Bayesian Methods for Semi-supervised Text Annotation

Kristian Miok[1,2], **Gregor Pirš**[1] **and Marko Robnik-Šikonja**[1]

[1] University of Ljubljana, Faculty of Computer and Information Science, Slovenia
Email: {gregor.pirs, marko.robnik}@fri.uni-lj.si
[2] West University of Timisoara, Computer Science Department, Romania
Email: kristian.miok@e-uvt.ro

Abstract

Human annotations are an important source of information in the development of natural language understanding approaches. As under the pressure of productivity annotators can assign different labels to a given text, the quality of produced annotations frequently varies. This is especially the case if decisions are difficult, with high cognitive load, requires awareness of broader context, or careful consideration of background knowledge. To alleviate the problem, we propose two semi-supervised methods to guide the annotation process: a Bayesian deep learning model and a Bayesian ensemble method. Using a Bayesian deep learning method, we can discover annotations that cannot be trusted and might require reannotation. A recently proposed Bayesian ensemble method helps us to combine the annotators' labels with predictions of trained models. According to the results obtained from three hate speech detection experiments, the proposed Bayesian methods can improve the annotations and prediction performance of BERT models.

1 Introduction

Recent successful applications of artificial intelligence in various fields, including natural language processing, are often due to long hours of human annotation when preparing datasets for machine learning. The annotation process transfers human knowledge to machine learning models but it is often done under time pressure and with inadequate instructions or with insufficiently trained annotators.Aiming to make the annotation process easier, we study the possibility of designing a data labeling process which requires less human supervision.

In practice, a fairly standard procedure in the annotation quality control is to recheck the labels that are wrongly classified by using several prediction models. As an alternative, Bayesian inference produces a distribution of possible decisions and can improve the selection of instances requiring reannotation (Miok et al., 2020). Most neural networks do not support the assessment of predictive uncertainty. The Bayesian inference framework can be helpful, however, most techniques do not scale well in neural networks with high dimensional parameter space (Izmailov et al., 2019). Various methods were proposed to overcome this problem (Myshkov and Julier, 2016), one of the most efficient being Monte Carlo Dropout (MCD) (Gal and Ghahramani, 2016a). Its idea is to use the dropout mechanism in neural networks as a regularization technique (Srivastava et al., 2014) and interpret it as a Bayesian optimization approach that samples from the approximate posterior distribution.

A common problem in text annotations is that annotators are not always sure about correct labels due to uncertainty in the text (Vincze, 2015; Szarvas et al., 2008). On difficult texts, annotators frequently give ambiguous labels and their annotations can be biased. Instead of asking annotators to label the raw text, it would be easier for them if they were proposed answers accompanied by probabilistic scores from an ensemble of predictive models. Ensemble methods produce robust models that frequently provide significantly better predictions than individual models. The key strength of ensembles is that they can overcome errors and shortcomings of individual ensemble members. However, diversity in combining different predictions and reliability of individual predictions need to be better understood and

The 14th Linguistic Annotation Workshop, pages 1–12
Barcelona, Spain, December 12, 2020.

evaluated (Zhou, 2012). A recently published ensemble method Multivariate Normal Mixture Conditional Likelihood Model (MM) (Pirš and Štrumbelj, 2019) tries to understand the predictors on the distributional level and use Bayesian inference to combine them. In this work, we evaluate MM's performance when combining predictive models on the hate speech detection task. We show that our methodology can serve as a helpful tool in the data annotation process.

Recently, the most successful approach in text classification is to use transformer neural networks (Vaswani et al., 2017), pretrained on large monolingual corpora, and then fine-tune them for a specific task, such as text classification. For example, BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2019) uses masked language modeling and order of sentences prediction tasks to build a general language understanding model. During the fine-tuning for a specific downstream task, additional layers are added to the BERT model, and the model is trained on the data of interest to capture the specific knowledge required to perform the task.

The main aims of the paper is to propose methods that can save time and resources during the text annotation process and improve prediction performance. As a test domain we use hate speech detection in tweets, news comments and Facebook comments. We investigate two performance improving techniques which can be summarized as our main contributions as follows.

1. We remove instances with uncertain classifications from the training set and show that fine-tuning on the cleaned dataset improves the performance of the BERT model. Less certain classifications can be selected for reannotation.

2. We combine predictions of machine learning models using the MM probabilistic ensemble method. The approach is beneficial for predictive performance.

The paper consists of five further sections. In Section 2, we present related works on prediction uncertainty and hate speech detection. In Section 3, we propose the methodology for uncertainty assessment of deep neural networks using attention layers and MCD. In Section 4, we describe the tested datasets and evaluation scenarios. The obtained results are presented in Section 5, followed by conclusions and ideas for further work in Section 6.

2 Related Work

In this section, we introduce related work split into four topics. First, we present the work on semi-supervised learning that can be used in text annotation, followed by the related research on Bayesian learning for text classification. In the third subsection, we describe probabilistic ensemble methods and in the fourth, we outline the related work on hate speech detection.

2.1 Semi-supervised Learning for Text Annotation

The performance of supervised learning depends on the availability of a sufficient amount of labeled data. However, manual labeling is expensive and difficult to scale up to large amounts of data. Semi-supervised learning tries to utilize large amounts of unlabeled data available for many problems by combining them with small amounts of labeled data (Zhu, 2005). The goal of semi-supervised learning is to understand how combining labeled and unlabeled data can change the learning behavior, and design algorithms that take advantage of such a combination (Zhu and Goldberg, 2009). Most semi-supervised learning strategies extend either unsupervised or supervised learning to include additional information typical of the other learning paradigm. The transductive learning is related to the semi-supervised learning, but assumes that the test set is known in advance and its goal is to optimize the generalization ability on this (unlabeled) test set (Zhou and Li, 2010). In the non-transductive setting, Acharya et al. (2013) combine probabilistic classifiers. They take class labels from existing classifiers and cluster labels from a clustering ensemble. The consensus labeling is assigned to the target data.

2.2 Bayesian Methods for Text Classification

Although, recent works on prediction uncertainty mostly investigate deep neural networks, many other probabilistic classifiers were analyzed in the past (Platt, 1999; Niculescu-Mizil and Caruana, 2005; Zhang

et al., 2013; Cao et al., 2015; He et al., 2018). Prediction reliability is an important issue for black-box models like neural networks as they do not provide interpretability or reliability information about their predictions. Most existing reliability scores for deep neural networks are contructed using Bayesian inference. The most popular exception is the work of Lakshminarayanan et al. (2017), who proposed to use deep ensembles to estimate the prediction uncertainty.

A computationally efficient simulation of Bayesian inference uses Monte Carlo dropout (Gal and Ghahramani, 2016a). The first implementation of dropout in recurrent neural networks (RNNs) was in 2013 (Wang and Manning, 2013) but further research revealed a negative impact of dropout in RNNs (Bluche et al., 2015). Later, the dropout was successfully applied to language modeling by Zaremba et al. (2014) who used it only in fully connected layers. Gal and Ghahramani (2016b) implemented the variational inference based dropout which can regularize also recurrent layers. In this way method mimics Bayesian inference by combining probabilistic parameter interpretation and deep RNNs. Several other works investigate how to estimate prediction uncertainty within different data frameworks using RNNs (Zhu and Laptev, 2017; Miok et al., 2019b), e.g., Bayes by Backpropagation (BBB) was applied to RNNs (Fortunato et al., 2017). Monte Carlo dropout was also introduced into variational autoencoders (Miok et al., 2019a; Miok et al., 2019c) and for estimating prediction intervals (Miok, 2018).

To our knowledge, Bayesian deep learning models were not yet used to detect less certain text classifications and remove them from a train dataset to improve the prediction performance.

2.3 Probabilistic Ensembles

Most methods used for text classification can produce probabilistic predictions which are rarely exploited beyond classification into a discrete class. As probabilistic predictions provide additional information compared to the discrete outcome, we use ensembles that can model predictive distributions. Ensemble methods can be divided into two main groups. The first group of methods estimates the performance of individual classifiers and weights them accordingly. The second group of methods learns the structure of predictions and bases their forecasts on it.

The first group of methods can be further divided into methods that are able to combine full posterior distributions and methods that only combine probabilistic point predictions. The advantage of the former is that they are more expressive, and a disadvantage is that they require inputs in the form of a full distribution, which is not always available. Bayesian model averaging (Hoeting et al., 1999) combines models by their marginal posterior probability. This method is suitable if one of the candidate models is the true data generating process, otherwise its performance decreases (Cerquides and De Mántaras, 2005). Bayesian stacking (Yao et al., 2018) is also useful in these cases. It is based on weighing the posterior predictive distributions of individual models by estimating their leave-one-out cross-validation performance. Linear opinion pool (Cooke, 1991) is a classical approach to combine classifiers and can be included into the second group of methods. It combines predictions as a linear combination of individual models by maximizing the likelihood. An example of Bayesian approach from the second group is the agnostic Bayesian learning of ensembles (Lacoste et al., 2014), which weighs the models by estimated probabilities of them being the best model; the estimates are based on holdout computation of generalization performance.

Methods that model the structure of predictions are especially useful in case of complex relationships between individual models' predictions and the response variable. Independent Bayesian classifier combination (IBCC) (Kim and Ghahramani, 2012) combines non-probabilistic predictions by estimating the probability mass of predictions with a categorical distribution, conditional on the true label. It provides probabilities for a new observation that are proportional to the probability mass of new inputs for each true label. Nazábal et al. (2016) has extended the IBCC to probabilistic predictions by using the Dirichlet distribution. Supra-Bayesian methods (Lindley, 1985) combine probabilistic predictions using the log-odds of probabilities and modeling them with the multivariate normal (MVN) distribution, conditional on the true label. They use the common covariance matrix over all true labels but vary the means.

Ensemble modeling has recently been studied also within the text classification area (Li et al., 2018; Silva et al., 2010; Kilimci and Akyokus, 2018), but not within the context of probabilistic ensemble

models. In our work, we investigate Bayesian ensemble modeling for hate speech detection and how this can improve individual model predictions.

2.4 Hate Speech Detection

Analyzing sentiments and extracting emotions from the text are useful natural language processing applications (Sun et al., 2018). Being one of the wide range of applications where machines tend to understand human sentiments, hate speech detection is gaining importance with the rise of social media. We regard hate speech as written or oral communication that abuses or threatens a specific group or target (Warner and Hirschberg, 2012).

Detecting abusive language for less-resourced languages is difficult, hence, multilingual and cross-lingual methods are employed to improve the results (Stappen et al., 2020). This is especially the case when the involved languages are morphologically or geographically similar (Pamungkas and Patti, 2019). In our work, we investigate and compare hate speech detection methods for English, Croatian, and Slovene. English is by far the most researched language with plenty of resources (Malmasi and Zampieri, 2017; Davidson et al., 2017; Waseem and Hovy, 2016). Recently, hate speech detection studies were done also on neighbouring Slavic languages Croatian (Kocijan et al., 2019; Ljubešić et al., 2018) and Slovene (Fišer et al., 2017; Ljubešić et al., 2019; Vezjak, 2018).

Hate speech detection is usually treated as a binary text classification problem, and is approached with supervised learning methods. In the past, the most frequently used classifier was the Support Vector Machine (SVM) method (Schmidt and Wiegand, 2017), but recently deep neural networks showed superior performance, first through recurrent neural networks (Mehdad and Tetreault, 2016), and recently using large pretrained transformer networks (Mozafari et al., 2019; Wiedemann et al., 2020). In this work, we use the recent state-of-the-art pretrained (multilingual) BERT model.

3 Methods

We describe two approaches to the assessment of prediction reliability, Bayesian Attention Networks and Bayesian Probabilistic Ensembles.

3.1 Bayesian Attention Networks

The work (Miok et al., 2020) that introduce method named 'Bayesian Attention Networks' (BAN), proposes the dropout layers to be active also during the prediction phase. In this way, predictions are rather random 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 it can offer additional information on the certainty of a given prediction. During the prediction phase, all layers of the network except the dropout layers are deactivated. The forward pass on such partially activated architecture is repeated for a fixed number of samples, each time producing a different outcome that can be combined into the final probability, or inspected as a probability distribution.

Monte Carlo dropout was adapted for the BERT model in the same way as for BANs. MCD can provide multiple predictions during the test time without any additional training (Gal, 2016). Training a neural network with dropout spreads the information contained in the neurons across the network. Hence, during the prediction, such a trained neural network will be robust; using the dropout principle, a new prediction is created in each forward pass, and a sufficiently large set of such predictions can be used to estimate prediction reliability. The BERT models are trained with 10% of dropout in all of the layers by default. Therefore, it allows for multiple predictions with the fine-tuned model. We call this model MCD BERT. 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. We leave this question for further work as it requires long and costly training of several BERT models.

3.2 Bayesian Probabilistic Ensemble

To alleviate the drawbacks of individual classification models, we propose the use of MM (Pirš and Štrumbelj, 2019), a Bayesian ensemble method suitable for combining correlated probabilistic predictions.

MM is an extension of IBCC (Kim and Ghahramani, 2012), which combines non-probabilistic predictions. The method is based on finding the latent structure of combined predictions and provides new probabilities based on its distribution. Let m be the number of classes and r the number of individual models we are combining. The main idea is similar to Supra-Bayesian ensembles (Lindley, 1985), as we first transform individual probabilistic predictions with the inverse logistic transformation (log-odds) to move from [0,1] space to the $\mathbb{R}$ space. We merge the transformed predictions of individual models and get a $(m-1)r$-variate distribution. We model this latent distribution with multivariate normal mixtures, conditional on the true label in a similar fashion as in the case of linear discriminant analysis. Let θ represent estimated parameters and θ_t the subset of parameters estimated for observations with true label t. Let $T^* \in \{1, 2, ..., m\}$ be the response random variable for a new observation and $u^* \in \mathbb{R}^{(m-1)r}$ the transformed and merged predictions for this new observation. Probabilistic predictions for unseen data can then be generated by calculating the densities of merged predictions for new data:

$$p(T^* = t|u^*, \theta) = \frac{p(u^*|\theta_t)(\gamma_t n_t)}{\sum_{i=1}^{r} p(u^*|\theta_i)(\gamma_i n_i)},$$

where p is the MVN mixture probability density, γ_t is the frequency prior for class t, and n_t is the number of true labels in class t in the training dataset. The method uses a regularization term, which increases the variance in any dimension that is difficult to model or has a detrimental effect on the results, effectively decreasing its effect. For a complete Bayesian specification and the derivation of the Gibbs sampler, we refer the reader to (Pirš and Štrumbelj, 2019). We used the same priors as proposed in this paper.

MM is well-suited for combining biased classifiers, or classifiers with systematic errors. It can serve as a calibration tool for an individual classifier by learning its latent distribution. Since BERT is usually accurate but less well calibrated, the MM method has the potential to alleviate miscalibration, while improving or at least preserving the classification performance.

4 Experimental Setting

We first introduce the three phases of our experiments, followed by the used datasets and implementation details. The experimental setting consists of three phases:

1. We categorize classifications to trusted and untrusted based on the uncertainty measure from MCD BERT. In this way, we can detect borderline classification that make a false impression of certainty.

2. We remove the instances with uncertain classifications from the *training set* to improve the dataset on which the BERT model is fine-tuned. This provides better quality data for training and shall improve the quality of the resulting prediction model.

3. We use Bayesian ensemble to combine automatic predictions with annotators' decisions to remove low-quality training instances.

4.1 Datasets

To test the proposed methodology in the multilingual context, we trained the presented classification models on three different datasets, summarized in Table 1.

1. The **English** dataset[1] is extracted from the hate speech and offensive language detection study of Davidson et al. (2017). We used the subset of data consisting of 5,000 tweets. We took 1,430 tweets labeled as hate speech and randomly sampled 3,670 tweets from the collection of the remaining 23,353 tweets.

2. The **Croatian** dataset was provided by the Styria media company within the EU Horizon 2020 EMBEDDIA project[2]. The texts were extracted from user comments in the news portal Večernji list[3].

[1] https://github.com/t-davidson/hate-speech-and-offensive-language
[2] http://embeddia.eu
[3] https://www.vecernji.hr

The original dataset consists of 9,646,634 comments from which we selected 8,422 comments of which 50% are labeled as hate speech by human moderators and the other half was randomly chosen from the non-problematic comments.

3. **Slovene** dataset is a result of the Slovenian national project FRENK[4]. Our dataset comes from two studies on Facebook comments (Ljubešić et al., 2019). The first study deals with LGBT homophobia topics while the second analyzes anti-migrants posts. We used all 2,188 hate speech comments, and randomly sampled 3,812 non-hate speech comments.

Table 1: Characteristics of the used datasets: type and number of instances, as well as the input embeddings for each of the datasets.

Dataset	type	Size	Hate	Non-hate	LSTM embeddings
English	tweets	5000	1430	3670	sentence
Croatian	news comments	8422	4211	4211	fastText
Slovene	Facebook comments	6000	2188	3812	fastText

4.2 Implementation

The two Bayesian methods that were proposed in this paper to improve the annotation process have full implementation within their original papers. All of the particularities of how MCD BERT was implemented in PyTorch library[5] are presented in (Miok et al., 2020). The implementation details of the MM method are clearly explained in (Pirš and Štrumbelj, 2019) methods section and in this paper we provide full R code [6].

5 Results

In this section, we present three groups of results: removing uncertain instances from the training set, creating a cleaner training set, and improving annotations using he Bayesian ensembles.

5.1 Removing Uncertain Instances

Using MCD BERT, we obtain multiple predictions for each test set instance, and compute their mean and variance. Using the mean, we determine the classification (hate speech or not), while the variance reports on the certainty of the BERT for this specific instance. Based on the variance, we group classifications into certain and uncertain. Unsurprisingly, removing the uncertain test set instances improves the prediction performance as shown in Table 2, but also leaves a portion of borderline instances unclassified.

From Table 2 we can conclude that the variance of MCD BERT predictions is correlated with the performance of models: the more variance there is in the predictions the less accurate the model. Thus, removing the uncertain classifications can seemingly improve the performance of the test set. A practical benefit of this is that uncertain classification could be passed back to annotators to recheck them.

5.2 Creating Cleaner Training Sets

While the removal of uncertain instances from the test set might just sweep the problematic instances under the carpet, a more practical benefit is to use the uncertainty information to create a better training set. The test tweets/comments were removed based on how variate are their predictions. Thus, we repeatedly train the MCD BERT model on part of the dataset and use this model to obtain multiple predictions on the other part of the training dataset. In such a way, we collect multiple predictions for all original training tweets or comments and remove observations with the highest prediction variance. As a result of this procedure, 15 and 18 percent of the most uncertain predictions were removed for the English and Slovene dataset respectively. Croatian dataset contains a lot of comments with high variability in their predictions

[4] `http://nl.ijs.si/frenk/` (Research on Electronic Inappropriate Communication)
[5] `https://github.com/KristianMiok/Bayesian-BERT`
[6] `https://github.com/gregorp90/MM`

Table 2: Performance of multilingual BERT model, after removing uncertain instances from the test set of 1000 comments.

Language	Metric	Full dataset	200 removed	500 removed	700 removed
EN	**Accuracy**	0.91	0.96	0.996	0.997
	Precision	0.90	0.95	0.992	0.994
	Recall	0.89	0.95	1	1
	F1	0.88	0.95	0.995	0.997
CRO	**Accuracy**	0.72	0.76	0.84	0.87
	Precision	0.68	0.71	0.80	0.85
	Recall	0.54	0.69	0.78	0.75
	F1	0.61	0.70	0.79	0.83
SLO	**Accuracy**	0.71	0.76	0.83	0.87
	Precision	0.60	0.65	0.70	0.65
	Recall	0.56	0.64	0.66	0.54
	F1	0.58	0.65	0.68	0.59

so for this dataset we removed around 35% of the most uncertain comments. The details of how many instances were removed for each of the three datasets are presented in Table 3.

Table 3: Sizes of the datasets before and after the removing: original number of instances, number of instances removed and final training data size.

Dataset	Training Size	Number of removed	Final Size	Percent removed
English	4000	719	3281	18 %
Croatian	7422	2615	4807	35%
Slovene	5000	731	4269	15 %

Using prediction certainty to remove the uncertain instances from the training can improve the fine-tuning of BERT. For neural network models, during training or fine-tuning their performance is evaluated on a separate validation set. In Table 4, we can observe how the prediction accuracy on the validation set is improved with number of training epochs. We can see that fine-tuning BERT on the cleaner dataset improves its performance. We hypothesize that when the uncertainty due to unreliable labels is reduced, the decision boundary is easier to determine.

Table 4: Performance (measured using F_1 score) on the validation sets during training for original and cleaned datasets.

	English		Croatian		Slovene	
	Original	**Cleaned**	**Original**	**Cleaned**	**Original**	**Cleaned**
Epoch1	0.92	0.98	0.68	0.77	0.70	0.64
Epoch2	0.92	0.98	0.69	0.77	0.70	0.77
Epoch3	0.92	0.98	0.68	0.78	0.71	0.79
Epoch4	0.92	0.98	0.70	0.79	0.72	0.81

Results for the model fine-tuned on the cleaned dataset are contained in Table 5. Compared to the results in Table 2 (see the "Full dataset" column), the prediction results for Croatian and Slovenian datasets are improved while for the English dataset this is not the case. We explain this by the fact that the English dataset is well-annotated with high-quality predictions. On the other hand, we believe that the Croatian and Slovenian datasets are less clean and contain several questionable annotations. This can be confirmed for the Croatian dataset, which was created within the project we participate in, so we are well-informed about the annotation process.

Table 5: Test set performance (F_1 score) of the models trained on the cleaned datasets.

Metrics	English	Croatian	Slovene
Accuracy	0.87	0.74	0.72
Precision	0.88	0.73	0.62
Recall	0.81	0.60	0.55
F1	0.85	0.66	0.59

5.3 Improving Annotations using Bayesian Ensembles

We propose a Bayesian ensemble as a support method for the annotation process. As annotators can be distracted, biased, or influenced, we propose to use the MM method to provide them a hint of how shall they annotate the instances. From Table 6, we can observe that by combining probabilistic predictions of BERT, random forest, and support vector machines, we can further improve the predictive performance. The MM ensemble not only improves BERT's results but also provides better calibrated predictions as evidenced from Figure 1.

Table 6: The F_1 score of the hate speech classifiers and their ensemble.

Method	English	Croatian	Slovene
BERT	0.91	0.72	0.71
RF	0.83	0.67	0.65
SVM	0.86	0.71	0.69
MM	**0.92**	**0.74**	**0.72**

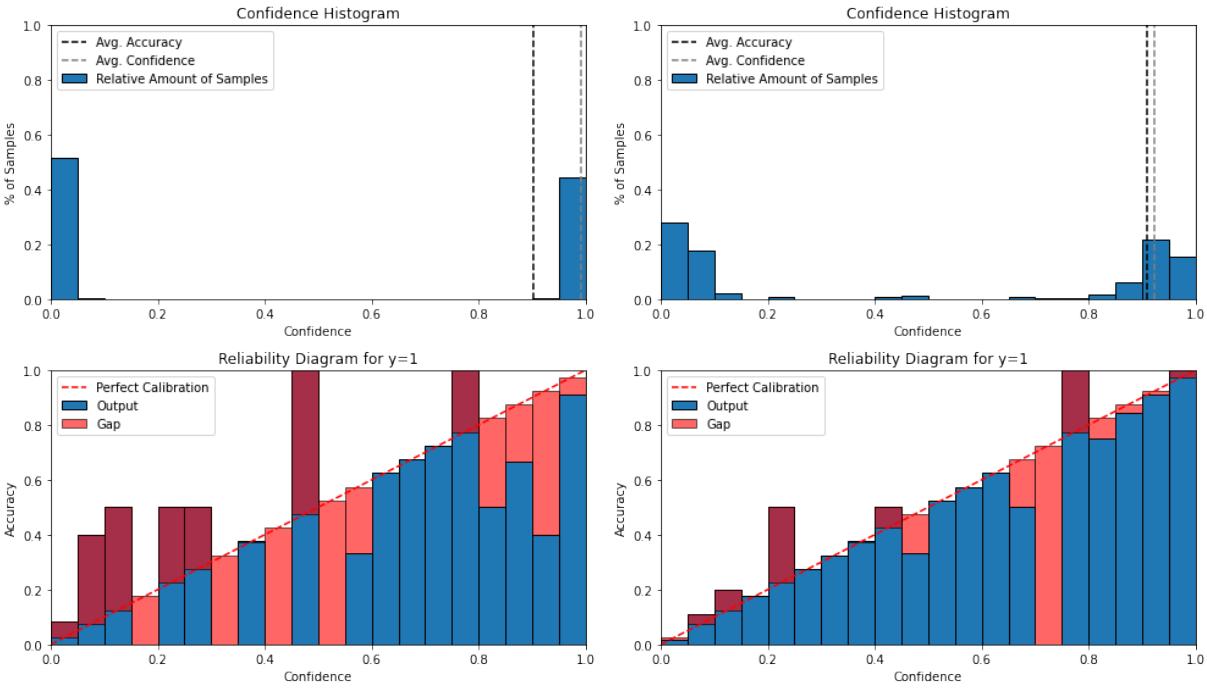

Figure 1: Calibration for the BERT predictions (left) and MM model predictions (right).

6 Conclusions and Further Work

A large amount of currently available textual data allows and requires modeling with machine learning methods. To apply the supervised methods, the text data has to be annotated, and effective learning requires accurate annotations, which may be expensive for organizers and difficult for human annotators. For this reason, the annotation process is often coupled with semi-supervised machine learning and classification reliability estimation.

We presented several machine learning approaches, based on Bayesian inference, that can improve the data annotation process. First, multiple predictions obtained with MCD BERT can identify instances with questionable labeling. Second, removing training instances with unreliable labels can improve the quality of the training set, making it more homogeneous and cleaner, thereby improving the predictive performance of BERT models. Third, probabilistic ensemble combinations can help annotators to better label the data by providing more accurate and better calibrated prediction probabilities. In conclusion, Bayesian methods can improve the annotation process and shall be further investigated and improved for this task.

In further work, we will focus on improving our method on how to remove uncertain instances. We will construct and test a workflow for semi-supervised text annotation in a real-world setting. Testing different dropout levels in the BERT model may provide a better understanding of its uncertainty and calibration.

Acknowledgements

This paper was supported by European Union's Horizon 2020 Programme project EMBEDDIA (Cross-Lingual Embeddings for Less-Represented Languages in European News Media, grant no. 825153). The research was supported by the Slovenian Research Agency through research core funding no. P6-0411, project CANDAS (Computer-assisted multilingual news discourse analysis with contextual embeddings, grant no. J6-2581), and Young researcher grant (Gregor Pirš).

References

Ayan Acharya, Eduardo R Hruschka, Joydeep Ghosh, Badrul Sarwar, and Jean-David Ruvini. 2013. Probabilistic combination of classifier and cluster ensembles for non-transductive learning. In *Proceedings of the 2013 SIAM International Conference on Data Mining*, pages 288–296. SIAM.

Théodore Bluche, Christopher Kermorvant, and Jérôme Louradour. 2015. Where to apply dropout in recurrent neural networks for handwriting recognition? In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, pages 681–685. IEEE.

Keyan Cao, Guoren Wang, Donghong Han, Jingwei Ning, and Xin Zhang. 2015. Classification of uncertain data streams based on extreme learning machine. *Cognitive Computation*, 7(1):150–160.

Jesús Cerquides and Ramon López De Mántaras. 2005. Robust Bayesian linear classifier ensembles. In *European Conference on Machine Learning*, pages 72–83. Springer.

Roger Cooke. 1991. *Experts in uncertainty: opinion and subjective probability in science*. Oxford University Press on Demand.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Eleventh international AAAI conference on web and social media*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Darja Fišer, Tomaž Erjavec, and Nikola Ljubešić. 2017. Legal framework, dataset and annotation schema for socially unacceptable online discourse practices in slovene. In *Proceedings of the first workshop on abusive language online*, pages 46–51.

Meire Fortunato, Charles Blundell, and Oriol Vinyals. 2017. Bayesian recurrent neural networks. *arXiv preprint arXiv:1704.02798*.

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

Yarin Gal and Zoubin Ghahramani. 2016b. A theoretically grounded application of dropout in recurrent neural networks. In *Advances in Neural Information Processing Systems*, pages 1019–1027.

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

Lirong He, Bin Liu, Guangxi Li, Yongpan Sheng, Yafang Wang, and Zenglin Xu. 2018. Knowledge base completion by variational bayesian neural tensor decomposition. *Cognitive Computation*, 10(6):1075–1084.

Jennifer A. Hoeting, David Madigan, Adrian E. Raftery, and Chris T. Volinsky. 1999. Bayesian model averaging: a tutorial. *Statistical science*, pages 382–401.

Pavel Izmailov, Wesley J Maddox, Polina Kirichenko, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. 2019. Subspace inference for bayesian deep learning. *arXiv preprint arXiv:1907.07504*.

Zeynep H Kilimci and Selim Akyokus. 2018. Deep learning-and word embedding-based heterogeneous classifier ensembles for text classification. *Complexity*, 2018.

Hyun-Chul Kim and Zoubin Ghahramani. 2012. Bayesian Classifier Combination. In *International Conference on Artificial Intelligence and Statistics*, pages 619–627.

Kristina Kocijan, Lucija Košković, and Petra Bajac. 2019. Detecting hate speech online: A case of croatian. In *International Conference on Automatic Processing of Natural-Language Electronic Texts with NooJ*, pages 185–197. Springer.

Alexandre Lacoste, Mario Marchand, François Laviolette, and Hugo Larochelle. 2014. Agnostic Bayesian learning of ensembles. In *International Conference on Machine Learning*, pages 611–619.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in neural information processing systems*, pages 6402–6413.

Ming Li, Peilun Xiao, and Ju Zhang. 2018. Text classification based on ensemble extreme learning machine. *arXiv preprint arXiv:1805.06525*.

Dennis V. Lindley. 1985. Reconciliation of discrete probability distributions' in Bernado. *JM, DeGroot, MH, Lindley, DV, and Smith, AFM, Eds., Bayesian Statistics II. North Holland, Amsterdam*, pages 375–391.

Nikola Ljubešić, Tomaž Erjavec, and Darja Fišer. 2018. Datasets of slovene and croatian moderated news comments. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 124–131.

Nikola Ljubešić, Darja Fišer, and Tomaž Erjavec. 2019. The FRENK datasets of socially unacceptable discourse in Slovene and English. In *International Conference on Text, Speech, and Dialogue*, pages 103–114. Springer.

Shervin Malmasi and Marcos Zampieri. 2017. Detecting hate speech in social media. *arXiv preprint arXiv:1712.06427*.

Yashar Mehdad and Joel Tetreault. 2016. Do characters abuse more than words? In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 299–303.

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

Kristian Miok, Dong Nguyen-Doan, Blaž Škrlj, Daniela Zaharie, and Marko Robnik-Šikonja. 2019b. Prediction uncertainty estimation for hate speech classification. In *International Conference on Statistical Language and Speech Processing*, pages 286–298. Springer.

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

Kristian Miok, Blaz Skrlj, Daniela Zaharie, and Marko Robnik-Sikonja. 2020. To ban or not to ban: Bayesian attention networks for reliable hate speech detection. *arXiv preprint arXiv:2007.05304*.

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

Marzieh Mozafari, Reza Farahbakhsh, and Noel Crespi. 2019. A BERT-based transfer learning approach for hate speech detection in online social media. In *International Conference on Complex Networks and Their Applications*, pages 928–940. Springer.

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

Alfredo Nazábal, Pablo García-Moreno, Antonio Artés-Rodríguez, and Zoubin Ghahramani. 2016. Human activity recognition by combining a small number of classifiers. *IEEE journal of biomedical and health informatics*, 20(5):1342–1351.

Alexandru Niculescu-Mizil and Rich Caruana. 2005. Predicting good probabilities with supervised learning. In Luc De Raedt and Stefan Wrobel, editors, *Proceedings of the 22nd International Machine Learning Conference*. ACM Press.

Endang Wahyu Pamungkas and Viviana Patti. 2019. Cross-domain and cross-lingual abusive language detection: A hybrid approach with deep learning and a multilingual lexicon. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 363–370.

Gregor Pirš and Erik Štrumbelj. 2019. Bayesian combination of probabilistic classifiers using multivariate normal mixtures. *J. Mach. Learn. Res.*, 20:51–1.

John C. Platt. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in large margin classifiers*, pages 61–74. MIT Press.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10.

Catarina Silva, Uros Lotric, Bernardete Ribeiro, and Andrej Dobnikar. 2010. Distributed text classification with an ensemble kernel-based learning approach. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 40(3):287–297.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

Lukas Stappen, Fabian Brunn, and Björn Schuller. 2020. Cross-lingual zero-and few-shot hate speech detection utilising frozen transformer language models and axel. *arXiv preprint arXiv:2004.13850*.

Xiao Sun, Xiaoqi Peng, and Shuai Ding. 2018. Emotional human-machine conversation generation based on long short-term memory. *Cognitive Computation*, 10(3):389–397.

György Szarvas, Veronika Vincze, Richárd Farkas, and János Csirik. 2008. The bioscope corpus: annotation for negation, uncertainty and their scope in biomedical texts. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, pages 38–45.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Boris Vezjak. 2018. Radical hate speech: The fascination with Hitler and fascism on the Slovenian webosphere. *Solsko Polje*, 29.

Veronika Vincze. 2015. *Uncertainty detection in natural language texts*. Ph.D. thesis, szte.

Sida Wang and Christopher Manning. 2013. Fast dropout training. In *International Conference on Machine Learning*, pages 118–126.

William Warner and Julia Hirschberg. 2012. Detecting hate speech on the world wide web. In *Proceedings of the second workshop on language in social media*, pages 19–26. Association for Computational Linguistics.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93.

Gregor Wiedemann, Seid Muhie Yimam, and Chris Biemann. 2020. Uhh-lt & lt2 at semeval-2020 task 12: Fine-tuning of pre-trained transformer networks for offensive language detection. *arXiv preprint arXiv:2004.11493*.

Yuling Yao, Aki Vehtari, Daniel Simpson, Andrew Gelman, et al. 2018. Using stacking to average Bayesian predictive distributions. *Bayesian Analysis*, 13(3):917–1007.

Wojciech Zaremba, Ilya Sutskever, and Oriol Vinyals. 2014. Recurrent neural network regularization. *arXiv preprint arXiv:1409.2329*.

Xunan Zhang, Shiji Song, and Cheng Wu. 2013. Robust bayesian classification with incomplete data. *Cognitive Computation*, 5(2):170–187.

Zhi-Hua Zhou and Ming Li. 2010. Semi-supervised learning by disagreement. *Knowledge and Information Systems*, 24(3):415–439.

Zhi-Hua Zhou. 2012. *Ensemble methods: foundations and algorithms*. CRC press.

Xiaojin Zhu and Andrew B Goldberg. 2009. Introduction to semi-supervised learning. *Synthesis lectures on artificial intelligence and machine learning*, 3(1):1–130.

Lingxue Zhu and Nikolay Laptev. 2017. Deep and confident prediction for time series at Uber. In *IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 103–110.

Xiaojin Jerry Zhu. 2005. Semi-supervised learning literature survey. Technical report, University of Wisconsin-Madison Department of Computer Sciences.