forumBERT: Topic Adaptation and Classification of Contextualized Forum Comments in German

Ayush Yadav
International Institute
of Information Technology
Bangalore, India
yadavayush.ay42@gmail.com

Benjamin Milde
Language Technology Group
Dept. of Informatics
Universität Hamburg, Germany
milde@informatik.uni-hamburg.de

Abstract

Online user comments in public forums are often associated with low quality, hate speech or even excessive demands for moderation. To better exploit their constructive and deliberate potential, we present forumBERT. forumBERT is built on top of the BERT architecture and uses a shared weight and late fusion technique to better determine the quality and relevance of a comment on a forum article. Our model integrates article context with comments for the online/offline comment moderation task. This is done using a two step procedure: self-supervised BERT language model fine tuning for topic adaptation followed by integration into the forumBERT architecture for online/offline classification. We present evaluation results on various classification tasks of the public One Million Post dataset, as well as on the online/offline comment moderation task on 998,158 labelled comments from NDR.de, a popular German broadcaster's website. forumBERT significantly outperforms baseline models on the NDR dataset and also outperforms all existing advanced baseline models on the OMP dataset. Additionally we conduct two studies on the influence of topic adaptation on the general comment moderation task.

1 Introduction

Online user comments, such as those on journalistic content or product features are often associated with low quality, hate speech or even excessive demands for moderation. Automating this moderation or aspects of it can be considered to be of high practical interest. One of the key challenges of forum comment moderation is the specificity of category of classification. Forum comments have to be moderated for hate-speech, discrimination, spam among many other generally discussed classification tasks. Additionally comments on forum articles must also be moderated for relevance and contribution to the discourse.

Previous work by Schabus et al. (2017) and Schabus and Skowron (2018) introduces the idea of applied classification, wherein comments are annotated across multiple forum specific categories and classification models are created for each category. In this paper we focus on the more general "comment moderation task" on news forum comments. In this task, comments can be classified into one of two categories, either online or offline, where an online classification represents a comment that is accepted by the forum moderators and an offline classification represents comments that have been taken down by the forum moderators.

In recent years, the Natural Language Processing community has experienced a substantial shift towards using pre-trained models. Their usage on large corpora has proved to be beneficial in learning general language representations and has shown improvement in text classification and many other NLP tasks, which has also helped avoid training large language models from scratch. However, the lack of portability of NLP models to new conditions is a central issue in NLP. For many target applications like comment moderation on niche public forums, labelled data might be lacking and there might not be enough unlabelled data to train a general language model. These conditions press us to visit domain adaptation to improve the language model.

Therefore, in this paper we present forumBERT, a modification to the BERT architecture which uses two weight shared BERT models and a late fusion technique to better determine a comment's quality and relevance on a forum article. We also extend the work by Rietzler et al. (2020) and investigate the influence of a domain adapted BERT language model on the downstream comment moderation accuracy as a function of labelled downstream training examples. In particular, the contributions of our paper are:

- We present the forumBERT architecture to determine a comment's quality and relevance on a forum post.
- We introduce the NDR dataset which is used for the comment moderation task.
- We show that forumBERT outperforms baseline models on the comment moderation task. forumBERT achieves state of the art results on seven classification tasks on the One Million Posts Dataset.
- We analyse the influence of topic adaptation on the forumBERT architecture by varying the number of labelled datapoints in the comment moderation task.
- We also analyse the influence of the number of training steps of the BERT language model and the results on the downstream comment moderation classification task.

This paper has been structured in the following way: Section 2 introduces the BERT architecture and mentions existing comment moderation architectures and some relevant BERT model adaptations. Section 3 describes the NDR dataset and the NDR topic datasets. Section 4 introduces forumBERT and the training procedure followed. Section 5 evaluates forumBERT and BERT on the NDR dataset and the OMP dataset. Section 6 contains our topic adaptation experiments on the effectiveness of topic adaptation and the influence of topic adaptation as a function of labelled training examples.

2 Related Work

Pre-trained models using large corpora have dominated the task of text classification. This began with pre-trained word embeddings such as Word2Vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014) and now in the current paradigm, pre-trained models like ELMo (Peters et al., 2018), BERT (Devlin et al., 2019), GPT/GPT2 (Radford et al., 2019), XLNet (Yang et al., 2019), have achieved state of the art results in a wide spectrum of NLP taks including text classification.

BERT (Devlin et al., 2019) is an amalgamation of several key findings in NLP research such as contextualized word representations, sub word tokenization (Wu et al., 2016) and transformers (Vaswani et al., 2017). The main innovations are the unique learning methods adopted by BERT. The BERT language model is trained to optimize on two tasks, i.e Masked Language Modelling (MLM) and Next Sentence Prediction.
Masked language modeling is a fill-in-the-blank task, where a model uses the context words surrounding a [MASK] token to try to predict what the [MASK] word should be. Next Sentence Prediction is a classification task, in which the BERT model receives a pair of sentences as input and learns to predict if the second sentence in the pair is the subsequent sentence in the original document.

2.1 Comment Moderation Architectures and BERT Adaptations

Pavlopoulos et al. (2017a) introduced an RNN based method for the comment moderation task on a Greek news sports portal. This method was improved by Pavlopoulos et al. (2017b) to include dataset specific user-embeddings, generated by accounting for the number of accepted and rejected comments of every user on the sports portal. Risch and Krestel (2018) have proposed a semi-automatic approach to comment moderation using a comment, user and article information to create a high recall logistic regression model.

Large pre-trained BERT language models have been incorporated into many task specific architectures. Sentence-BERT (Reimers and Gurevych, 2019) is one such modification of the BERT network using Siamese and Triplet networks that is able to derive semantically meaningful sentence embeddings where semantically similar sentences are closer in the vector space. SentiBERT (Yin et al., 2020) is a BERT variant that effectively captures compositional sentiment semantics by incorporating BERT's contextualized representation with binary constituency parse tree to capture semantic composition.

However, in the current paradigm, pre-trained language models are generalized and their portability to new conditions still remains an issue. To this end, work by Rietzler et al. (2020) and Xu et al. (2019) shows that in the aspect target sentiment task, the performance of models that are pre-trained on a general language corpus can be improved by fine tuning the language model on a domain specific corpus. We build on this and in Section 6 show that even in the comment moderation task on niche forums, the performance of models that are pre-trained on a German general language corpus can be improved by finetuning the

language model on each specific forum topic.

3 Datasets

To verify the topic adaptation capabilities in German news forum datasets, we procured the NDR dataset[1] which consists of almost one million labelled user comments and their adjoining articles from the NDR news website. This dataset can be obtained directly from NDR for academic and research use. To evaluate the performance of our forumBERT architecture on an already existing dataset, we use the One Million Posts Dataset (Schabus et al., 2017).

3.1 NDR Dataset

The NDR dataset consists of a collection of 998,158 labelled comments on 65,261 articles on the NDR website. All comments were collected between five and a half year span from 2014-05-09 to 2019-12-12. The dataset consists of the following attributes for every comment:

- Headline: The title of the article
- URL: A URL to the article on the NDR website
- Comment: The comment text
- Date: The date of posting the comment
- Label: A binary offline/online label, which represents the final status of the comment on the website. Offline labelled comments are considered non-desirable content on the forum.

On average the length of a comment on the NDR dataset is 59.15 words. The quartile comment lengths are shown in Table 1 and the distribution of comment lengths is plotted in Figure 1.

quartile	comment length
0.25	22
0.50	43
0.75	79
1.00	1308

Table 1: comment length at every quartile in the NDR dataset

[1] https://www.ndr.de/index.html

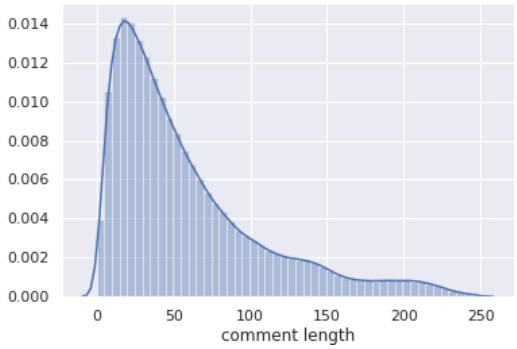

Figure 1: Distribution of comment length on the NDR dataset (clipped to a maximum comment length of 250 words.)

3.1.1 Topic Segmentation

News datasets are very general in nature where discussions range from politics, sports to technology and scientific news. Therefore, we used the URL attribute to segment the entire dataset into different topics for topic adaptation. Specifically, by splicing the URL attribute, topic information was obtained for each comment. For example, in "http://relaunch.ndr.de/sport/handball/bundesliga/" the url contains the topic of the article, which in this case is sport. This is used to segment the entire dataset into topics. The number of comments per topic are shown in Table 2.

Topic	online	offline	offline %
Nachrichten (News)	613061	215656	26.02%
Sport	73151	12258	14.35%
Kultur (Culture)	21231	5485	20.53%
Fernsehen (TV)	12218	2998	19.70%
Info	20020	5491	21.52%
Radio	1986	295	12.93%
Rest	11177	2996	21.11%

Table 2: number of online and offline examples in all topic forums

We applied topic adaptation (Rietzler et al., 2020) to two topics, "sport" and "kultur" (Culture), as both had among the most labelled training datapoints, as shown in Table 2). "Nachrichten" (News) is too general to be considered a forum topic and thus was omitted.

3.2 One Million Posts (OMP)

The One Million Posts dataset (OMP Schabus et al. (2017)) contains a selection of user comments posted to the Austrian Newspaper website "Der

Standard". The comments have been selected from a 12 month time span between 2015-06-01 and 2016-05-31. There are 11,773 freely labelled posts on nine categories (not all labelled comments are labelled in every category) and 1,000,000 unlabelled posts in the data set. The amount of labelled data for each of the nine categories has been mentioned in Table 3

Category	Does Apply	Does Not Apply	Percentage
Sentiment Negative	1691	1908	47%
Sentiment Neutral	1865	1734	52%
Sentiment Positive	43	3556	1%
Off-Topic	580	3019	16%
Inappropriate	303	3296	8%
Discriminating	282	3317	8%
Possibly Feedback	1301	4737	22%
Personal Stories	1625	7711	17%
Arguments Used	1022	2577	28%

Table 3: number of labelled examples in each category in the OMP dataset (Schabus et al., 2017)

4 Methodology

This section presents forumBERT, which is an extension of BERT for contextual classification tasks like general comment moderation task. We use a German language pre-trained BERT language model as a basis and approach this task using a three-step procedure. In the first step we finetune the pre-trained weights of the language model in a self-supervised way on a topic-specific corpus. In the second step we incorporate this finetuned language model into the forumBERT architecture. The final step is the supervised training of forumBERT for the online/offline classification end-task. A schema for this process is depicted in Figure 2

In the following subsections, we discuss how we finetune the BERT language model and then the forumBERT architecture.

4.1 BERT: Language Model Finetuning and Topic Adaptation

To create our forumBERT model, our first step deals with finetuning a pretrained BERT language model using a topic specfic corpora. As described in Section 3.1.1 we split the NDR dataset into multiple topics. We adopt post-training of BERT (Xu et al., 2019) on a topic dataset which is algorithmically the same as pretraining the model. The Masked Language Modelling task is used to learn topic knowledge and remove any biases learnt from the pretraining datasets. Next Sentence Prediction helps BERT learn contextualized embeddings that

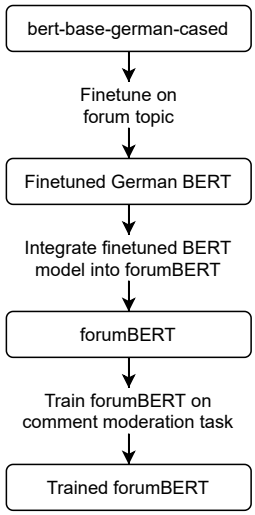

Figure 2: Schema diagram for the construction of forumBERT.

are beyond word level. This is important since, at a high level we wish to generate similar embeddings for comments that are in the same context as it's adjoining article's context. Finetuning the language model helps mitigate the problem of having less labelled data, which is the case in many online forums. This finetuned language model is then incorporated into the forumBERT architecture.

Other than using the topic adapted BERT language model to create the forumBERT model, we also investigate the limitations of language model finetuning for the comment moderation task through two tasks described in Section 6.

4.2 forumBERT: A Weight Shared BERT Model

forumBERT is an extension of BERT for topic-knowledge learning and forum-comment classification. The model must be able to compare the article and the comment on the article to determine its quality and relevance on the forum. Inappropriate and discriminatory comments must be removed from the forum irrespective of the corresponding articles, but the model must also remove comments that are off-topic/irrelevant and digress too far from the topic of the article. To achieve this we use the forumBERT architecture.

We adapt the finetuned $BERT_{BASE}$ model for forum comment classification by using two finetuned $BERT_{BASE}$ models, one which takes in as input the headline of the article and another which takes the comment on the corresponding article as input. To mitigate the problem of a parameter explosion be-

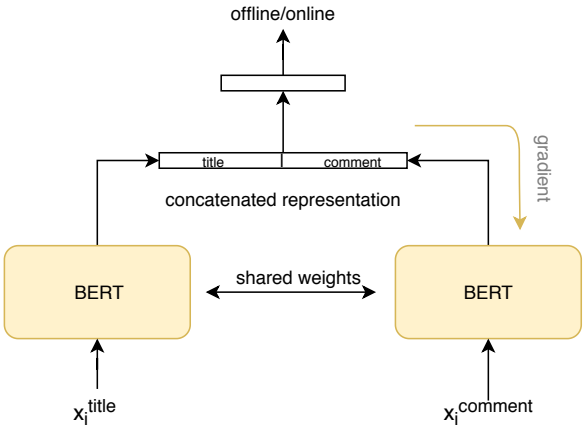

Figure 3: forumBERT architecture

cause of using 2 BERT models and to add implicit regularization we share the weight between the two BERT models (shown in Figure 3).

We follow Devlin et al. (2019) and consider the final hidden state corresponding to the [CLS] input token for both the BERT models. The pair of article and comment representations thus obtained are both of dimensions 768×1. The pair of embeddings are concatenated at the output of the BERT model (late fusion). Late fusion is preferred rather than concatenating the input tokens and passing them through the network, to allow the network to fully separate out the differences between the article and the comment. The dimensions of the concatenated vector is 1536×1. The fused vector is then passed through 2 fully connected layers with weights $W_t \in \mathbb{R}^{2n \times n}$ and $W_{t'} \in \mathbb{R}^{n \times k}$ respectively, where n is the dimension of the comment/headline embedding (n = 768) and k is the number of labels (k = 2). A softmax function is applied to the final k length vector.

$$\text{out} = \text{softmax}(W_{t'}(W_t(x))) \quad (1)$$

Here, x represents the fused representation vector. We optimize the cross-entropy loss.

4.3 Implementation Details

As a base for all our experiments we use the $\text{BERT}_{\text{BASE}}$ model which consists of 12 layers (transformer blocks), 12 attention heads 768 hidden dimensions per token amounting to a total of 110 million parameters. The parameters of this model are initialized using bert-base-german-cased[2], which has been pretrained on the German Wikipedia Dump (6 GB), the German OpenLegalData dump (2.4 GB) and German news articles (3.6 GB) and released by deepset.ai[3]. For the BERT language model finetuning we use 32 bit floating point computations using the Adam optimizer (Kingma and Ba, 2015). The batchsize is set to 8 while the learning rate is set to $3 \cdot 10^{-5}$. The maximum input sequence length is set to 512 tokens, which amounts to about 11 sentences per sequence on average. For all experiments except Experiment 6.1 we use a forumBERT model in which we integrate a topic adapted BERT language model which is trained for 13 epochs on the entire topic with a learning rate of $3 \cdot 10^{-5}$.

For the down-stream online/offline classification task we use 32 bit floating point computations and the Adam optimizer. The models are trained for 7 epochs, with a learning rate of $2 \cdot 10^{-6}$ for the two epochs and $6.31 \cdot 10^{-7}$ for the remaining 5 epochs. The validation accuracy converges after about 3 epochs.

For all experiments and results on the NDR dataset, we split the topic dataset in a 9:1 ratio. The larger portion of the dataset is used for language modelling and for training on downstream tasks and the smaller portion is used only for testing on downstream tasks.

5 Results

5.1 Comment Moderation Task on NDR Dataset

Meas.	BOW com.	D2V tit.+com.	BERT com.	BERT tit.+com.	fBERT tit./com.
Prec.	0.65	0.60	**0.73**	0.71	0.698
Rec.	0.27	0.15	0.38	0.42	**0.431**
F1.	0.38	0.24	0.50	0.527	**0.533**
Acc.	0.786	0.767	0.810	0.814	**0.819**

Table 4: Results of the comment moderation task on the entire NDR dataset (without any topic segmentation). Precision, Recall, F1-score are all computed on the minority class (offline).

For the comment moderation task, we compare the performance of forumBERT with following baseline models trained on the NDR sport and kulur topic datasets: 1) Logistic regression on count vectorizer (BOW model); 2) logistic regression on doc2vec[4] representation (D2V model); 3) 3 layer DNN (dense neural network) (3DNN model) built

[2] https://huggingface.co/bert-base-german-cased

[3] https://deepset.ai/german-bert

[4] The doc2vec document embedding (Le and Mikolov,

Sports Topic										
No. Training Ex.	**1024**			**8192**			**All Topic Data**			
Model	Prec.	Rec.	F1	Prec	Rec	F1	Prec	Rec.	F1	Wins
log-reg (count)	0.222	0.639	0.329	**0.586**	0.212	0.311	**0.591**	0.212	0.311	2
log-reg (D2V)	0.203	0.618	0.305	0.220	0.649	0.328	0.428	0.046	0.083	0
3DNN (D2V)	0.131	0.496	0.207	0.141	0.460	0.216	0.290	0.241	0.263	0
BERT (comment)	0.267	0.633	0.375	0.318	0.692	0.435	0.584	0.369	0.452	0
BERT (title + comment)	0.283	0.650	0.394	0.337	0.689	0.452	0.571	0.406	0.475	0
forumBERT (title/comment)	**0.295**	**0.697**	**0.414**	0.328	**0.741**	**0.457**	0.483	**0.547**	**0.513**	7
Kultur Topic										
No. Training Ex.	**1024**			**8192**			**All Topic Data**			
Model	Prec.	Rec.	F1	Prec	Rec	F1	Prec	Rec.	F1	Wins
log-reg (count)	0.264	0.617	0.370	0.331	0.678	0.445	0.513	0.327	0.339	0
log-reg (D2V)	0.210	0.637	0.316	0.253	0.636	0.362	0.578	0.056	0.102	0
3DNN (D2V)	0.193	0.636	0.296	0.202	0.607	0.302	0.292	**0.476**	0.362	1
BERT (comment)	0.319	**0.751**	0.447	0.318	**0.803**	0.455	0.552	0.417	0.475	2
BERT (title + comment)	0.358	0.652	0.462	**0.439**	0.638	**0.520**	0.650	0.363	0.465	2
forumBERT (title/comment)	**0.367**	0.643	**0.467**	0.398	0.643	0.468	**0.706**	0.375	**0.490**	4

Table 5: Comment moderation task results on the NDR sport topic dataset and the culture topic dataset. The results have been computed for three quantities of uniformly sampled training examples with the first two being 1024 and 8192. The final quantity is all training comments from that particular topic. Precision, recall and F1-score are computed on the minority class (offline).

on doc2vec representations; 4) two BERT models. For all models other than forumBERT and a BERT model, contextualized input of the form "TITLE [title] COMMENT [comment]", is provided as input. To test the importance of providing context, we also train a BERT model using only comment text as input.

We report performance measures on Table 5. forumBERT significantly outperforms all other models and has the highest F1 scores in both the sports and kultur topic datasets, even in few shot conditions (1024/8192 training examples). From this table, it can be seen that our approach significantly outperforms the standard BERT model, improving the F1 scores from 0.475 to 0.513 (8% increase) in the sports topic and an improvement from 0.465 to 0.490 (a 5.3% increase) in the kultur dataset. Also if we compare forumBERT to a standard BERT model with only comment input the F1 scores increase from 0.452 to 0.513 (a 13.4% performance gain) on the sport topic and an improvement from 0.475 to 0.490 (a 3.15% gain).

Table 4 represents the effectiveness of the design architecture of the forumBERT model. The forumBERT model considered here uses a pretrained BERT language model without performing topic adaptation. We see that forumBERT outperforms all other methods, giving the best recall value, F1 score and the best accuracy on the entire dataset.

5.2 Classification on the OMP Dataset

We also compare the performance of: 1) forumBERT; 2) BERT with contextualized input 3) BERT without contextualized input; 4) the baselines reported in Schabus et al. (2017); 5) advanced baseline for doc2vec (Le and Mikolov, 2014) (D2V) vector representation and a support vector machine (Cortes and Vapnik, 1995) with Radial Basis Function (RBF) kernel for classification as reported in Schabus and Skowron (2018). To compare with the published results, all results have been computed using stratified 10-fold cross validation. The forumBERT model considered here uses a pretrained BERT language model without topic adaptation. The results for each category are reported in Table 6.

From Table 6, it can be seen that for categories that do not require additional context from the article (i.e Sentiment Negative and Discriminating) "BERT with only input comment text" performs among the best. Providing contextualized input in the form of article title and comment dilutes the

2014) was first trained on the NDR dataset, prior to training any models for the comment moderation task.

information input to the model leading to worse predictions.

For categories that require contextualized input (i.e offtopic, inappropriate, Possibly Feedback and Personal Stories) it can be seen that "BERT with contextualized inputs" gives best results and slightly outperforms forumBERT in almost all categories to establish the state of the art results. Upon further investigation, we found that 10 articles account for a majority of the annotated comments in OMP. More precisely, 10 articles are the source of 72.1% of all "OffTopic" and "Inappropriate" annotated comments, 58.3% of all "Personal Stories" annotated comments and 45.1% of all "Possibly Feedback" comments. Without diversity in the article input to the forumBERT model, it tends to perform slightly worse than BERT. This was not the case with the NDR dataset, where there was enough diversity in the articles (65,261 articles) to promote better classification.

Nonetheless forumBERT exceeds all baseline and advanced baseline results and still offers competitive results on the OMP dataset.

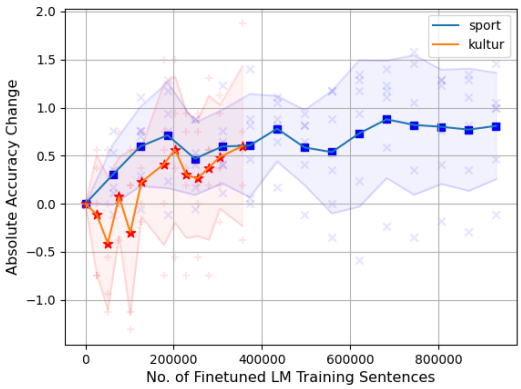

Figure 4: Absolute accuracy percentage improvements on the downstream offline/online classification task as a function of the number of training sentences the BERT language model was fine-tuned on. The ■ and ★ symbols represent the average over 6 runs of finetuning and classification on the "Sport" and "Kultur" topics of the NDR dataset (Section 3.1) respectively. The 'x' and the '+' represent individual runs. The filled-in portions represent the standard deviation over these 6 runs ($\mu \pm \sigma$). The absolute accuracy improvements are measured from 85.94% for the "Sport" topic and 81.64% for the "Kultur" topic.

Categ.	Meas.	BOW com.	D2V top.+com.	LSTM com.	BERT com.	BERT tit.+com.	fBERT tit./com.
Neg.	Prec.	0.552	0.621	0.534	0.664	0.663	**0.711**
	Rec.	0.510	0.483	**0.719**	0.642	0.709	0.646
	F1	0.530	0.544	0.613	0.654	**0.685**	0.677
Offtop.	Prec.	0.275	0.252	0.274	0.513	**0.537**	0.565
	Rec.	0.237	**0.453**	0.263	0.253	0.337	0.272
	F1	0.255	0.324	0.268	0.339	**0.415**	0.368
Inappr	Prec.	0.162	0.143	0.196	0.360	**0.411**	0.346
	Rec.	0.111	**0.412**	0.108	0.188	0.147	0.178
	F1	0.132	0.212	0.140	**0.247**	0.217	0.235
Disc	Prec.	0.184	0.154	0.113	**0.368**	0.325	0.304
	Rec.	0.102	**0.283**	0.141	0.112	0.052	0.112
	F1	0.132	0.200	0.126	**0.171**	0.089	0.163
Feed.	Prec.	0.655	0.531	0.630	0.741	**0.798**	0.792
	Rec.	0.580	0.735	0.628	0.698	**0.765**	0.762
	F1	0.616	0.617	0.630	0.719	**0.781**	0.771
Pers.	Prec.	0.698	0.589	0.638	0.836	**0.834**	0.832
	Rec.	0.592	0.850	0.665	0.828	**0.854**	0.841
	F1	0.640	0.696	0.651	0.832	**0.844**	0.836
Arg.	Prec.	0.610	0.545	0.568	0.716	**0.742**	0.733
	Rec.	0.512	0.763	0.645	0.733	0.754	**0.769**
	F1	0.526	0.636	0.604	0.725	0.748	**0.750**

Table 6: Classification results for multiple categories on the OMP dataset (Schabus et al., 2017). Precision, Recall and F1-score have been computed for the minority class for each category.

6 Experiments

We aim to answer the following research questions through our experiments:

Q1. How does the number of training iterations in the BERT language model finetuning stage influence the general comment moderation endtask performance on German topic forum datasets?

Q2. What is the influence of topic adaptation on the comment moderation endtask as a function of labelled endtask training examples?

6.1 Topic Adaptation

To answer Q1, we first split the topic datasets into a 9:1 ratio. The larger portion is used for BERT language model finetuning (topic adaptation) and the remaining is used for online/offline classification after every epoch of the language model finetuning. The results are shown in Figure 4.
Figure 4 and Table 2 empirically show that BERT is capable of learning topic specific forum comment knowledge even with less than 100,000 unlabelled training examples. We trained the BERT language model for 15 epochs individually on the sport and culture topic.
We also infer that topic based BERT language model finetuning improves the general downstream offline/online task. We see that the performance improves immediately in the case of the more specific sports topic, whereas for the more general

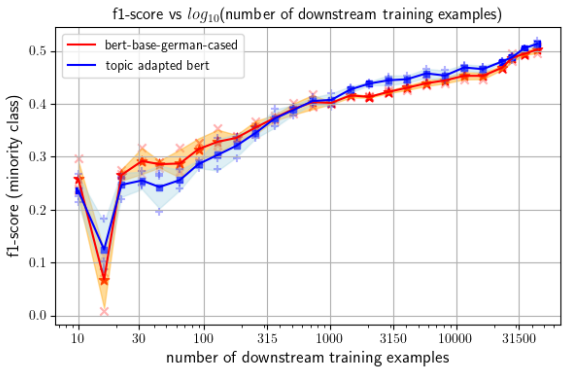

Figure 5: Average online/offline classification F1 score (for the minority "offline" class) computed on the sports topic using a pretrained forumBERT model (using bert-base-german-cased) and a sports topic adapted forumBERT model as a function of the number of downstream classification examples. The x-axis is represented on a log_{10} scale. The ■ and ★ symbols represent the average over 3 runs of online/offline classification on the sport topic of the NDR dataset (Section 3.1). The 'x' and '+' markers represent the individual runs. The filled-in portions represent the standard deviation over the 3 runs ($\mu \pm \sigma$)

culture topic, initially downstream classification performs worse (till 100,000 training sentences), but starts to see a steady gain in performance as it is trained after training on 150,000 sentences. Due to high variance in results, we average the results of 6 runs on each topic dataset and measure and plot the standard deviation to measure the improvements in performance.

6.2 Effectiveness of Topic Adaptation

To test the effectiveness of topic adaptation and answer Q2, we modelled the following experiment. We trained a pretrained forumBERT model and a sports topic-adapted forumBERT model on the comment moderation endtask using varying number of labelled endtask examples. Due to high variance in few shot results we average the results over 3 runs and measure and plot the standard deviation to generate reliable insights. The results of our experiment are shown in Figure 5.

From Figure 5 we see that the pretrained forumBERT model slightly outperforms the topic-adapted forumBERT model in very few shot learning situations ($<$ 300 training examples). However, it can be seen that in the range of 315-1000 labelled training examples, the topic-adapted forumBERT model performs as well as the pretrained forumBERT model. Beyond this ($>$ 1000 labelled training examples), the performance of topic adapted forumBERT clearly exceeds the pretrained forumBERT without topic adaptation. We also observe that the performance of both models starts converging beyond 10000 training examples.

From this experiment, we conclude that the effectiveness of topic adaptation reduces as the number of labelled training examples increase in the downstream task since labelled training examples consist of both task information and topic information, they provide much richer information to the model. As our experiment shows, with more than 10000 labelled training examples the advantage of using a topic adapted model diminishes.

7 Conclusion

In this paper, we introduced forumBERT, a simple architecture designed to determine comment's relevance in a discourse using 2 weight shared BERT models and a late fusion technique on BERT comment and article representations. Also, to mitigate the problem of portability of large NLP language models to niche language domains (in our case small news forums), we adopted a topic adaptation technique to learn better BERT representations.

We empirically showed that forumBERT outperforms all other baseline models on the NDR dataset. Our adaptation significantly outperforms the standard BERT model, improving the F1 scores from 0.475 to 0.513 (an 8% relative increase) on the sports topic dataset and an F1 score improvement from 0.465 to 0.490 (a 5.3% relative increase) on the culture topic dataset. The model also outperforms all existing advanced baseline results on the OMP dataset. Further analysis also shows the importance of topic adaptation as a function of labelled training examples. We would like to extend the application of forumBERT to other NLP tasks applications involving context dependent classification. Our implementation uses PyTorch (Paszke et al., 2019) and is publicly available.[5]

Acknowledgments. This work was partly funded by Hamburg's ahoi.digital program in the Forum 4.0 project. We would also like to thank German broadcaster Norddeutscher Rundfunk (NDR) for giving us access to an extensive collection of moderated NDR.de user comments.

[5]See https://github.com/ayushyadav99/forumBERT.

References

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20(3):273–297.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14, pages 1188–1196, Beijing, China.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Lake Tahoe, Nevada, USA.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Vancouver, BC, Canada.

John Pavlopoulos, Prodromos Malakasiotis, and Ion Androutsopoulos. 2017a. Deep learning for user comment moderation. In *Proceedings of the First Workshop on Abusive Language Online*, pages 25–35, Vancouver, BC, Canada. Association for Computational Linguistics.

John Pavlopoulos, Prodromos Malakasiotis, Juli Bakagianni, and Ion Androutsopoulos. 2017b. Improved abusive comment moderation with user embeddings. In *Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism*, pages 51–55, Copenhagen, Denmark. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China.

Alexander Rietzler, Sebastian Stabinger, Paul Opitz, and Stefan Engl. 2020. Adapt or get left behind: Domain adaptation through BERT language model finetuning for aspect-target sentiment classification. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 4933–4941, Marseille, France.

Julian Risch and Ralf Krestel. 2018. Delete or not delete? semi-automatic comment moderation for the newsroom. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 166–176, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Dietmar Schabus and Marcin Skowron. 2018. Academic-industrial perspective on the development and deployment of a moderation system for a newspaper website. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, Miyazaki, Japan. European Languages Resources Association (ELRA).

Dietmar Schabus, Marcin Skowron, and Martin Trapp. 2017. One million posts: A data set of german online discussions. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 1241–1244, Tokyo, Japan.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Long Beach, CA, USA.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, ukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation.

Hu Xu, Bing Liu, Lei Shu, and Philip Yu. 2019. BERT post-training for review reading comprehension and aspect-based sentiment analysis. 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 2324–2335, Minneapolis, Minnesota. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. XLNet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems 32*, pages 5753–5763. Vancouver, BC, Canada.

Da Yin, Tao Meng, and Kai-Wei Chang. 2020. SentiBERT: A transferable transformer-based architecture for compositional sentiment semantics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3695–3706, Online.