Dialogue*, pp. 103–114. Springer, 2019.

McInnes, L., Healy, J., and Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Miok, K. Estimation of prediction intervals in neural network-based regression models. In *20th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC)*, pp. 463–468, 09 2018.

Miok, K., Nguyen-Doan, D., Robnik-Šikonja, M., and Zaharie, D. Multiple imputation for biomedicaldata using monte carlo dropout autoencoders. In *7th IEEE International Conference on E-Health and Bioengeneering (EHB)*. IEEE, 2019a.

Miok, K., Nguyen-Doan, D., Škrlj, B., Zaharie, D., and Robnik-Šikonja, M. Prediction uncertainty estimation for hate speech classification. In *International Conference on Statistical Language and Speech Processing*, pp. 286–298. Springer, 2019b.

Miok, K., Nguyen-Doan, D., Zaharie, D., and Robnik-Šikonja, M. Generating data using monte carlo dropout. In *2019 IEEE 15th International Conference on Intelligent Computer Communication and Processing (ICCP)*, pp. 509–515. IEEE, 2019c.

Myshkov, P. and Julier, S. Posterior distribution analysis for bayesian inference in neural networks. *Advances in Neural Information Processing Systems (NIPS)*, 2016.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008, 2017.

Xu, Y., Qiu, X., Zhou, L., and Huang, X. Improving bert fine-tuning via self-ensemble and self-distillation. *arXiv preprint arXiv:2002.10345*, 2020.

A. Calibration of BAN and BERT

Figure 2 shows how calibration of prediction scores change during training of AN. The red line represents the performance of the fully trained network. It is apparent that additional calibration is necessary – the dotted line represents perfect calibration. Surprisingly, initial training iterations show better calibrated scores. This can be due to the definition of ECE measure: in case that both accuracy and predicted scores are low, this would lead to low ECE value.

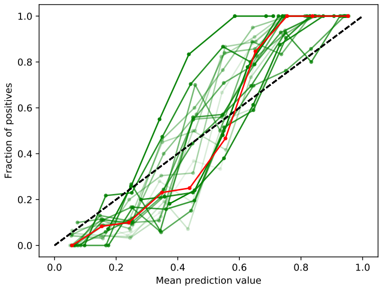


Figure 2. Calibration plot for MCD AN after each epoch (green) based on the validation set of the best performing architecture. The more transparent the calibrations the earlier the training stage (fewer epochs). The final calibration is in red.

In the Tables 4, 5, and 6 calibration results for different calibration approaches of MCD AN are presented: no calibration, isotonic regression, and Platt's method, combined with the adaptive threshold or not. It can be observed that for all three languages both calibration methods improve the ECE score, and Platt's method seems to produce the best calibration scores. Adaptive threshold slightly improves the ECE score for the uncalibrated (raw) results. This is especially true for the Slovenian comments where the ECE score was reduced from the 0.794 to the 0.621. Nevertheless, we can conclude that calibration using adaptive threshold heuristics is beneficial but cannot be compared with the improvements brought by proper calibration techniques.

In order to compare the calibration results for different BAN and BERT architectures, we plotted their ECE scores in Figure 3. It can be observed that calibration methods substantially improve the MCD AN calibration; however, the BERT model is even better calibrated.

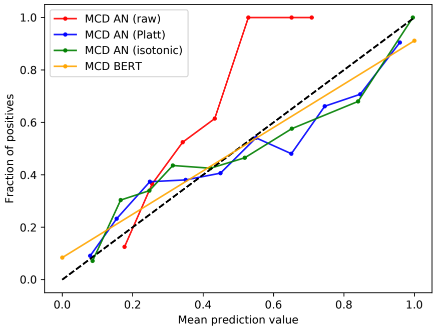


Figure 3. Calibration plots based on English test set performance for BERT and MCD AN architecture using different calibration algorithms.

Table 4. Calibration scores of MCD AN with different calibration approaches on English tweets. The results are presented based on whether they were calibrated and whether the adaptive threshold (AT) was applied

Calibration	AT	Accuracy	F1	ECE
Raw	False	0.83 (0.02)	0.82 (0.03)	0.547
Raw	True	0.83 (0.01)	**0.83** (0.041)	0.54
Isotonic	False	**0.84** (0.01)	0.82 (0.01)	0.230
Isotonic	True	0.83 (0.01)	0.82 (0.02)	0.234
Platt's	False	0.84 (0.02)	0.82 (0.02)	**0.225**
Platt's	True	0.83 (0.01)	0.82 (0.01)	0.232

Table 5. Calibration scores of MCD AN with different calibration approaches on Croatian user news comments. The results are presented based on whether they were calibrated and whether the adaptive threshold (AT) was applied

Calibration	AT	Accuracy	F1	ECE
Raw	False	0.61 (0.02)	0.47 (0.03)	0.681
Raw	True	0.62 (0.02)	**0.50** (0.04)	0.663
Isotonic	False	0.60 (0.01)	0.49 (0.04)	0.206
Isotonic	True	0.61 (0.01)	0.50 (0.03)	0.206
Platt's	False	0.61 (0.02)	0.48 (0.02)	0.198
Platt's	True	**0.62** (0.02)	0.49 (0.02)	**0.197**

Table 6. Calibration scores of MCD AN with different calibration approaches on Slovenian Facebook comments. The results are presented based on whether they were calibrated and whether the adaptive threshold (AT) was applied

Calibration	AT	Accuracy	F1	ECE
Raw	False	0.59 (0.01)	0.33 (0.05)	0.794
Raw	True	0.59 (0.02)	0.48 (0.05)	0.621
Isotonic	False	0.58 (0.02)	0.49 (0.03)	0.212
Isotonic	True	0.58 (0.02)	**0.49** (0.03)	0.213
Platt's	False	0.58 (0.03)	0.48 (0.02)	0.206
Platt's	True	**0.59** (0.02)	0.47 (0.04)	**0.204**

B. Additional reliability graphs

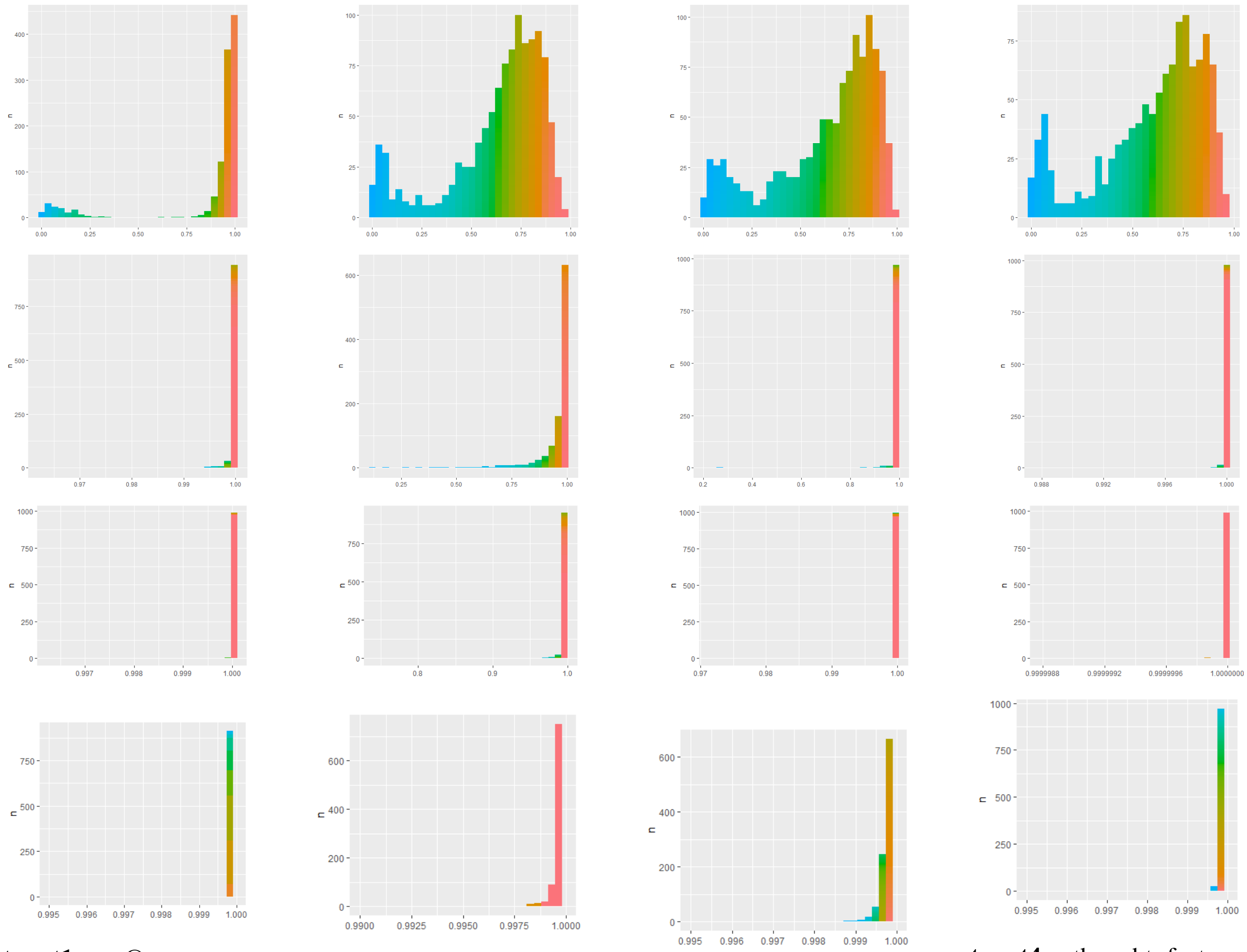
B.1. Visualization of Uncertainty

Obtaining multiple predictions for a specific test tweet can improve understanding of the final prediction. The mean of the distribution is used to estimate the probability and the variance informs us about the spread and certainty of the prediction. We can inspect the actual distribution of prediction scores with histogram plots as demonstrated on Figures 4 and 6 for a few correctly classified instances and on Figures 5 and 7 for a few misclassified instances.

Histograms presented in the Figures 4 and 5 for English tweets and Figures 6 and 7 for Croatian comments visually display the prediction certainty of the specific tweet or comment. It can be observed that results for BERT are going to the extremes, especially for the predictions when model seems to be sure. Also, it can be observed that AN architecture with 10% of dropout provide the similar spread of values as the BERT. On the other hand, introducting 30 % of dropout in AN for examples where attention network model is not sure will influence the spread and make the prediction more uncertain.

Apart from visualization text tweet separately the multiple prediction provide opportunity for understanding the contextual dependencies with the other test tweets. Following (Miok et al., 2019b) we visualize the embeddings of the output. The key idea of the visualization can be summarized as follows. First, 1000 samples are obtained for each prediction. The space of such distributions across individual test-set texts is next *embedded* into two dimensions by using Uniform Manifold Projections method (McInnes et al., 2018). this way, a two dimensional space corresponding to the initial 1000 dimensional space of predictions is obtained. Next, Gaussian kernel estimation is used to identify equivalent regions, which are connected with closed curves. Finally, the shapes and sizes of individual predictions are adapted based on the classification error and certainty of a given prediction. The goal of using such visualization is to discover potentially larger structures within the space of emitted probabilities, potentially offering insights into the given probabilistic neural network's drawbacks and limitations. The results of such visualization are shown in Figures 8 and 9. In Figure 8 the plot displays the position of the certain and uncertain test tweets in the latent space while in the Figure 9 the differences based on the mean probability are displayed.

It can be observed that (Figure 8) after only a few epoch of training, majority of the prediction are in the middle layer (yellow) that corresponds to the predictions that are uncertain (high spread of the whole predictive distribution). On the contrary, in Figure 9, where the same learning setting was considered for 50 epochs, the probability space distinctly separates into two main *components*, indicating that there are predictions for which the neural network is certain (and were indeed correct), however for some predictions, especially related to the instances that are not hate speech, the network is less certain (albeit still correct). The two examples demonstrate how the space of probabilities *separates* into distinct components once the neural network is trained. The visualizations also indicate that some of the instances are more problematic than others, potentially facilitating the debugging process for a developer (and inspection of convergence).



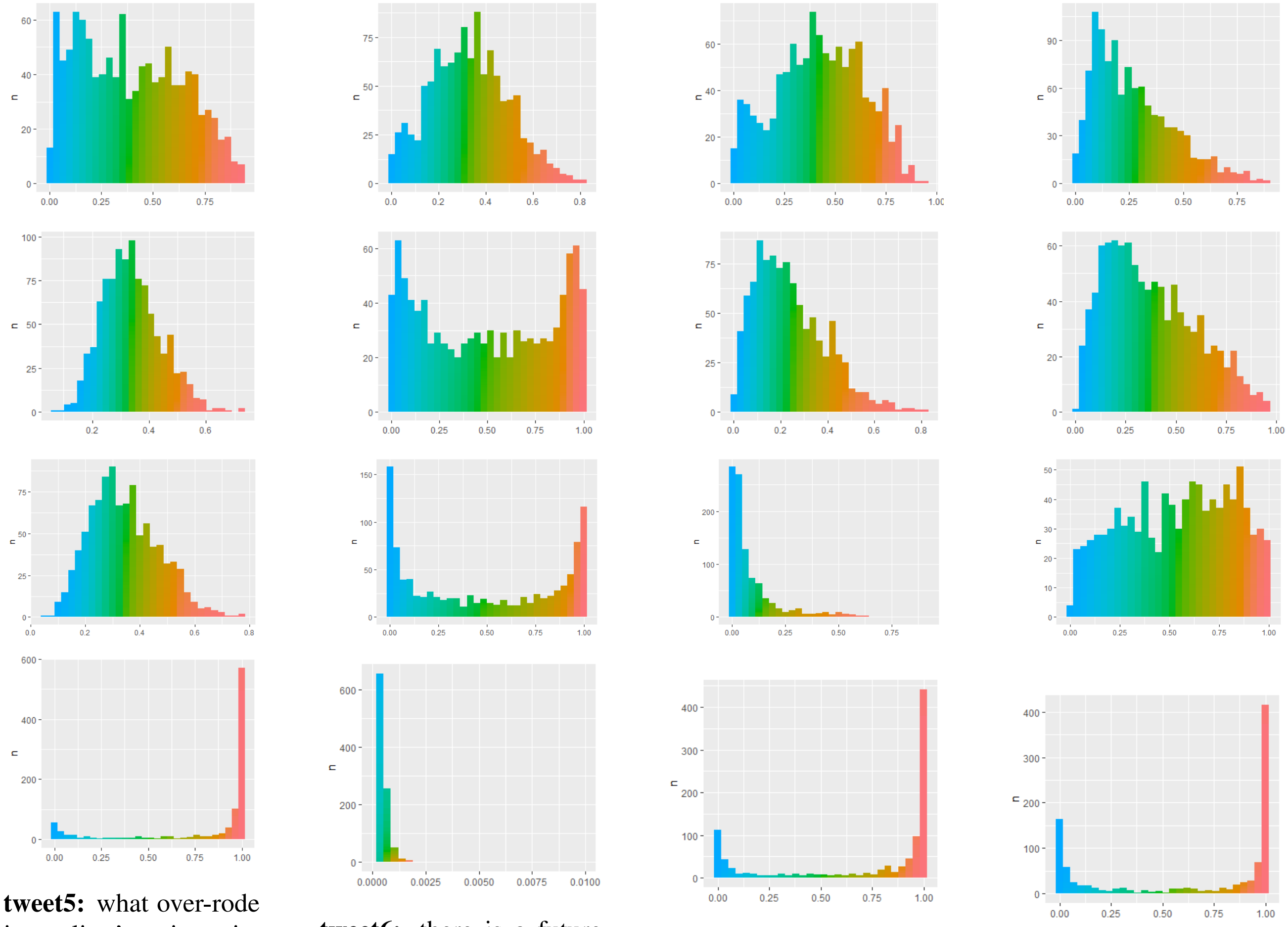
tweet1: @user you might be a libtard if... #libtard #sjw #liberal #politics

tweet2: #paid #kkk to #fabricate #stories to push the producers #narrative #cancel

tweet3: carl paladino, trump ally, wishes obama dead of mad cow disease

tweet4: thought factory: left-right polarisation! #trump #uselections2016 #leadership #politics #brexit #blm

Figure 4. English data set - Comparing the multiple prediction distributions for MCD LSTM (first row), MCD AN with 30% dropout (second row), MCD AN with 10% dropout (third row) and MCD BERT (fourth row) for 4 test tweets where hate speech was **correctly** predicted. Note that the x axis showing predicted probability distributions are different for each tweet. Results of BERT are concentrated in much narrower interval compared to MCD LSTM and MCD AN.



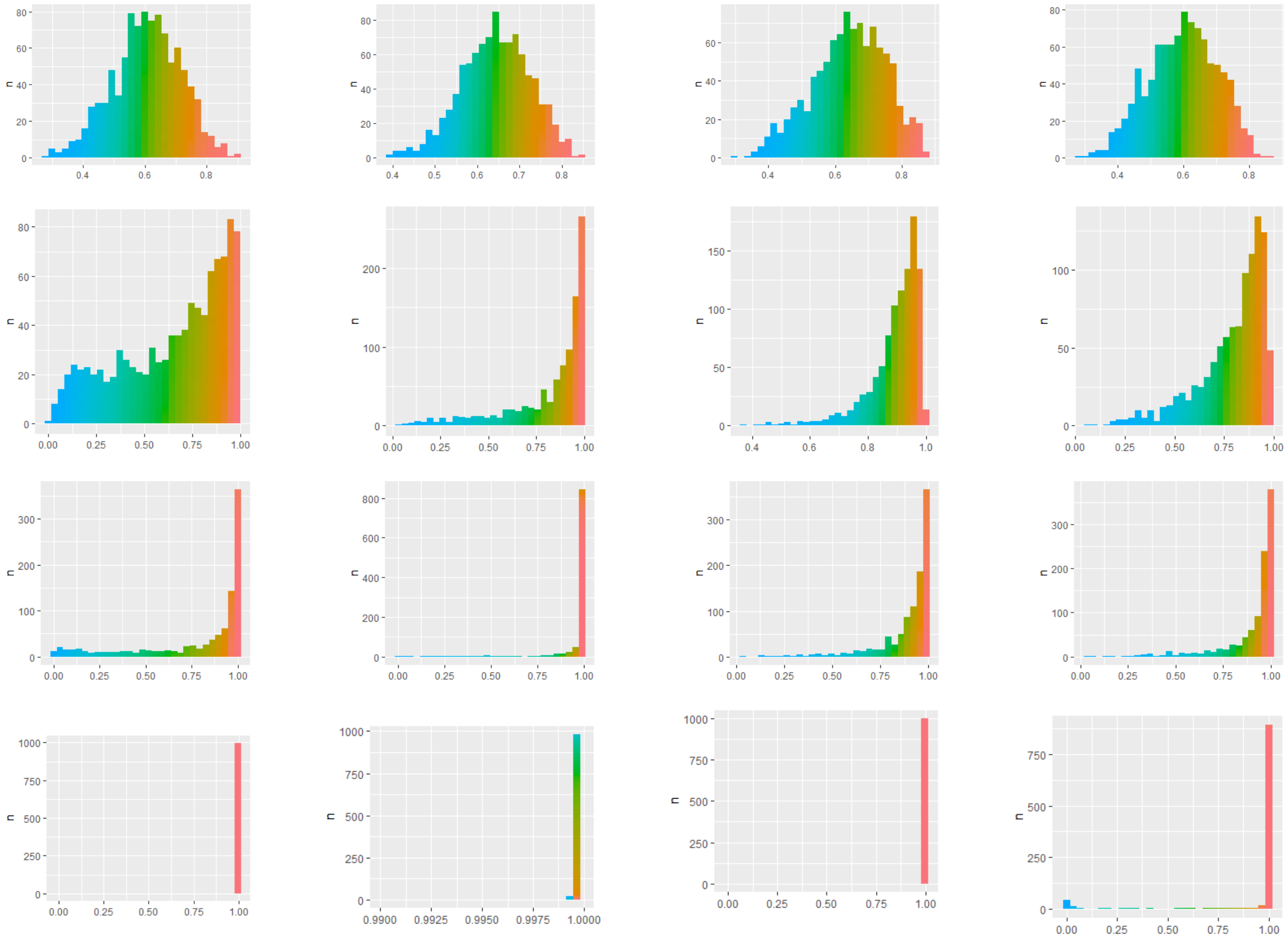
tweet5: what over-rode journalists' integrity was greed and ambition, along with a total absence of courage (true label = hate speech)

tweet6: there is a future nina and right now you are pretty much throwing yours away. (true label = non hate speech)

tweet7: @user @user someone really respects eu guidelines!!! (true label = non hate speech)

tweet8: the scars left by us waime camps mt @user (true label = hate speech)

Figure 5. English data set - Comparing the multiple prediction distributions for MCD LSTM (first row), MCD AN with 30% dropout (second row), MCD AN with 10% dropout (third row) and MCD BERT (fourth row) for 4 test tweets where hate speech was not clearly predicted. Note that the x axis showing predicted probability distributions are different for each tweet. Results of BERT are concentrated in much narrower interval compared to MCD LSTM and MCD AN.



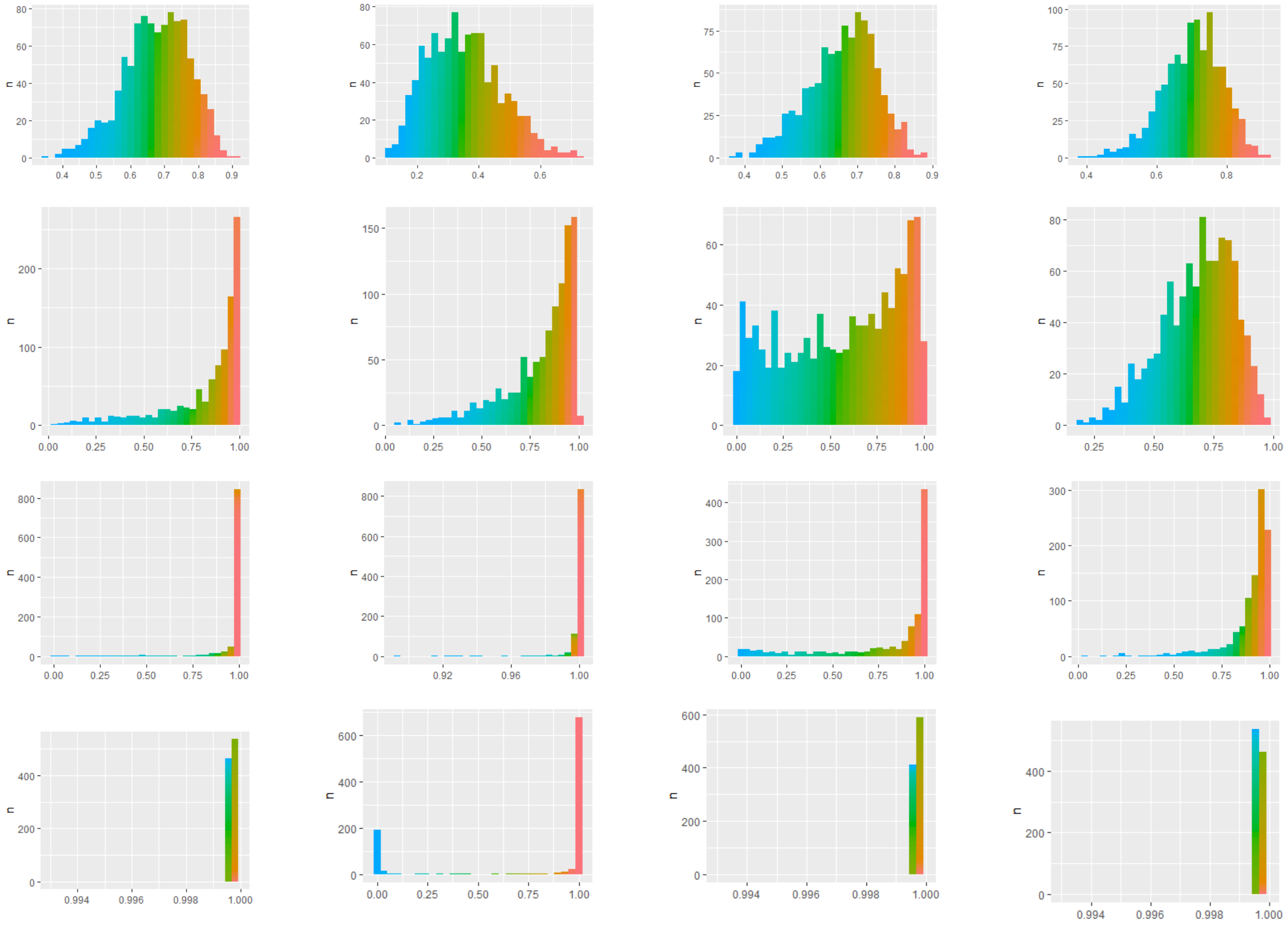
81 - kamo sreće da jeste...al mi u HR nismo imali pojam ”rodjo” dok se vi niste naselili ovdje

17 - Drazi Mihailovicu je podignut spomenik i u Usa.Kada vas Pavelic dobije spomenik,bilo gde u svetu,onda se javite.Istina boli vise od batina,zar ne?

2 - prvo treba najuriti pola sabora na celu sa cedom pupijo . poz... od HOSa ZDS

24 - I mi srbi volimo hrvatice, jedva cekam vikend i njihov dolazak u Bg.

Figure 6. Croatian data set - Comparing the multiple prediction distributions for MCD LSTM (first row), MCD AN with 10% dropout (second row), MCD AN with 30% dropout (third row) and MCD BERT (fourth row) for 4 test tweets where hate speech was not clearly predicted. Note that the x axis showing predicted probability distributions are different for each tweet. Results of BERT are concentrated in much narrower interval compared to MCD LSTM and MCD AN.



18 - Eh vi Hrvati vas problem je sto vi sve nas Srbe smatrate cetnicima,ima nas i komunista,a i pacifista,ja sam pacifisticki nastrojen zbog toga kazem Pruzimo Miru sansu,i ne dirajmo dvojezicne table jer su one simbol jedinstva SrboHrvata! (true label = non hate speech)

31 - mali sloba na reveru nosi zastavu srbije a naš milanče je bio neodlučan– (true label = non hate speech)

56 - Rubinet@ / Pa bas i da ste slobodni i niste.Jednog gospodara Beograd zamenili ste sa novim Brisel,kao kroz citavu vasu povijest.Hej jos se i dice time da imaju gospodara ! (true label = hate speech)

19 - Posledica suživota sa nama?ahhhhahhh ćuti bre konjušaru (true label = hate speech)

Figure 7. Croatian data set - Comparing the multiple prediction distributions for MCD LSTM (first row), MCD AN with 30% dropout (second row), MCD AN with 10% dropout (third row) and MCD BERT (fourth row) for 4 test tweets where hate speech was not clearly predicted. Note that the x axis showing predicted probability distributions are different for each tweet. Results of BERT are concentrated in much narrower interval compared to MCD LSTM and MCD AN.

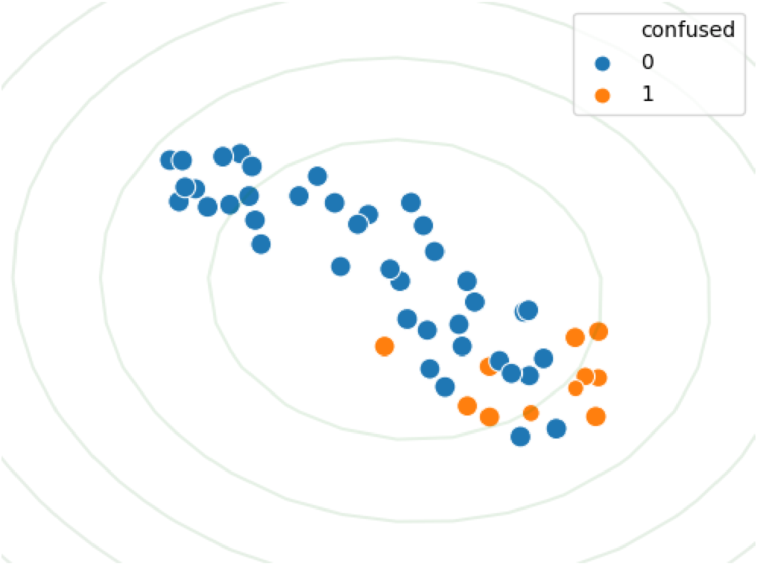


Figure 8. Visualization of the 100 test tweets in two dimensions. Tweets that are found to be certain are colored in blue (0) while tweets that are confused in orange (1). It can be observed that uncertain tweets get clustered.

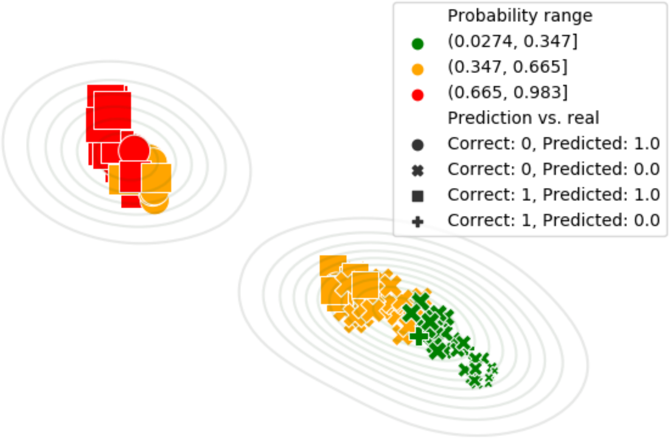


Figure 9. Visualization of the outcome probability space for 100 tweets from the test set. The test tweets are colored in the green, yellow and red depending to which interval belongs the mean probability of the 1000 predictions. It can be observed that the predictions with very high confidence form an isolated part of the probability space.