Aspect-Target Sentiment Classification for Cyberbullying Detection

Stanford CS224N Custom Project

Cong Kevin Chen †
ckchen95@stanford.edu

Sharan Ramjee †
sramjee@stanford.edu

Joseph Wang †
wangjoe@stanford.edu

Abstract

Cyberbullying detection is a challenging task to tackle, given the complex nature of the problem and the lack of Natural Language Processing (NLP) literature when it comes to addressing this issue. For a piece of text to be considered as cyberbullying, it not only has to be associated with a negative sentiment, but must also be targeted. This motivates the use of Aspect-Target Sentiment Classification (ATSC), which evaluates the sentiment of a given piece with respect to an aspect-target in the text. In particular, we make use of the BERT-ADA transformer architecture, fine-tuned on the hatespeech-twitter dataset, to demonstrate its superior ability in detecting cyberbullying in comparison to other state-of-the-art sentiment analysis baselines. Additionally, we make use of Named Entity Recognition (NER) in order to extract aspect-targets from tweets that do not explicitly "@" username handles of other users. The code is available on GitHub: `https://github.com/sharanramjee/cyberbullying-atsc`

1 Introduction

Sentiment Analysis [1] is an important task in Natural Language Processing and has a wide rage of real-word applications. Typical Sentiment Analysis methods focus on predicting the polarity of a given sentence i.e. whether the sentence reflects a positive or negative sentiment. Building on this, a more complex subtask in Sentiment Analysis is predicting the sentiment towards a certain aspect or word mentioned in a sentence. This subtask is known as Aspect-Based Sentiment Analysis (ABSA) [2]. ABSA comes in two variants that are implemented as two-step procedures as illustrated in Fig. 1.

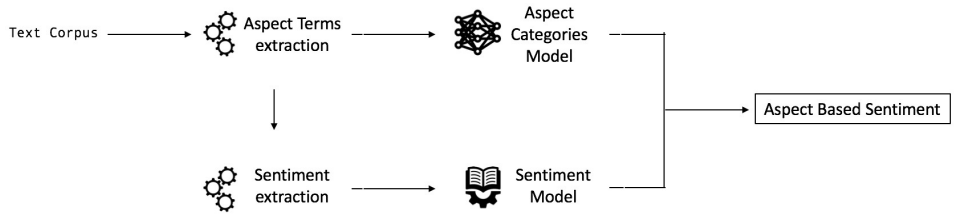


Figure 1: Variants of Aspect-Based Sentiment Analysis

The first variant consists of Aspect-Category Detection (ACD) [3], which uses an Aspect Terms Extraction model to extract the aspect-targets in the sentence followed by Aspect-Category Sentiment Classification (ACSC) [4], which uses an Aspect Categories model to categorize the aspect-targets. The second variant consists of Aspect-Target Extraction (ATE) [5], which uses a Sentiment Extraction model to obtain sentiments for aspect-targets in the sentence followed by Aspect-Target Sentiment Classification (ATSC) [6], which uses a Sentiment Model to estimate the sentiment of the sentence with respect to a specific aspect-target in the sentence.

† Department of Computer Science, Stanford University

In the context of our project, which is the application of Sentiment Analysis for cyberbullying detection, we focus on the second variant, primarily, ATSC. ATSC, as mentioned earlier, is the task of determining the polarity associated with an aspect-target or a specific word in a sentence. For instance, as examined by He *et al.*[7], given an input sentence "The appetizers are ok, but the service is slow." and an aspect-target "appetizers", the outputted sentiment of the sentence towards the aspect-target "appetizers" would be neutral, regardless of the overall sentiment of the sentence. However, in the same example if the given aspect-target is "service", the outputted sentiment would be negative.

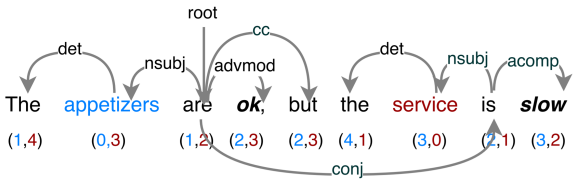


Figure 2: Illustration of a dependency tree of an example sentence. The numbers indicate the distances from the word to the two targets respectively along the syntactic path. [7]

ATSC is a particularly challenging task since the targeted sentiment score towards a particular aspect-target in a sentence can be drastically different from the overall sentiment of the sentence in consideration. As illustrated in Fig. 2 and as investigated by He *et al.*[7], the sentiment of the sentence towards a certain aspect-target drifts towards the overall sentiment of the sentence as a function the distances from the word to the aspect-targets along the syntactic path give by the dependency tree. Therefore, it requires the Sentiment Analysis algorithm to pay particular attention to the local context of the aspect-target, which motivates the use of transformer-based models with attention mechanisms such as the Bidirectional Encoder Representations from Transformers (BERT) [8].

As investigated by Zhao *et al.*[9], cyberbullying detection is a challenging task to tackle, given the complex nature of the problem and the lack of literature with respect to Natural Language Processing when it comes to addressing this issue. Delving deeper into the problem, online forums such as Twitter and Reddit are popular sites for trolls to thrive, especially in this modern era, where it is extremely easy to spread negativity. In particular, given the nuances of internet-chat jargon, it is hard to determine whether a negative piece of text is negative in general or is directed towards a person or a group. For a piece of text to be considered cyberbullying, it not only has to be associated with a negative sentiment, but must also be targeted. For instance, a retweet towards an original tweet can be negative, however, this cannot simply be classified as cyberbullying towards the author of the original tweet using conventional Sentiment Analysis methods because while such a method can determine the negative sentiment of the retweet, it cannot determine whether or not the negative sentiment was directed at the original tweet. This motivates the use of Aspect-Target Sentiment Classification for cyberbullying detection. In particular, we make use of a pre-trained BERT-ADA [10] transformer model, fine-tune it on the hatespeech-twitter dataset [11], and use it for the ATSC downstream task.

Aspect-target extraction is a simple sentence parsing task when it comes to pieces of text like certain tweets, where the aspect-targets can be obtained using the "@" username handles that the tweet mentions. However, there are also tweets that cyberbully without explicitly "@"ing someone, in addition to other platforms where cyberbullying can occur without explicitly mentioning the username handles of other users. This motivates the use of Named Entity Recognition (NER) [12] for aspect-target extraction. In particular, we make use a pre-trained BERT-large [8] transformer model fine-tuned on the ConLL-2003 dataset [13] for NER, as examined in the subsequent sections.

2 Related Work

Xu et al. [14] argues BERT can be fine-tuned on domain-specific corpora for improved performance on ATSC tasks, and hones in on its impact on two specific datasets: Yelp restaurant reviews [15] and Amazon laptop reviews [16]. They make up for the lack of training data in the target domains by using various permutations of cross-domain and in-domain fine-tuning across the restaurant and laptop datasets. The proposed model, BERT-ADA, achieves state of the art accuracy (87.14%) when trained in-domain (i.e. fine-tuned, trained, and tested on data from the same domain) on the restaurant dataset only, while making slightly incremental improvement upon the benchmark model performance [14]. Cross-domain (fine-tuning and testing on either the restaurant or laptop domain while training on the

other) and joint-domain (training on both domains and testing on either one) training are also able to match or exceed the baselines set by the BERT-base [17] and XLNet [18] models.

Agrawal *et al.*[19] apply a variety of CNN and LSTM models to datasets from three different sources: Formspring, Twitter, and Wikipedia, containing examples labeled as either "bullying" or "non-bullying". Due to the class imbalance in the dataset, "bullying" examples were oversampled by a factor of 3 during training in comparison to the "non-bullying" examples. This allowed all the models to output dramatically higher F-1 scores than the standard machine learning baselines with n-gram word representations: logistic regression, support vector machine, random forest, naive Bayes. The strong performance of bidirectional LSTMs with attention led us to the hypothesis that a BERT-based approach would perform well on our objective. The authors also determined that for cross-domain classification, taking word embeddings (the highest performing type is GloVe) learned on one dataset (e.g. Wikipedia) and fine-tuning them on another (e.g. Twitter) could further improve test accuracy compared to training on Wikipedia posts and directly attempting to infer on tweets. This finding inspired some of the fine-tuning techniques we ultimately attempted.

Yadav *et al.*[20] train BERT-base on the aforementioned datasets and use Agrawal *et al.*[19] as a baseline. Interestingly, they did not run any experiments on the Twitter dataset available to them, nor did they report fine-tuning results of any sort, leaving room for us to expand on a larger dataset. When Yadav *et al.*[20] oversampled the "bullying" label for the Formspring dataset, they beat the F-1 score that Agrawal *et al.*[19] achieved on the similarly oversampled dataset using bidirectional LSTMs with attention.

3 Technical Approach

3.1 Baselines

We make use of pre-trained state-of-the-art non-ATSC Sentiment Analysis transformer models for our baselines. Specifically, we compare the performance of the ATSC-based BERT-ADA transformer model with two general purpose models - BERT-base [21] and DistilBERT [22], as well as two models specialized for Twitter data - twitter-roBERTa-base [23] and BERTweet [24]. BERT-base and twitter-roBERTa-base each output either positive or negative sentiment, while DistilBERT and BERTweet can output positive, neutral, or negative. In each case, when the model classifies a tweets as negative, we equate that to the "Cyberbullying" label in our dataset (see Data section).

BERT-base BERT-base is the smaller of the two architectures (110 million parameters as opposed to 340 million) proposed by Devlin *et al.*[8]. This baseline test takes the implementation of BERT-base used in the TextAttack sub-package of the HuggingFace transformers library [21], taking in a tweet as input and outputting the cyberbullying prediction.

DistilBERT DistilBERT is a lightweight version of BERT with only half as many layers and 40% fewer parameters, thus speeding up inference by 60% on average. DistilBERT under-performs in comparison to BERT by 0.6% on IMDb classification accuracy [25] and by 2.7 points according to the F1 score on the SQuAD question-answering task [26], which makes it the gold standard when it comes to computationally efficient alternatives to BERT.

twitter-roBERTa-base twitter-roBERTa-base is a roBERTa-base [27] architecture that has been pre-trained on a corpus of 60 million tweets and applied to a variety of social media specific tasks such as sentiment analysis, hate speech detection, irony detection, and stance analysis. twitter-roBERTa-base achieved validation accuracies of 79.6% on hate speech detection and 77.7% on offensive language detection, the two tasks most relevant to our cyberbullying objective.

BERTweet BERTweet is the first NLP model to be trained exclusively on tweets. It follows the same architecture as BERT-base[17] and has been pre-trained in the same manner as roBERTa-base [27] on a corpus of 868 million tweets. For the sentiment polarity classification task, the model was fine-tuned on the SemEval 2017-4A training set[28] and achieved an accuracy of 72.0%. This beat roBERTa-base, which was not fine-tuned, on the same dataset, despite roBERTa-base being pre-trained on twice as many GB of data.

3.2 Named Entity Recognition

Named Entity Recognition (NER) [12] is a fundamental task in information extraction. Souza *et al.*[29] formally define NER as the task of detecting and classifying mentions of domain-relevant entities such as persons and organizations. For instance, given a tweet "Ousted WeWork founder Adan Neumann lists his Manhattan penthouse for $37.5 million.", the NER model aims to detect the tokens "WeWork", "Adam Neumann", "Manhattan", and "$37.5 million" and classify them as "organization", "person", "location", and "monetary value", respectively, as illustrated in Fig. 3.

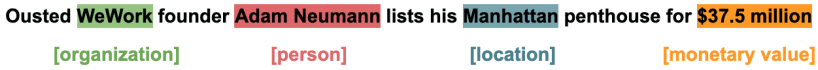


Figure 3: Example of NER on a tweet

As mentioned earlier, aspect-target extraction is a simple sentence parsing task in scenarios where cyberbullying occurs online by directly mentioning another user's username handle. For instance, the tweet "@BarackObama is the worst president ever" can be easily parsed to extract "@BarackObama" as the aspect-target using the "@" symbol. However, cyberbullying can take many forms in online forums such as Twitter and Reddit. In particular, cyberbullying can occur without explicitly "@"ing someone. In order to address such scenarios, we extend Aspect-Target Sentiment Classification by incorporating NER into it in order to detect aspect-targets in such pieces of text.

In particular, we make use of a BERT-large [8] transformer model for NER, pretrained on the CoNLL-2003 [13] NER dataset. BERT-large consists of 24 layers, 1024 hidden dimensions per token, and 16 attention heads, which results in a total of 340 million parameters. We further fine-tune the BERT-large model on the Broad Twitter Corpus (BTC) [30] to facilitate better NER performance on domain-specific twitter data, as detailed by Sun *et al.*[31].

For the output layer, we feed the final hidden representation h_i of each toekn i into the softmax function. The probability P is computed as follows:

$$P(t|h_i) = softmax(W_0 h_i, b_0)$$

where $T = \{O, B-MIS, I-MIS, B-PER, I-PER, B-ORG, I-ORG, B-LOC, I-LOC\}, t \in T$, as outlined by Devlin *et al.*[32]. Here, W_o and b_0 are the weight parameters and during fine-tuning, the default training configuration along with the default hyperparameters as specified by Devlin *et al.*[32] with the categorical cross-entropy loss function were used since we found this configuration to work best empirically. Furthermore, other NER models such as BERT-base [17] and T5 [33] were also considered as alternatives. However, as given in Table. 1, we found BERT-large to outperform them across all the BTC dataset metrics as detailed by Derczynski *et al.*[30].

Table 1: Evaluation of NER models across BTC dataset metrics

Method	dev-f1	dev-precision	dev-recall	test-f1	test-precision	test-recall
T5	0.917	0.916	0.920	0.889	0.886	0.893
BERT-base	0.873	0.871	0.976	0.845	0.844	0.847
BERT-large	**0.951**	**0.950**	**0.953**	**0.913**	**0.907**	**0.919**

3.3 Aspect-Target Sentiment Classification

We make use of a pre-trained BERT-base [8] architecture, which consists of 12 layers, 768 hidden dimensions per token, and 12 attention heads, which results in a total of 110 million parameters. The ATSC task will be approached using a two-step procedure, as illustrated in Fig. 4. In the first step, the pre-trained weights of the language model are fine-tuned in a self-supervised way on the domain-specific corpus, which, in our case, is the hatespeech-twitter dataset [11]. In the second step, the fine-tuned language model is trained in a supervised way on the ATSC end-task.

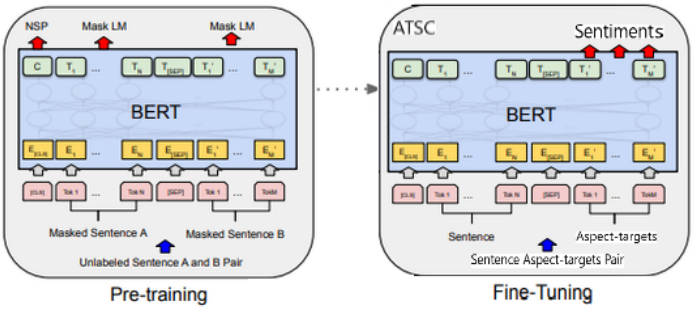


Figure 4: BERT Language Model fine-tuning and Aspect-Target Sentiment Classification Training

BERT Language Model Fine-Tuning Fine-tuning BERT involves optimizing an objective that consists of two parts: the masked language model objective, where the model learns to predict randomly masked tokens from their context, and the next-sequence prediction objective, where the model learns to predict if a sequence B would naturally follow the previous sequence A in order to better capture long-term dependencies. The training input representation for language model fine-tuning consists of two sequences s_A and s_B in the format of "[CLS]$_{s_A}$ [SEP]$_{s_B}$ [SEP]", where [CLS] is a dummy token used for down-stream classification and [SEP] are separator tokens in accordance to [34].

As for the masked language model objective, the sequences A and B have tokens randomly masked out in order for the model to learn to predict them. This allows domain-specific fine-tuning to alleviate bias from pre-training. For instance, the [MASK] in the sentence "The touchscreen is an [MASK] device" could take on the values of "input" or "amazing" based on the corpus domains of fact-based, and review-based, respectively.

As for the next-sentence prediction objective, the model is trained to predict whether sequence B follows sequence A. If so, the sequences are jointly sampled from the same document in the order they appear naturally and otherwise, the sequences are sampled randomly from the training corpus.

Aspect-Target Sentiment Classification Training The ATSC task aims to classify the sentiment polarity as positive, neutral, or negative with respect to an aspect-target. As previously mentioned, the inputs to the classifier are a tokenized sentence $s = s_{1:n}$ and an aspect-target $t = s_{j:j+m}$ contained in the sentence, where $j < j + m \leq n$. Again, the input is transformed into a format that is compatible with BERT sequence-pair classification tasks: "[CLS]$_s$ [SEP]$_s$ [SEP]". The last hidden representation of the [CLS] token is given by $h_{[CLS]} \in \mathbb{R}^{768\times 1}$ since the BERT architecture structurally maintains the position fo the token embeddings after each multi-head attention layer. The number of sentiment polarity classes is 3 (positive, neural, negative) and thus, a distribution $p \in [0, 1]^3$ over these classes is predicted using a fully-connected layer with 3 output neurons on top of $h_{[CLS]}$, followed by a softmax activation function:

$$p = \text{softmax}(W \cdot h_{[CLS]} + b)$$

where $b \in \mathbb{R}^3$ and $W \in \mathbb{R}^{3\times 768}$. Furthermore, the cross-entropy loss is used as the training loss. Here, BERT is used for classifying the sentiment polarities in a way that is equivalent to how BERT is used for sequence-pair classification tasks performed in the original paper by Devlin *et al.*[8].

4 Experiments

4.1 Data

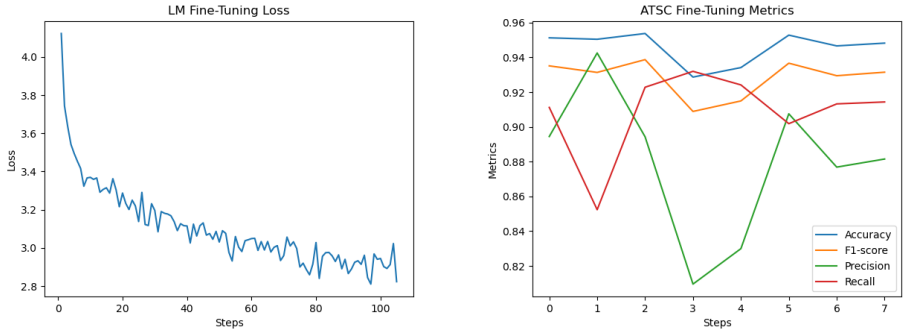
The hatespeech-twitter dataset curated by Founta *et al.*[11] consists of approximately 100k tweets collected via crowd-sourcing. Each tweet bears one of the following labels: "Normal" (54%), "Abusive" (27%), "Spam" (14%), "Hateful" (5%). Due to the class imbalance, we combined the "Abusive" and "Hate" categories into a single label, "Cyberbullying", while discarding all tweets labeled "Spam", which are likely to be generated by a bot. This left us approximately 86,000 tweets in the dataset, with 63% labelled "Normal" and 37% labelled "Cyberbullying".

For preprocessing, we remove the emojis and URLs from all tweets replace the HTML codes with the character they represent ('&' = '&', '>' = '>', '<' = '<'). Here, references to other "@" username handles are treated as the aspect-targets. There are approximately 48,000 tweets that contain "@"s, which is roughly half of the dataset and will be used for the ATSC task with a train-validation-test split of 80-10-10. For data entries that do not contain "@", we use Named Entity Recognition (NER) to extract the aspect targets and assign the sentiment of the overall sentence as its sentiment. We also use two separate sets of data for the two different fine-tuning steps: Language Model (LM) fine-tuning and Aspect-Target Sentiment Classification (ATSC) fine-tuning. For LM fine-tuning, we concatenate all the tweets that do not contain "@" into one single text file to train the model on twitter-generic corpus, which is roughly half of the overall dataset. For ATSC fine-tuning, we use the dataset that only contain "@" symbols, as mentioned earlier. An example entry would be `{sentence: "@Trump is a moron", term: "Trump", polarity = "negative"}`. We split the data in this manner in order to enable the model to learn generic tweet features in the LM fine-tuning process and learn target-specific tweet features in the ATSC fine-tuning process.

4.2 Evaluation Method

For the LM fine-tuning step, we use the loss function of the masked language model to keep track of the learning curve. The BERT loss function takes in the prediction of the masked value and calculate cross entropy loss, or log loss, based on the output of the softmax layer. For ATSC fine-tuning, because of the considerable amount of class imbalance present in the dataset, the evaluation metrics of choice that we will use to compare our model to the baselines on the test set are: Accuracy, Precision, Recall, and F1-score. Considering the nature of the problem, we would like to ensure that we capture as many true positives as possible, even if capturing false positives may be inconvenient and as such, we will pay close attention to the Recall scores of the model and the baselines.

4.3 Experimental Details



(a) LM fine-tuning loss (b) ATSC fine-tuning metrics

Figure 5: BERT-ADA fine-tuning curves

We make use of a pre-trained BERT-base [8] architecture, which consists of 12 layers, 768 hidden dimensions per token, and 12 attention heads (110 million parameters) and fine-tune it on the hate-speech train set, as detailed in 3.3. The Language Model fine-tuning process took about 2.5 hours to fine-tune. We then proceeded to fine-tune the BERT-base model on the downstream ATSC task. The ATSC fine-tuning process took about 1.5 hour to fine-tune. Fine-tuning resulted in a model size of 500MB with a vocabulary size of 30522. Both fine-tuning processed we performed on a Tesla K80 GPU. We used a gradient accumulation steps of 1, which allows the model to learn after every step, thus facilitating the model to capture more granular details of the text. The detailed fine-tuning configurations are specified in Table 2 and the BERT-ADA fine-tuning curves for LM fine-tuning and ATSC fine-tuning are given in Figs. 5(a) and 5(b), respectively.

Table 2: Training configuration for LM finetuning and task-specific finetuning

Finetune Step	L.Rate	Optimizer	Decay	MaxSeqLen	Epochs	BatchSize	TrainSteps	WarmUp
Language Model	$3e^{-5}$	Adam	0.01	256	2	16	N/A	0
Task-specific	$3e^{-5}$	Adam	0.0	128	1	32	1200	120

4.4 Results

The quantitative analysis of BERT-ADA (before and after fine-tuning) and the baselines is given in Table 3. BERT-ADA outperforms BERT-base and DistilBERT on the test set across all the evaluation metrics without any fine-tuning. However, twitter-RoBERTa-base and BERTweet outperform BERT-ADA (before fine-tuning) in terms of accuracy, precision, and F1-score. This can be attributed to the fact that both of these baselines are fine-tuned on a domain-specific twitter corpus - the tweeteval dataset [28]. Most importantly, BERT-ADA (before and after fine-tuning) outperforms all the baselines in terms of the recall score, which, for our application, is the most important metric of consideration. This can be attributed the to fact that the baselines considered do not make use of ATSC in order to evaluate the sentiment of the tweet with respect to the specific aspect-target(s) in the tweet, which may cause them to miss true positives if the overall sentiment of the tweet is positive.

Table 3: Quantitative analysis of BERT-ADA and baselines

Method	NER	Accuracy	Precision	Recall	F1 Score
BERT-base	–	0.608	0.888	0.432	0.582
DistilBERT	–	0.600	0.892	0.414	0.565
twitter-roBERTa-base	–	0.825	0.944	0.768	0.847
BERTweet	–	0.784	0.848	0.763	0.817
BERT-ADA	No	0.662	0.627	0.882	0.733
BERT-ADA (fine-tuned)	No	0.920	0.868	0.882	0.940
BERT-ADA	Yes	0.694	0.695	0.917	0.791
BERT-ADA (fine-tuned)	Yes	**0.952**	**0.936**	**0.944**	**0.940**

Furthermore, we perform an ablation study with and without using NER to obtain aspect-targets from tweets in the test set in order to evaluate the effect of NER on the performance of the models. This ablation study does not consider the baselines since they do not make use of aspect-targets to classify the sentiment of tweets. We found that the performance of BERT-ADA (before and after fine-tuning) improves significantly after the application of NER as observed in Table 3. This can be attributed to the ability of BERT-ADA to more accurately classify the sentiment of a tweet with respect to a certain aspect-target in comparison classifying the sentiment of the tweet as a whole, which arises out of the ATSC fine-tuning that was performed on the BERT-ADA model.

5 Analysis

Qualitative evaluation of BERT-ADA in comparison to the other baselines is given in Table. 4. Here, the **Ref.** column specifies the tweet reference from the dataset, the **aspect** column specifies the aspect-targets used by BERT-ADA for sentiment analysis, and the **Gold** column specifies the ground truth sentiment ("+" for positive and "–" for negative) of the tweet with respect to the aspect-target. It is important to note here that the baselines considered do not make use of the aspect-targets and merely perform general Sentiment Analysis on the tweet, as a whole. Furthermore, we include the performance of BERT-ADA before fine-tuning and after fine-tuning (f-t) in order to observe the performance gain achieved as a result of fine-tuning.

For tweets 19949 and 18089, we see that BERT-ADA (before and after fine-tuning) and all the baselines correctly classify the sentiments of the tweets. The negative sentiment of tweet 19949 mainly stems from the word "[obscenity]", which has been censored for this paper, whereas the positive sentiment of tweet 18089 mainly stems from the word "Proud". These words sway all the models to classify these tweets correctly, regardless of whether or not the aspect-target has been considered, as evidenced by all the baselines correctly classifying the sentiments.

For tweets 28880 and 90190, we see that BERT-ADA (fine-tuned) is the only model that correctly classifies the sentiments of the tweets. Tweet 28880 contains the words "mad" and "hell" which led to baselines incorrectly classifying the sentiment as negative, despite the fact that the tweet is clearly not an attempt to cyberbully "@Kobe". Tweet 90190, while does not explicitly use words that could sway the models towards a negative sentiment, is a form of microaggression as detailed by Breitfeller *et al.*[35] and BERT-ADA (fine-tuned) was the only model capable of picking up this nuance.

Ref.	Tweet	Aspect	BERT-base	DistilBERT	roBERTa-base	BERTweet	BERT-ADA	BERT-ADA (f-t)	**Gold**
19949	@RonPaul is a god damn joke who has no business being in politics!!! What a [obscenity][1] moron!!!	@RonPaul	–	–	–	–	–	–	–
18089	@JenniferLawrence's New Dior Ad Campaign Would Make @Katniss @Everdeen Proud	@Jennifer Lawrence	+	+	+	+	+	+	+
28880	Damn @Kobe was mad as hell	@Kobe	–	–	–	–	–	+	+
90190	RT @honeybwee: he is so hot/cute	@honeybwee	+	+	+	+	+	–	–
79640	If the raiders get @Lynch.. i might aswell just buy my superbowl tickets RN even if @Car breaks both his legs this season.. LYNCH IS A SAVAGE	@Lynch	–	–	–	–	–	–	+
37284	@coollikedan acucanu i wanna know how to pronounce your name tbh also dodie gif idk her but she's cute	@coollikedan	+	+	+	+	+	+	–

Table 4: Qualitative analysis of BERT-ADA and baselines

For tweets 79640 and 37284, we see that none of the models correctly classify the sentiments of the tweets. Tweet 79640 contains the word "SAVAGE", which is a popular slang with positive connotations. However, none of the models were able pick up this nuance due to the use of the word in negative contexts in various datasets the models picked up during pre-training. Tweet 37284, similar to tweet 90190, is a form of microaggression [35] and as such, all the baselines were unable to pick up on this. Interestingly, BERT-ADA (fine-tuned) as also unable to pick up on this nuance. We found that BERT-ADA is unable to correctly detect and classify microaggressions in longer tweets, perhaps due to the lack of a more sophisticated attention mechanism which is specialized for microaggression detection as examined by Tsvetkov *et al.*[36].

6 Conclusion

Cyberbullying detection is a challenging task, given the complex nature of the problem and the lack of Natural Language Processing literature when it comes to addressing this issue. Through our extensive experiments, we showed that Aspect-Target Sentiment Classification models such as BERT-ADA outperform general Sentiment Analysis models for cyberbullying detection. Furthermore, we addressed the issue of aspect-target detection in pieces of text where the aspect-targets are not apparent through the use of Named Entity Recognition models.

However, Aspect-Target Sentiment Classification models are still far from perfect, as evidenced by the qualitative analysis performed in the previous section. They are unable to detect nuances in internet-chat jargon, especially when the domain of the pre-training corpus sufficiently differs from that of the fine-tuning corpus. Furthermore, cyberbullying can also occur in the form of microaggressions and are challenging for such models to address, given the drawbacks of such models - lack of incorporating socio-cultural knowledge representation, lack of demoting social biases during modeling, and lack of interpretability in evaluating and characterizing model behaviors. Cyberbullying detection is a pressing issue to tackle, especially in this modern era, and perhaps future work in this field can address these shortcomings of Aspect-Target Sentiment Classification models for better cyberbullying detection.

[1] tweet has been censored for this paper

References

[1] Ronen Feldman. Techniques and applications for sentiment analysis. *Communications of the ACM*, 56(4):82–89, 2013.

[2] Hai Ha Do, PWC Prasad, Angelika Maag, and Abeer Alsadoon. Deep learning for aspect-based sentiment analysis: a comparative review. *Expert Systems with Applications*, 118:272–299, 2019.

[3] Kim Schouten, Onne Van Der Weijde, Flavius Frasincar, and Rommert Dekker. Supervised and unsupervised aspect category detection for sentiment analysis with co-occurrence data. *IEEE transactions on cybernetics*, 48(4):1263–1275, 2017.

[4] Jiajun Cheng, Shenglin Zhao, Jiani Zhang, Irwin King, Xin Zhang, and Hui Wang. Aspect-level sentiment classification with heat (hierarchical attention) network. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 97–106, 2017.

[5] Chuhan Wu, Fangzhao Wu, Sixing Wu, Zhigang Yuan, and Yongfeng Huang. A hybrid unsupervised method for aspect term and opinion target extraction. *Knowledge-Based Systems*, 148:66–73, 2018.

[6] Dehong Ma, Sujian Li, Xiaodong Zhang, and Houfeng Wang. Interactive attention networks for aspect-level sentiment classification. *arXiv preprint arXiv:1709.00893*, 2017.

[7] Ruidan He, Wee Sun Lee, Hwee Tou Ng, and Daniel Dahlmeier. Effective attention modeling for aspect-level sentiment classification. In *Proceedings of the 27th international conference on computational linguistics*, pages 1121–1131, 2018.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[9] Rui Zhao, Anna Zhou, and Kezhi Mao. Automatic detection of cyberbullying on social networks based on bullying features. In *Proceedings of the 17th international conference on distributed computing and networking*, pages 1–6, 2016.

[10] Alexander Rietzler, Sebastian Stabinger, Paul Opitz, and Stefan Engl. 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, May 2020. European Language Resources Association.

[11] Antigoni-Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. Large scale crowdsourcing and characterization of twitter abusive behavior. In *11th International Conference on Web and Social Media, ICWSM 2018*. AAAI Press, 2018.

[12] David Nadeau and Satoshi Sekine. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26, 2007.

[13] Erik F Sang and Fien De Meulder. Introduction to the conll-2003 shared task: Language-independent named entity recognition. *arXiv preprint cs/0306050*, 2003.

[14] Hu Xu, Bing Liu, Lei Shu, and Philip S. Yu. 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*, jun 2019.

[15] Nabiha Asghar. Yelp dataset challenge: Review rating prediction. *arXiv preprint arXiv:1605.05362*, 2016.

[16] Ruining He and Julian McAuley. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, page 507–517, Republic and Canton of Geneva, CHE, 2016. International World Wide Web Conferences Steering Committee.

[17] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[18] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. Xlnet: Generalized autoregressive pretraining for language understanding, 2020.

[19] Sweta Agrawal and Amit Awekar. Deep learning for detecting cyberbullying across multiple social media platforms, 2018.

[20] J. Yadav, D. Kumar, and D. Chauhan. Cyberbullying detection using pre-trained bert model. In *2020 International Conference on Electronics and Sustainable Communication Systems (ICESC)*, pages 1096–1100, 2020.

[21] John X. Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in nlp, 2020.

[22] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

[23] Francesco Barbieri, Jose Camacho-Collados, Leonardo Neves, and Luis Espinosa-Anke. Tweeteval: Unified benchmark and comparative evaluation for tweet classification, 2020.

[24] Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. BERTweet: A pre-trained language model for English Tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14, 2020.

[25] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.

[26] Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for SQuAD. In *Association for Computational Linguistics (ACL)*, 2018.

[27] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[28] Sara Rosenthal, Noura Farra, and Preslav Nakov. Semeval-2017 task 4: Sentiment analysis in twitter. In *Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017)*, pages 502–518, 2017.

[29] Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. Portuguese named entity recognition using bert-crf. *arXiv preprint arXiv:1909.10649*, 2019.

[30] Leon Derczynski, Kalina Bontcheva, and Ian Roberts. Broad Twitter corpus: A diverse named entity recognition resource. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1169–1179, Osaka, Japan, December 2016. The COLING 2016 Organizing Committee.

[31] Cong Sun and Zhihao Yang. Transfer learning in biomedical named entity recognition: An evaluation of bert in the pharmaconer task. In *Proceedings of The 5th Workshop on BioNLP Open Shared Tasks*, pages 100–104, 2019.

[32] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

[33] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*, 2019.

[34] Chi Sun, Luyao Huang, and Xipeng Qiu. Utilizing bert for aspect-based sentiment analysis via constructing auxiliary sentence. *arXiv preprint arXiv:1903.09588*, 2019.

[35] Luke Breitfeller, Emily Ahn, David Jurgens, and Yulia Tsvetkov. Finding microaggressions in the wild: A case for locating elusive phenomena in social media posts. 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 1664–1674, 2019.

[36] Yulia Tsvetkov, Vinodkumar Prabhakaran, and Rob Voigt. Socially responsible natural language processing. In *Companion Proceedings of The 2019 World Wide Web Conference*, pages 1326–1326, 2019.