Jigsaw @ AMI and HaSpeeDe2: Fine-Tuning a Pre-Trained Comment-Domain BERT Model

Alyssa Lees and **Jeffrey Sorensen** and **Ian Kivlichan**
Google Jigsaw
New York, NY
(alyssalees|sorenj|kivlichan)@google.com

Abstract

The Google Jigsaw team produced submissions for two of the EVALITA 2020 (Basile et al., 2020) shared tasks, based in part on the technology that powers the publicly available PerspectiveAPI comment evaluation service. We present a basic description of our submitted results and a review of the types of errors that our system made in these shared tasks.

1 Introduction

The HaSpeeDe2 shared task consists of Italian social media posts that have been labeled for hate speech and stereotypes. As Jigsaw's participation was limited to the A and B tasks, we will be limiting our analysis to that portion. The full details of the dataset are available in the task guidelines (Bosco et al., 2020).

The AMI task includes both raw (natural Twitter) and synthetic (template-generated) datasets. The raw data consists of Italian tweets manually labelled and balanced according to misogyny and aggressiveness labels, while the synthetic data is labelled only for misogyny and is intended to measure the presence of unintended bias (Elisabetta Fersini, 2020).

2 Background

Jigsaw, a team within Google, develops the PerspectiveAPI machine learning comment scoring system, which is used by numerous social media companies and publishers. Our system is based on distillation and uses a convolutional neural-network to score individual comments according to several attributes using supervised training data labeled by crowd workers. Note that PerspectiveAPI actually hosts a number of different models that each score different attributes. The underlying technology and performance of these models has evolved over time.

While Jigsaw has hosted three separate Kaggle competitions relevant to this competition (Jigsaw, 2018; Jigsaw, 2019; Jigsaw, 2020) we have not traditionally participated in academic evaluations.

3 Related Work

The models we build are based on the popular BERT architecture (Devlin et al., 2019) with different pre-training and fine-tuning approaches.

In part, our submissions explore the importance of pre-training (Gururangan et al., 2020) in the context of toxicity and the various competition attributes. A core question is to what extent these domains overlap. Jigsaw's customized models (used for the second HaSpeeDe2 submission, and both AMI submissions) are pretrained on a set of one billion user-generated comments: this imparts statistical information to the model about comments and conversations online. This model is further fine-tuned on various toxicity attributes (toxicity, severe toxicity, profanity, insults, identity attacks, and threats), but it is unclear how well these should align with the competition attributes. The descriptions of these attributes and how they were collected from crowd workers can be found in the data descriptions for the Jigsaw Unintended Bias in Toxicity Classification (Jigsaw, 2019) website.

A second question studied in prior work is to what extent training generalizes across languages (Pires et al., 2019; Wu and Dredze, 2019; Pamungkas et al., 2020). The majority of our training data is English comment data from a variety of sources, while this competition is based on Italian Twitter data. Though multilingual transfer has been studied in general contexts, less is known about the specific cases of toxicity, hate speech,

misogyny, and harassment. This was one of the focuses of Jigsaw's recent Kaggle competition (Jigsaw, 2020); i.e., what forms of toxicity are shared across languages (and hence can be learned by multilingual models) and what forms are different.

4 Submission Details

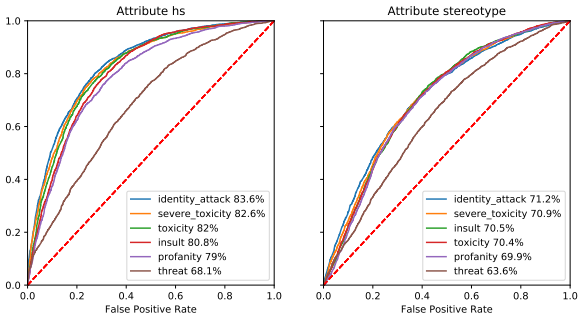

Figure 1: ROC curves for the PerspectiveAPI multilingual teacher model attributes compared to the HaSpeeDe2 attributes (hate speech and stereotype).

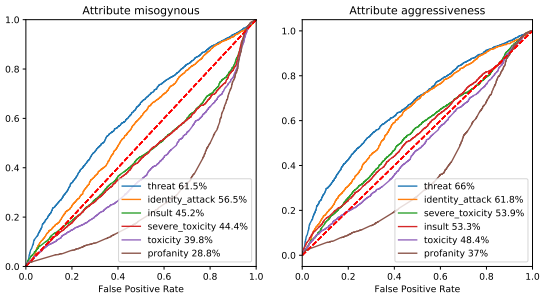

Figure 2: ROC curves for PerspectiveAPI multilingual teacher model attributes compared to the AMI attributes (misogyny and aggressiveness).

As Jigsaw has already developed toxicity models for the Italian language, we initially hoped that these would provide a preliminary baseline for the competition despite the independent nature of the development of the annotation guidelines. Our Italian models score comments for toxicity as well as five additional distinct toxicity attributes: severe toxicity, profanity, threats, insults, and identity attacks. We might expect some of these attributes to correlate with the HaSpeeDe2 and AMI attributes, though it is not immediately clear whether any of these correlations should be particularly strong.

The current Jigsaw PerspectiveAPI models are typically trained via distillation from a multilingual teacher model (that is too large to practically serve in production) to a smaller CNN. Using this large teacher model, we initially compared the EVALITA hate speech and stereotype annotations against the teacher model's scores for different attributes. The results are shown in Figure 1 for the training data. Perspective is a reasonable detector for the hate speech attribute, but performs less well for the stereotype attribute, with the identity attack model performing the best.

Using these same models on the AMI task, shown in Figure 2 for detecting misogyny proved even more challenging. In this case, the aggressiveness attribute was evaluated only on the subset of the training data labeled misogynous. In this case, the most popular attribute of "toxicity" is actually counter-indicative of the misogyny label. The best detector for both of these attributes appears to be the "threat" model.

As can be seen, the existing classifiers are all poor predictors of both attributes for this shared task. Due to errors in our initial analysis, we did not end up using any of the models used for PerspectiveAPI in our final submissions.

Category	Submission	hate speech	stereotype
news	1	0.68	0.64
	2	0.64	0.68
tweets	1	0.72	0.67
	2	0.77	0.74

Table 1: Macro-averaged F1 scores for Jigsaw's HaSpeeDe2 Submissions.

4.1 HaSpeeDe2

The Jigsaw team submitted two separate submissions that were independently trained for Tasks A and B.

4.1.1 First Submission

Our first submission, one that did not perform very well, was based on a simple multilingual BERT model fine-tuned on 10 random splits of the training data. For each split, 10% of the data was held out to choose an appropriate equal-error-rate threshold for the resulting model.

The BERT fine-tuning system used the 12 layer model (Tensorflow Hub, 2020), a batch size of 64 and sequence length of 128. A single dense layer is used to connect to the two output sigmoids which are trained using a binary cross-entropy loss

using stochastic gradient descent with early stopping, which is computed using the AUC metric computed using the 10% held out slice. This model is implemented using Keras (Chollet and others, 2015).

To create the final submission, the decisions of the ten separate classifiers were combined in a majority voting scheme (if 5 or more models produced a positive detection, the attribute was assigned true).

4.1.2 Second Submission

Our second submission was based on a similar approach of fine-tuning a BERT-based model, but one based on a more closely matched training set.

The underlying technology we used is the same as the Google Cloud AutoML for natural language processing product that had been employed in similar labeling applications (Bisong, 2019).

The remaining models built for this competition and in the subsequent section are based on a customized BERT 768-dimension 12-layer model pretrained on 1B user-generated comments using MLM for 125 steps. This model was then fine-tuned on supervised comments in multiple languages for six attributes: toxicity, severe toxicity, obscene, threat, insult, and identity hate. This model also uses a custom wordpiece model (Wu et al., 2016) comprised of 200K tokens representing tokens from hundreds of languages.

Our hate speech and misogyny models use a fully connected final layer that combines the six output attributes and allows weight propagation through all layers of the network. Fine-tuning continues on using the supervised training data provided by the competition hosts using the ADAM optimizer with a learning rate of 1e–5.

Figure 3 displays the ROC curve for our second submission for each of the news and the tweets datasets as well as for both the hate speech and stereotype attributes.

Our second submission for HaSpeeDe2 consisted of fine-tuning a single model with the provided training data with a 10% held-out set. The custom BERT model was fine-tuned on TPUs using a relatively small batch size of 32.

4.2 AMI

Our submissions for the AMI task only considered the unconstrained case, due to the use of pretrained models. All AMI models were fine-tuned on TPUs using the customized BERT checkpoint and custom wordpiece vocabulary from Section 4.1.2. However, a larger batch-size of 128 was specified. All models were fine-tuned simultaneously on misogynous and aggressive labels using the provided data, where zero aggressiveness weights were assigned to data points with no misogynous labels.

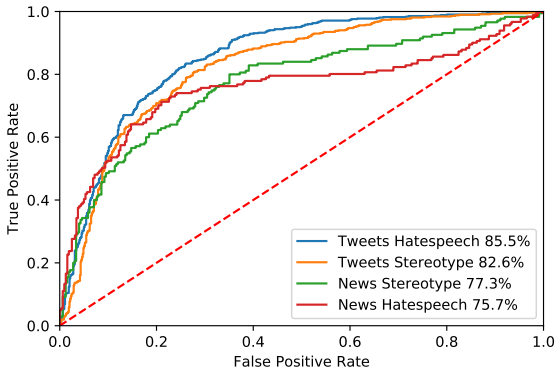

Figure 3: ROC plots for HaSpeeDe2 Test Set Labels.

Both submissions were based on ensembles of partitioned models evaluated on a 10% held-out test set. We explored two different ensembling techniques, which we discuss in the next section.

AMI submission 1 does not not include synthetic data. AMI submission 2 includes the synthetic data and custom biasing mitigation data selected from Wikipedia articles. Table 2 clearly shows that the inclusion of such data significantly improved the performance on Task B for submission 2. Interestingly, the inclusion of synthetic and bias mitigation data slightly improved the performance in Task A as well.

Task	Submission	Score
A	1	0.738
	2	0.741
B	1	0.649
	2	0.883

Table 2: Misogynous and Aggressiveness Macro-averaged F1 scores for Jigsaw's AMI Submissions.

The two Jigsaw models ranked in first and second place for Task A. The second submission ranked first among participants for Task B.

4.2.1 Ensembling Models

Both the first and second submissions for AMI were ensembles of fine-tuned custom BERT models constructed from partitioned training data. We explored two ensembling techniques (Brownlee, 2020):

- Majority Vote: Each partitioned model was evaluated using a model specific threshold. The label for each attribute was determined by majority vote among the models.
- Average: The raw models probabilities are averaged together. The combined model calculates the labels via custom thresholds determined by evaluation on a held-out set.

Thresholds for the individual models in the majority vote and average ensemble were calculated to optimize for the point on the held-out data ROC curve where $|\text{TPR} - (1 - \text{FPR})|$ is minimized.

The majority voting model performed slightly better for both the misogynous and aggressive task on the held-out sets. As such, both submissions use majority vote.

4.2.2 First Submission

Using the same configuration as Section 4.1.2, we partitioned the raw training data into ten randomly chosen partitions and fine-tuned nine of these using the 10% held out portion to compute thresholds. No synthetic or de-biasing data was included in this submission.

We include ROC curves for half of these models in Figure 4, to illustrate that they are similar but with some variance when used to score the test data.

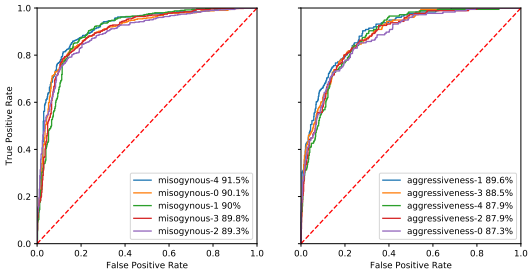

Figure 4: ROC plots for AMI test set labels for models pre-ensemble.

Our first unconstrained submission using majority vote for AMI achieved scores of 0.738 for Task A and 0.649 for Task B. The poorer score for Task B is not surprising given that no bias mitigating data or constraints were included in training.

4.2.3 Second Submission

In order to mitigate bias, we decided to augment the training data set using sentences sampled from the Italian Wikipedia articles that contain the 17 terms listed in the identity terms file provided with the test set data. These sentences were labeled as both non-misogynous and non-aggressive. 11K sentences were used for this purpose, with the term frequencies summarized in Table 3.

Identity Term	Sentence Count
donna	4306
donne	3100
femmine	1275
femmina	652
fidanzata	538
nonna	378
mamma	269
casalinga	256
casalinghe	187
compagne	132
compagna	34
mamme	24
fidanzate	12
nonne	11
matrone	9
matrona	8
morosa	6

Table 3: Term frequency in Wikipedia sampled sentences for bias mitigation.

The second submission employed the same partitioning of data with a held-out set. However the unconstrained data included the raw training data, the provided synthetic data and our de-biasing term data. As with submission 1, majority vote was used with custom thresholds determined by evaluation on the held-out set.

Our first unconstrained submission for AMI achieved scores of 0.741 for Task A and 0.883 for Task B.

5 Error Analysis

We discuss an informal analysis of the errors we observed with each of these tasks. Aside from the typical questions regarding data annotation quality, and the small sample sizes, we observed some particular instances of avoidable errors.

5.1 HaSpeeDe2 Errors

Looking at the largest incongruities as shown in Table 4 it is clear that context, which is unavailable to our models, and presumably to the mod-

ID	Comment	HS	Score
11355	@user @user @user Giustissimo, non bisogna mai nascondersi nelle ideologie, sopratutto oggi perché non esistono più. Sta di fatto, che le cose più aberranti che leggi oggi sui giornali hanno sempre@a@che fare con stranieri... o rom URL	1	.00001
10803	#Simone di #CasalBruciato, #Roma: "Certi rom sono cittadini italiani, ma non sono uguali a noi. Uguali non è il termine più giusto da usare". URL	1	.00003
11288	I SOLDI DEI DISABILI AI MIGRANTI La regione Emilia Romagna destina la metà dei fondi destinati alle fasce deboli a progetti per i richiedenti asilo A Reggio Emilia il 69% delle risorse stanziate sono state utilizzate ai richiedenti asilo #PRIMAGLIITALIANI URL	1	.00003
10420	#MeNeFottoDi questi sfigati #facciamorete che continuano a giustificare ogni crimine commesso da immigrati... non fate rete, FATE SCHIFO... #facciamociFURBI	0	0.99996
11189	@user Naturalmente in questo caso alla faccia dei comunisti e dei migranti stitici!	0	0.99996
10483	@user SCHIFOSA IPOCRITA SPONSORIZZI I MUSSULMANI E POI VOI DARE I DIRITTI ALLE DONNE SI VEDE CHE SEI POSSEDUTA DAL DIAVOLO SEI BUGIARDA BOLDRINA SAI SOLO PROTESTARE POI TI CHIEDI PERCHÉ IL VERO ITALIANO TI ODIA PERCHÉ SEI UNA SPORCA IPOCRITA	0	0.99995

Table 4: Largest Errors for hate speech classifier on HaSpeeDe2 Tweet data

erators, is important for determining the author's intent. The use of humor and the practice of quoting text from another author are also confounding factors. As this task is known to be hard (Vigna et al., 2017; van Aken et al., 2018), the edge cases display these confounding reasons. Additionally, as evidenced by the last comment, the subtlety of hate speech that is directed toward the designated target for this challenge has not been well captured.

The BERT model that we fine-tuned for this application is cased, and we see within our errors frequent use of all-caps text. However, lower casing the text has almost no effect on the scores, suggesting that the BERT pre-training has already linked the various cased versions of the tokens in the vocabulary.

We analyzed the frequency of word piece fragments in the data and saw no correlation between misclassification and the presence of segmented words. This suggests that vocabulary coverage in the test set does not play a significant role in explaining our systems' errors.

Considering the sentence with the high model score for hate speech, several single terms are tagged by the model. For example the term "sfigati" occurs only once in the training data in a sentence that is marked as non-hate speech. However, this term is not in our vocabulary and gets split into pieces "sfiga##ti", and the prefix "sfiga" appears in two out of three training examples that are marked hate speech—exactly the kind of data sparsity that leads to unwanted bias. Using a larger amount of training data, even if it creates an imbalance, is one way to address this, as we did in the case of the AMI challenge.

5.2 AMI

Because we are using ensemble models trained on partitions of the training set, we observe that the highest-scoring test samples that are marked non-misogynous and non-aggressive, as well as the lowest-scoring misogynous and aggressive comments, vary from model to model. However, we display the most frequently occurring mistakes across all ten ensembles in Table 5.

Regarding the false alarms, these comments appear to be mislabeled test instances, and there is ample support for this claim in the training data. The first comment combines both uppercase and a missing space. While it's true that subjunctive mode is not well represented in the training data, lower casing this sentence produces high scores. This is also the case with the third example. The second error seems more subtle, perhaps an attempt at humor, but one with no salient misogyny terms.

6 Bias

Because the identity terms for AMI are not observed with a high frequency in the test data, we restrict our analysis to the synthetic data set. We find wide variation in the performance of our individual models, with one model exhibiting very poor performance across the subgroups. The summary of the AUC measurements for these models

ID	Comment	Error Type
5466	Odio Sakura per il semplice motivo che qualunque cosa faccia o dica Naruto lei lo prende a schiaffi o a pugniHA CHIESTO COME STA SAI DIOSANTO BRUTTA STRONZA MA CONTRALLI MADONNA SPERO CHE TI UCCIDANOscusami Sarada	Missed Detection
5385	porca troia la prenderei a schiaffi se non fosse mia madre	Missed Detection
5819	ma in tutto ciò lo scopo di anna qual è? far soffrire il mio protetto? IO TI AMMAZZO COI LANCIAFIAMME OH #TemptationIsland	Missed Detection
5471	@danielita8811 Che bel culo tutto da sfondare	False Alarm
5604	@coppiacalda2 Che bel culo da inculare	False Alarm

Table 5: Persistent errors for AMI across different ensembles.

are shown in Figure 5, Figure 6, and Figure 7 using the technique presented in (Borkan et al., 2019). There does not appear to be a systemic problem with bias in these models, but judging based only upon synthetic data is probably unwise. The single term "donna" from the test set shows a subgroup AUC that drops substantially from the background AUC for nearly all of the models, perhaps indicating limitations of judging based on synthetic data.

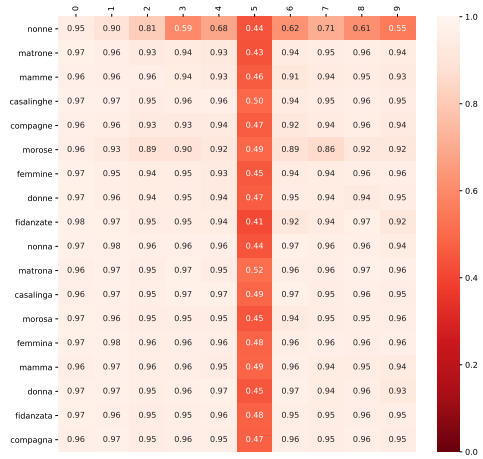

Figure 5: Subgroup AUC

7 Conclusions and Future Work

Both of these challenges dealt with issues related to content moderation and evaluation of user-generated content. While early research raised fears of censorship, the ongoing challenges platforms face have made it necessary to consider the potential of machine learning. Advances in natural language understanding have produced models that work surprisingly well, even ones that are able to detect malicious intent that users try to encode in subtle ways.

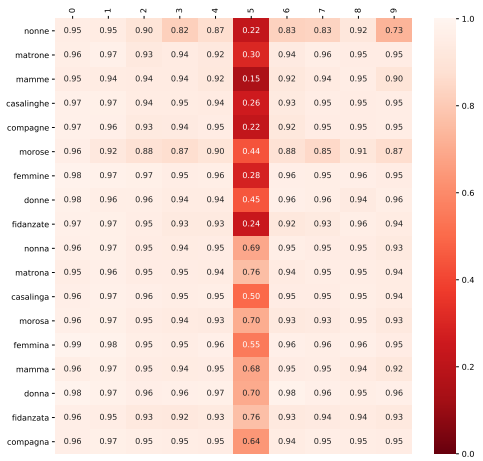

Figure 6: Background Positive, Subgroup Negative AUC

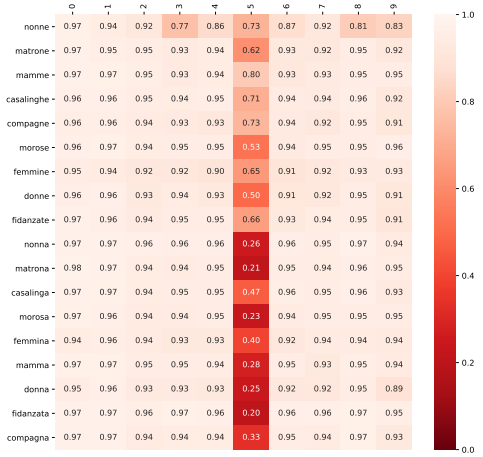

Figure 7: Background Negative, Subgroup Positive AUC

Our particular approach to the EVALITA challenges represented an unsurprising application of what has now become a textbook technique: leveraging the resources of large pre-trained models. However, many participants achieved nearly similar performance levels in the constrained task. We regard this as a more impressive accomplishment.

Jigsaw continues to apply machine learning to support publishers and to help them host quality online conversations where readers feel safe participating. The kinds of comments these challenges tagged are some of the most concerning and pernicious online behaviors, far outside of the norms that are tolerated in other public spaces. But humans and machines both still misinterpret profanity for hostility, and tagging humor, quotations, sarcasm, and other legitimate expressions for moderation remain serious problems.

Challenges like the AMI and HasSpeede2 competitions underscore the importance of understanding the relationships between the parties in a conversation, and the participants' intents. We are greatly encouraged that attributes that our systems do not currently capture were somewhat within the reach of our present techniques—but clearly much work remains to be done.

References

Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro. 2020. Evalita 2020: Overview of the 7th evaluation campaign of natural language processing and speech tools for italian. In Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro, editors, *Proceedings of Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020)*, Online. CEUR.org.

Ekaba Bisong, 2019. *Google AutoML: Cloud Natural Language Processing*, pages 599–612. Apress, Berkeley, CA.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion Proceedings of The 2019 World Wide Web Conference*, pages 491–500.

Cristina Bosco, Tommaso Caselli, Gloria Comandini, Elisa Di Nuovo, Simona Frenda, Viviana Patti, Irene Russo, Manuela Sanguinetti, and Marco Stranisci. 2020. Hate speech detection task second edition (haspeede2) at evalita 2020 task guidelines. `https://github.com/msang/haspeede/blob/master/2020/HaSpeeDe2020_Task_guidelines.pdf`.

Jason Brownlee. 2020. How to develop voting ensembles with python. `https://machinelearningmastery.com/voting-ensembles-with-python/`, September.

Francois Chollet et al. 2015. Keras.

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, June. Association for Computational Linguistics.

Paolo Rosso Elisabetta Fersini, Debora Nozza. 2020. Ami @ evalita2020: Automatic misogyny identification. In Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro, editors, *Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020)*, Online. CEUR.org.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online, July. Association for Computational Linguistics.

Jigsaw. 2018. Jigsaw toxic comment classification challenge. `https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge`, March.

Jigsaw. 2019. Jigsaw unintended bias in toxicity classification. `https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification`, July.

Jigsaw. 2020. Jigsaw multilingual toxic comment classification. `https://www.kaggle.com/c/jigsaw-multilingual-toxic-comment-classification`, July.

Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. 2020. Misogyny detection in twitter: a multilingual and cross-domain study. *Information Processing & Management*, 57(6):102360.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy, July. Association for Computational Linguistics.

Tensorflow Hub. 2020. Mulilingual L12 H768 A12 V2. `https://tfhub.dev/tensorflow/bert_multi_cased_L-12_H-768_A-12/2`, August.

Betty van Aken, Julian Risch, Ralf Krestel, and Alexander Löser. 2018. Challenges for toxic comment classification: An in-depth error analysis. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 33–42, Brussels, Belgium, October. Association for Computational Linguistics.

F. D. Vigna, A. Cimino, Felice Dell'Orletta, M. Petrocchi, and M. Tesconi. 2017. Hate me, hate me not: Hate speech detection on facebook. In *ITASEC*.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. 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 833–844, Hong Kong, China, November. Association for Computational Linguistics.

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. *CoRR*, abs/1609.08144.