fBERT: A Neural Transformer for Identifying Offensive Content

Diptanu Sarkar[1], Marcos Zampieri[1], Tharindu Ranasinghe[2], Alexander Ororbia[1]
[1]Rochester Institute of Technology, USA
[2]University of Wolverhampton, UK
ds9297@rit.edu

Abstract

Transformer-based models such as BERT, XLNET, and XLM-R have achieved state-of-the-art performance across various NLP tasks including the identification of offensive language and hate speech, an important problem in social media. In this paper, we present fBERT, a BERT model retrained on SOLID, the largest English offensive language identification corpus available with over 1.4 million offensive instances. We evaluate fBERT's performance on identifying offensive content on multiple English datasets and we test several thresholds for selecting instances from SOLID. The fBERT model will be made freely available to the community.

1 Introduction

To cope with the spread of offensive content and hate speech online, researchers have worked to develop automatic methods to detect such posts automatically. Early efforts included methods that used various linguistic features in tandem with linear classifiers (Malmasi and Zampieri, 2017) while, more recently, deep neural networks (DNNs) (Ranasinghe et al., 2019), transfer learning (Wiedemann et al., 2020; Abu Farha and Magdy, 2020), and pre-trained language models (Liu et al., 2019a; Ranasinghe and Zampieri, 2020, 2021) have led to even further advances. As evidenced in recent competitions, the performance of these models varies with the sub-task that they are designed to address as well as the datasets used to train them. For example, classical statistical learning models such as the support vector machine (SVM) have outperformed neural transformers in hate speech detection at HatEval 2019 (Basile et al., 2019) and in aggression detection at TRAC 2018 (Kumar et al., 2018). However, for both of these tasks in OffensEval 2019 and 2020 (Zampieri et al., 2019b, 2020), which focused on the identification of more general offensive language identification, pre-trained transformer-based models such as BERT (Devlin et al., 2019) outperformed other neural architectures and statistical learning methods.

The introduction of representations learned through the bidirectional encoding inherent to neural transformers (BERT) (Devlin et al., 2019) has been driven much progress in areas in NLP such as language understanding, named entity recognition, and text classification. The base model is pre-trained on a large English corpus, e.g., Wikipedia, BookCorpus (Zhu et al., 2015), using unsupervised masked language modeling and next sentence prediction objectives to adjust the model weights. Various other transformer-based models have also been introduced including RoBERTa (Liu et al., 2019b), XLNet (Yang et al., 2019), XLM-R (Conneau et al., 2019). All of these models, however, are trained on general-purpose corpora for better language understanding, generally lacking domain-specific knowledge. To cope with this limitation, more recently, domain-specific models have been trained and/or fine tuned to different domains such as finance (FinBERT) (Araci, 2019), law (LEGAL-BERT) (Chalkidis et al., 2020), scientific texts (SciBERT) (Beltagy et al., 2019), and microblogging (BerTweet) (Nguyen et al., 2020).

Caselli et al. (2021) recently released HateBERT, a BERT transformer model for abusive language detection trained on the Reddit Abusive Language English dataset (RAL-E). HateBERT achieves competitive performance on a few benchmark datasets but relies heavily on manually annotated labels. Moreover, HateBERT was trained on a task-specific dataset (aggression) instead of a more general dataset that encompasses multiple types of offensive language (e.g. hate speech, cyberbullying, profanity) like the popular OLID (Zampieri et al., 2019a) used in OffensEval 2019 at SemEval.

To address this gap, in this study, we present fBERT, a pre-trained BERT model trained on SOLID (Rosenthal et al., 2021), a recently released

Findings of the Association for Computational Linguistics: EMNLP 2021, pages 1792–1798

large dataset crated using OLID's general annotation model but using semi-supervised learning instead of manually annotated labels. SOLID contains over 1.4 million English tweets with offensive scores greater than 0.5. We show that the proposed fBERT outperforms both a plain BERT implementation and HateBERT on various offensive and hate speech detection tasks.

The contributions of this paper are as follows:

1. An empirical evaluation of transformer-based, semi-supervised learning techniques applied to offensive language identification with the clear potential application to many other text classification tasks.

2. A comprehensive evaluation of several BERT-based strategies and data selection thresholds for offensive language identification across multiple datasets.

3. The release of fBERT, a high-performing, state-of-the-art pre-trained model for offensive language identification.

2 Related Work

The use of large pre-trained transformer models has become widespread in NLP. This includes several recently developed offensive language identification systems based on transformer architectures such as BERT (Devlin et al., 2019). These systems have achieved top performance in popular competitions such as HASOC 2019 (Mandl et al., 2019), HatEval 2019 (Basile et al., 2019), OffensEval 2019 and 2020 (Zampieri et al., 2019b, 2020), and TRAC 2020 (Kumar et al., 2020). The great performance obtained by these systems provides further evidence that pre-trained transformer models are a good fit for the kind of semantic understanding required when identifying offensive content online.

Most of the top systems submitted to the aforementioned competitions (Ranasinghe et al., 2019; Wiedemann et al., 2020; Liu et al., 2019a), however, use models pre-trained on standard contemporary texts. User generated content and offensive language online, however, contain its own set of distinctive features that models trained on standard texts may fail to represent. Therefore, fine-tuning pre-trained models to this challenging domain is a promising but under explored research direction. To the best of our knowledge, a recent first attempt to fine-tune a BERT model to deal with offensive language online, HateBERT (Caselli et al., 2021), shows promising results for English on multiple datasets. In this paper, we address some of the limitations of HateBERT, discussed in the introduction of this paper, and present fBERT, a new BERT-based offensive language model made freely available to the research community.

3 Data

The limited size of existing datasets has been a bottleneck for offensive language identification. OLID (Zampieri et al., 2019a), the dataset used in OffensEval 2019 and arguably the most popular dataset for this task, contains only $14,100$ tweets. OLID is annotated using a hierarchical annotation taxonomy and, as a result, only a sub-set of the corpus is annotated in the lower levels of the taxonomy, i.e., only a few hundred instances.

More recently, following the OLID taxonomy, Rosenthal et al. (2021) released a large-scale offensive language identification dataset (SOLID) with over 9 million English tweets. The data is collected using the Twitter streaming API. The annotations include labels learned using semi-supervised methods. One important difference between SOLID and OLID is that SOLID is collected using random seeds, which has been shown to decrease topic bias compared to the target keywords used in OLID. All the usernames and URLs are replaced with placeholders and tweets with less than two words or 18 characters were discarded.

For retraining fBERT, we have selected over 1.4 million offensive instances from SOLID. We considered multiple offensive score thresholds from the SOLID dataset including all instances with scores between 0.5 and 1.0 arranged in five bins with 0.1 increments. The number of instances in each range is available in Table 1. We did not consider the range $0.9 - 1.0$ given that only a very small number of instances $(2,771)$ were in this bin .

Threshold	Instances
0.5 - 1.0	1,446,580
0.6 - 1.0	1,040,525
0.7 - 1.0	700,719
0.8 - 1.0	348,038
0.9 - 1.0	2,771

Table 1: Offensive instances from the SOLID dataset, organized according to threshold.

4 Model Architecture

4.1 Input Representation

We take the sentence input and tokenize it using WordPiece embeddings (Wu et al., 2016) with a 30,000 token vocabulary as described in (Devlin et al., 2019). The tokenized input is represented as:

$$\mathbf{X} = (\mathbf{x}_{[CLS]}, \mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n, \mathbf{x}_{[SEP]}) \quad (1)$$

where $\mathbf{x}_t$ is the V-dimensional one-hot encoding of the t-th token in a sequence of n symbols (vocabulary of size V). The tokenized input is then processed via $Bert(\mathbf{X})$ to generate contextualized embeddings as follows:

$$\mathbf{H} = Bert(\mathbf{X}) \quad (2)$$

$$\mathbf{H} = (\mathbf{h}_{[CLS]}, \mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_n, \mathbf{h}_{[SEP]}) \quad (3)$$

where $\mathbf{h}_t$ is the d-dimensional embedding for the t-th token $\mathbf{x}_t$ (resulting in n embeddings).

4.2 Retraining Procedure

The goal of the study is to adapt the BERT model for social media aggression detection tasks. We utilized a BERT base (uncased) model that consists of 12 bidirectional transformers encoders with 768 hidden layers and 12 self-attention heads. To use the general understanding of the English language and context, we initialize the transformer with pre-trained weights[1]. We used over 1.4 million offensive texts from the SOLID dataset to retrain the model. No cleaning was applied to preserve the incoherent composition of social media posts, such as the excessive use of mentions, emojis, or hashtags. We retrained the model using the masked language modeling objective to adapt the bidirectional representations of social media offensive language.

Masked Language Modeling (MLM) In MLM, we randomly mask a percentage of tokens and predict the masked inputs. As prescribed in the original BERT implementation, we randomly select 15% of the total tokens for replacement, 80% of the selected tokens are replaced with $[MASK]$, 10% are substituted with a random token chosen from the vocabulary, and 10% remain unchanged. The hidden vectors with masked tokens are fed into a softmax activation function to generate a probability distribution over each (masked) token $\mathbf{x}_t$:

$$p(\mathbf{x}_t|\mathbf{H}) = softmax(\mathbf{W} \cdot \mathbf{h}_t + \mathbf{b}) \quad (4)$$

[1]BERT Pre-trained weights: https://github.com/google-research/bert

where $\cdot$ is matrix multiplication, $\mathbf{W} \in \mathcal{R}^{d \times V}$, and $\mathbf{b} \in \mathcal{R}^{V \times 1}$. The model is trained to predict the original token by minimizing the Catergorical cross-entropy objective as follows:

$$\mathcal{L} = -\sum_{t=1}^{n} m_t \sum_{v} \Big(\mathbf{x}_t \otimes \log(p(\mathbf{x}_t|\mathbf{H}))\Big)[v] \quad (5)$$

where m_t is the binary scalar applied at time step t (1 if the word is masked, 0 otherwise). $[v]$ retrieves/indexes the vth item in the vector $\mathbf{x} \otimes \log(p(\mathbf{x}|\mathbf{H})$ and $\otimes$ indicates element-wise multiplication. A schematic representation of the BERT masked language model is presented in Figure 1.

Retraining Setup We trained the resulting fBERT for 25 epochs using the MLM objective with 0.15 probability to randomly mask tokens in the input. The language model is trained with a batch size of 32 and a 512 maximum token length using the Adam optimizer with a learning rate of $5e-5$. The training time took 5 days on a single Nvidia V100 GPU.

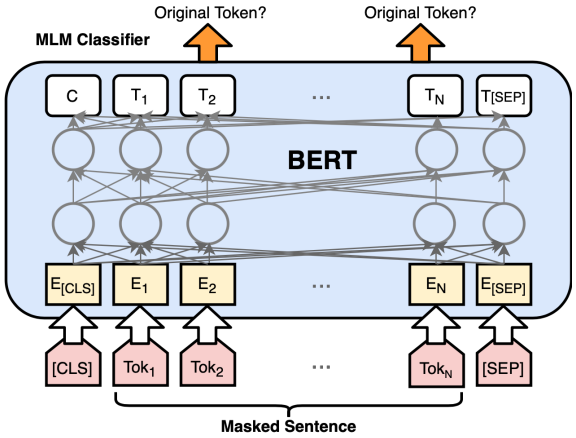

Figure 1: A schematic representation of the fBERT masked language model (the re-trained/tuned BERT).

5 Experiments

To determine the effectiveness and portability of the trained fBERT, we conducted a series of experiments using benchmark datasets and compared our model with a general-purpose BERT model. We used the same set of configurations for all the datasets evaluated in order to ensure consistency between all the experiments. This also provides a good starting configuration for researchers who intend to use fBERT on a new dataset.

We used a batch-size of eight, Adam optimiser with learning rate $1e{-}4$, and a linear learning rate warm-up over 10% of the training data. During the training process, the parameters of the transformer model, as well as the parameters of the subsequent layers, were updated. The models were trained using only training data. Furthermore, they were evaluated while training using an evaluation set that had one fifth of the rows in training data. We performed early stopping if the evaluation loss did not improve over ten evaluation steps. All the models were trained for three epochs. The rest of the parameters are shown in Table 2. These experiments were also conducted in a Nvidia V100 GPU.

Parameter	Value
learning rate	1e-4
adam epsilon	1e-8
warmup ratio	0.1
warmup steps	0
max grad norm	1.0
max seq. length	140
gradient accumulation steps	1

Table 2: Parameter Specifications.

HatEval 2019 In the SemEval 2019, HatEval (Basile et al., 2019) introduced the challenge of detecting multilingual hate speech against women and immigrants. The dataset for the task is collected from Twitter in both English and Spanish. In this work, we used only the English dataset comprised of $9,000$ training instances with $4,177$ hateful tweets. The development (dev) and test sets contain $1,000$ (123 instances are hateful) and $3,000$ examples ($1,380$ instances are hateful) respectively. In terms of pre-processing, we removed extra whitespaces, usernames and URLs were replaced with placeholders, the Emoji[2] package was used to convert the emojis to text, and the Word Segmentation[3] package was used to segment the words into hashtags. We applied the same preprocessing steps for all models to compare the test set macro F_1 score.

OLID We use OLID, the official dataset for OffensEval 2019 (Zampieri et al., 2019b), one of the the most popular offensive language identification shared tasks. The dataset has $13,240$ training and 860 test instances. There are $4,400$ and 240 offensive posts in the training and test dataset, respectively. For the experiment, we chose sub-task A, a binary classification task between offensive and non-offensive posts. We used 10% of the training data as development data and performed preprocessing and cleaning steps as described by Liu et al. (2019a). We trained fBERT for the offensive language detection task and compared its performance with other language models using the macro F_1 score.

Hate Speech and Offensive Language Detection (HS & O) In fine-grain aggression detection, classifying offensive language and hate speech is challenging. Hate speech contains explicit instances targeted towards a specific group of people intended to degrade or insult. Davidson et al. (2017) compiled a $24,783$ English tweets dataset annotated with one of three labels – "hate speech", "only offensive", and "neither". The dataset contains $1,430$ hate speech, $19,190$ only offensive, and $4,163$ instances that are neither. We further split the dataset into training, dev, and test sets in a 3:1:1 ratio. We applied the same preprocessing steps we applied to the HatEval 2019 dataset.

6 Results

We first present the results for the SOLID data selection thresholds in Table 3 in terms of F_1 Macro. For the three datasets tested, the 0.5 - 1.0 threshold, which includes the largest number of instances, yielded the best performance.

	Datasets		
Scores	HatEval	OLID	HS & O
0.5 - 1.0	**0.596**	**0.813**	**0.878**
0.6 - 1.0	0.562	0.808	0.871
0.7 - 1.0	0.550	0.802	0.867
0.8 - 1.0	0.554	0.801	0.865

Table 3: Macro F_1 scores for different SOLID threshold score values.

We then compare the performance of fBERT with BERT and HateBERT. In the HatEval Sub-task A, we see that fBERT has outperformed BERT by increasing the test macro F_1 score by over 23%. This empirically demonstrates the advantage and generalization power of our domain-specific retrained BERT model. The best model (Indurthi et al., 2019) in this task used an SVM model with a radial basis kernel, exploiting sentence embeddings from

[2]Emoji Package: https://pypi.org/project/emoji/

[3]Word Segmentation Package: https://pypi.org/project/wordsegmentation/

Google's Universal Sentence Encoder as features. The results are shown in Table 4.

The fBERT model also performs better than the generic BERT and abusive language HateBERT in OffensEval Sub-task A, achieving a test set macro F_1 score of 0.8132. We observe that the fBERT is also highly effective in fine-grain offensive and hate speech detection, obtaining a 10% increase in the F_1 score.

Dataset	Model	Macro F1
HatEval	**fBERT**	**0.596**
	HateBERT	0.525
	BERT	0.483
OLID	**fBERT**	**0.813**
	HateBERT	0.801
	BERT	0.794
HS & O	**fBERT**	**0.878**
	HateBERT	0.846
	BERT	0.806

Table 4: The test set macro F_1 scores for all datasets and models. Results are ordered by performance. Best results are shown in bold font.

Finally, as observed in the experimental results presented above, we observe that fBERT has outperformed the abusive language HateBERT model in all of the experiments. The proposed fBERT has also performed efficiently in all the aggression detection tasks. This validates the effectiveness of the proposed domain-specific transformer model for offensive and hateful language classification tasks. The proposed fBERT model is effective across different datasets and objectives, providing a powerful model to be used for hateful/offensive content identification.

7 Conclusion

Over the years, neural transformer models have outperformed previous state-of-the-art deep learning models across various NLP tasks including offensive and hate speech detection tasks. Nevertheless, these transformers are usually trained on general corpora which lack tweet and offensive language-specific cues. Previous studies have shown that domain-specific fine-tuning or retraining of models before attempting to solve downstream tasks can lead to excellent results in multiple domains. As discussed in this paper, fine-tuning/retraining a complex models to identify offensive language has not been substantially explored before and we address this gap by proposing fBERT, a *bert-base-uncased* model that has been learned using over 1.4 million offensive instances from the SOLID dataset. The shifted fBERT model better incorporates domain-specific offensive language and social media features. The fBERT model achieves better results in both OffensEval and HatEval tasks and in the HS & O dataset over BERT and HateBERT.

In future work, we would like to investigate the performance of fBERT both at the post- and token-level identification stages. Furthermore, we will expand fBERT to multiple languages. Since our approach is based on a semi-supervised dataset, it is easily expandable to other languages as well. We plan to extend this process to other transformer models such as XLNET (Yang et al., 2019), RoBERTa (Liu et al., 2019b) and ALBERT (Lan et al., 2020). Finally, fBERT is publicly available on Hugging Face model hub (Wolf et al., 2020).[4]

Ethics Statement

fBERT is essentially a BERT model for offensive language identification which is trained on multiple publicly available datasets. We used multiple datasets referenced in this paper which were previously collected and annotated. No new data collection has been carried out as part of this work. We have not collected or processed writers'/users' information nor have we carried out any form of user profiling protecting users' privacy and identity.

We understand that every dataset is subject to intrinsic bias and that computational models will inevitably learn biased information from any dataset. We believe that fBERT will help coping with biases in datasets and models as it features a freely available BERT model that other researchers can use to train new offensive language identification models on other datasets.

Acknowledgments

We would like to thank the HatEval and OffensEval shared task organizers for making the datasets used in this paper available. We further thank the anonymous EMNLP reviewers for their insightful feedback.

This research was partially supported by a seed fund award sponsored by RIT's Global Cybersecurity Institute (GCI).

[4]fBERT at HuggingFace: https://huggingface.co/diptanu/fBERT

References

Ibrahim Abu Farha and Walid Magdy. 2020. Multitask learning for Arabic offensive language and hatespeech detection. In *Proceedings of OSCAT*.

Dogu Araci. 2019. Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *Proceedings of SemEval*.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of EMNLP*.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2021. Hatebert: Retraining bert for abusive language detection in english. In *Proceedings of WOAH*.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The Muppets straight out of Law School. In *Findings of the ACL*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. In *Proceedings of ACL*.

Thomas Davidson, Dana Warmsley, Michael W. Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of ICWSM*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL*.

Vijayasaradhi Indurthi, Bakhtiyar Syed, Manish Shrivastava, Nikhil Chakravartula, Manish Gupta, and Vasudeva Varma. 2019. FERMI at SemEval-2019 task 5: Using sentence embeddings to identify hate speech against immigrants and women in Twitter. In *Proceedings of SemEval*.

Ritesh Kumar, Atul Kr Ojha, Shervin Malmasi, and Marcos Zampieri. 2018. Benchmarking Aggression Identification in Social Media. In *Proceedings of TRAC*.

Ritesh Kumar, Atul Kr. Ojha, Shervin Malmasi, and Marcos Zampieri. 2020. Evaluating Aggression Identification in Social Media. In *Proceedings of TRAC*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations. In *Proceedings of ICLR*.

Ping Liu, Wen Li, and Liang Zou. 2019a. NULI at SemEval-2019 task 6: Transfer learning for offensive language detection using bidirectional transformers. In *Proceedings of SemEval*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*.

Shervin Malmasi and Marcos Zampieri. 2017. Detecting Hate Speech in Social Media. In *Proceedings of RANLP*.

Thomas Mandl, Sandip Modha, Prasenjit Majumder, Daksh Patel, Mohana Dave, Chintak Mandlia, and Aditya Patel. 2019. Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages. In *Proceedings of FIRE*.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English tweets. In *Proceedings of EMNLP*.

Tharindu Ranasinghe and Marcos Zampieri. 2020. Multilingual Offensive Language Identification with Cross-lingual Embeddings. In *Proceedings of EMNLP*.

Tharindu Ranasinghe and Marcos Zampieri. 2021. MUDES: Multilingual Detection of Offensive Spans. In *Proceedings of NAACL*.

Tharindu Ranasinghe, Marcos Zampieri, and Hansi Hettiarachchi. 2019. Brums at hasoc 2019: Deep learning models for multilingual hate speech and offensive language identification. In *Proceedings of FIRE*.

Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Marcos Zampieri, and Preslav Nakov. 2021. SOLID: A Large-Scale Weakly Supervised Dataset for Offensive Language Identification. In *Findings of the ACL*.

Gregor Wiedemann, Seid Muhie Yimam, and Chris Biemann. 2020. UHH-LT at SemEval-2020 task 12: Fine-tuning of pre-trained transformer networks for offensive language detection. In *Proceedings of SemEval*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame,

Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of EMNLP*.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding. In *Proceedings of NeurIPS*.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019a. Predicting the type and target of offensive posts in social media. In *Proceedings of NAACL*.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019b. SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval). In *Proceedings of SemEval*.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. SemEval-2020 Task 12: Multilingual Offensive Language Identification in Social Media (OffensEval 2020). In *Proceedings of SemEval*.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of ICCV*.