Training Language Models under Resource Constraints for Adversarial Advertisement Detection

Eshwar Shamanna Girishekar **Shiv Surya** **Nishant Nikhil*** **Dyut Kumar Sil**
Sumit Negi **Aruna Rajan**
Amazon
{geshwar, shisurya, dyut, suminegi, rajarna}@amazon.com *i.nishantnikhil@gmail.com

Abstract

Advertising on e-commerce and social media sites deliver ad impressions at web scale on a daily basis driving value to both shoppers and advertisers. This scale necessitates programmatic ways of detecting unsuitable content in ads to safeguard customer experience and trust. This paper focusses on techniques for training text classification models under resource constraints, built as part of automated solutions for advertising content moderation. We show how weak supervision, curriculum learning and multi-lingual training can be applied effectively to fine-tune BERT and its variants for text classification tasks in conjunction with different data augmentation strategies. Our extensive experiments on multiple languages show that these techniques detect adversarial ad categories with a substantial gain in precision at high recall threshold over the baseline.

1 Introduction

All advertisements on e-commerce and social media platforms must be moderated to ensure regulatory and ethical standards in countries where they are being served. A tiered moderation workflow with automated components like cached lookup, ML models, rule based annotators complement human experts to ensure reliable content moderation for ads created by advertisers while scaling to e-commerce advertising volumes. The advertising platform currently enables ads to be created in various media formats like text, images and videos. In this work, we focus on detecting adversarial ads in one broad class of ads, where engagement is driven primarily through text and images. Such ads on e-commerce site serve as a casing for the product being advertised. The casing includes product text and image attributes along with optional custom captions provided by the advertiser. It is under the purview of moderation to check whether an ad contains prohibited content. Any ad containing prohibited content can have an adverse impact on the shopper experience and hence needs to be prevented from showing up. See Section 2.1 for a broad overview of the adversarial ad categories.

In this paper, we focus on techniques we use to train NLP models built as a part of this system. Training any ML model requires a good quality dataset that is representative of the policy being enforced. The quality of data available to train models targeting a defect, say detection of "adult and objectionable content" depends on several factors. Typically occurrences of such products are rare but the impact of such an ad on shopper experience is adverse. The uncommonness of these violations makes curating large in-domain monolingual corpora difficult. This problem is compounded in low resource languages where there are limited linguistic resources and the rarity of these violations are even more skewed. Further, it is expensive and time consuming to gather more labeled data.

Through this paper, we show different ways to train generalised language models when we have limited labeled data. We suggest various ways for data augmentation and empirically provide evidence suggesting when each of the approaches works best. We explore how we can leverage the product catalogue and user behaviour in weak and semi-weak supervision, curriculum learning and multilingual training strategies to train generalised language models like BERT (Devlin et al., 2019) and its variants. Our experiments show :

- Weak supervision for unlabelled data in the target domain provides an average gain of 10.88% in precision across languages.
- Curriculum strategies to augment labeled data from resource rich language by translation improves average true negative rate(TNR) by 24.25% in low resource setting.
- Multilingual training using labeled data in any

*Work done when at Amazon

available languages provides average gain of 24.32% in TNR over the baselines.

2 Background: Content moderation

2.1 Scope of content moderation

Online advertising platforms typically enable advertisers to create ads in various media formats like text, images and videos. Here we provide an overview of the broad categories which are generally restricted from advertising across these platforms.

Sculley et al. (2011) describe some of the adversarial categories which can compromise the user safety. These include ads which promote unsafe and illegal content or products. In addition to these categories, promotion of adult, profane, hate inciting and tobacco related products/content are restricted as well. All of these adversarial categories are under the purview for content moderation.

We primarily featurise the text attributes of the product in catalogue such as product title, description and optional custom text provided by the advertiser to detect aforementioned unsuitable content.

2.2 Dataset

A very small fraction of ads belong to the restricted categories referenced in Section 2.1. We perform all experiments on 5 such semantic categories shown in Table 1. For the positive class(defective ad), we consider all ads labelled by human experts. We split this data into train and validation set using multi-label stratification (Sechidis et al. (2011); Szymański and Kajdanowicz (2017)) on catalogue categorisation of the product. To enable training, we restrict the size of negative class by restricting the sample size to utmost 100 times the size of the positive class and augment it with 10% of hard negative samples that were caught by existing signals but approved by human experts. The validation set is used to tune model hyperparameters and determine the stopping criterion. We maintain a separate temporally distinct test set replicating production setting. A similar approach is taken when creating train and test set for low resource languages.

2.3 Baselines

BERT and M-BERT For all the experiments we make use of BERT (Bidirectional Encoder Representations from Transformers)(Devlin et al., 2019), a transformer based attention model that encodes an entire sequence at once using multiple attention based encoder layers. We use a linear classification layer applied on max-pooled version of last four attention layer outputs of BERT and finetune the model on limited labeled data. Because of the skew in the labels, we weight the binary cross entropy loss inversely based on label frequency and clip the scaling factor to improve stability of training. The model is trained using textual attributes of the products. Adam (Kingma and Ba, 2014) optimiser is used and the maximum sequence length is restricted to 512 during training and inference. For low resource languages we make use of M-BERT. We decide the hyper-parameters of the models by their performance on the validation set and maintain these hyper parameters across ablative experiments.

Word embedding based text classifier In the multi-lingual setting, we use another baseline. This is a linear classifier based on word embeddings similar to the setup in (Shen et al., 2018). We use fastext (Bojanowski et al., 2017) embeddings for German to get the word embeddings and combine them by taking a weighted average of the embeddings as described in Arora et al. (2017). This removes the special direction to generate the sentence embedding. We also obtain max-pooled embeddings that extracts salient features along vector dimensions. This is later stacked to the reference weighted average embedding and used to train a logistic regression classifier with the limited labeled data. We refer to this model as BOE_LIN.

3 Finetuning BERT under low resource constraints

We explore various techniques that can be used to train generalised language models(GLM) like BERT and multilingual variants with significant performance gains over baseline models described in Section 2.3. We look at resource constraints during training of machine learning models in a supervised setting attributed to the following cases:

- Lack of labeled data.
- Lack of large in-domain monolingual corpora.
- Linguistic resources insufficient for building reliable statistical NLP applications.

We leverage product catalog to source data for weak and semi-weak supervision training in monolingual setting. We also explore how curriculum strategies and multilingual training can benefit training text classifiers for low resource languages.

Our experiment show that generalised language models like BERT or multilingual variants like M-BERT can be trained using these techniques with significant performance gains over baseline model described in Section 2.3.

3.1 Semi-Supervision and Semi-Weak-Supervision

We employ two approaches as described in Yalniz et al. (2019) One is the conventional semi-supervised approach using teacher-student paradigm. The teacher model is trained using the limited labelled data (or strong data) and then used to get predictions for the unlabelled data. Top k% of the predicted samples for each of the class are used to pre-train the new student model. The student model is further fine-tuned using the limited labelled data. The second approach is semi-weakly-supervised approach. Here, the sourced data associated with weak labels is used to pre-train the teacher model before fine-tuning on the limited labelled data. Again top k% predicted samples by the this teacher model is used to pre-train student network prior to fine-tuning on the strong data. Yalniz et al. (2019) apply these two techniques for image and video classification tasks and achieve SOTA results using semi-weak-supervision. We explore these approaches applied to text classification task using a GLM like BERT.

3.1.1 Semi-Supervised(SS) Methodology

In this section we describe how we augment unlabelled/weakly labeled data. We leverage user behavioural data by using internal search engine to source products relevant to different categories from huge product catalog. We can query search using generic text phrases and pre-existing catalogue categorisation (CC). So we design relevant text phrases and pre-existing catalogue categorisation for a defect of interest. These attributes are filtered by a keyword list which is a combination of a curated list and word list sourced from models that use BoW as a feature. Table 1 provides the statistics of the proportion of number of products sourced using different approaches.

We use the augmentation for only defective class since the class skew is several orders larger. Once we have the augmented data for the defective category we treat it as unlabelled for semi-supervised setup. The teacher model which is BERT is trained only on the strong data. In case of very limited data like CAT4–5 we make use of fasttext classifier as teacher. The teacher model is used to score the augmented samples. Top k% of the augmented data based on model scores are picked to pre-train the new student BERT model. Later the student BERT model is fine-tuned using the strong labelled data. When training both teacher and student models we validate the model after each epoch on the same validation set and use the validation score as the stopping criteria.

Table 1: Statistics of deny list keywords, catalogue categorisation labels (CC) and relative scale of data for each label category

Defect category	CAT1	CAT2	CAT3	CAT4	CAT5
Count of keywords	315	240	36	45	50
Count of CC	26	111	27	1	20
Scale of Data	10	100	5	2	1

Table 2: Precision over Baseline, BERT(B) trained with limited labeled data, at our high Recall threshold for all models across defects.

Precision improvement at high Recall threshold over Baseline					
Method	CAT1	CAT2	CAT3	CAT4	CAT5
B_SS	+40.46	**+11.6**	+ 2.12	+4.85	+2.93
B_SWS	**+40.48**	+10.99	**+6.44**	**+8.02**	**+7.09**

3.1.2 Semi-Weak-Supervised(SWS) Methodology

Here we treat the augmented data as weakly labeled data and use it to pre-train teacher model before fine-tuning it with strong labeled data. This teacher model is used to score the top k% samples of the weakly labeled data which is used to pre-train new student model which is later fine-tuned using strong data. Here again while pre-training and fine-tuning teacher and student models we validate the model after each epoch on the same validation set and use the validation score as the stopping criteria.

3.1.3 Extension to low resource languages

We take the exact same approach of augmenting data for low resource languages and train the M-BERT model. With low resource languages we face two challenges. First, labelled data available here is less compared to English(EN). In German(DE) and French(FR), the scale of the positive class is of order 0.02-0.15 compared to scale of different defect categories for EN reported in Table 1. Second, keywords available for sourcing weakly labeled data is less which affects quality of sourcing weak data. To address these challenges we explore curriculum learning and multilingual training for low resource setting.

3.2 Curriculum for leveraging resource rich domains

In the above section we discussed augmenting data using weak signals. Here we explore how we can utilise large amounts of labeled data available in resource rich languages such as EN. We translate the ad creatives available in EN to the target language. Hence forth, this data is referred to as translated data. A trivial approach to utilise this data for tuning the model is to combine the strong and translated data and randomly sample mini-batches (B_TL_{RS}) from the unified set while training. Another possibility is to use the translated data to pretrain the classifier and fine-tune it with the strong data in target domain (B_TL_{FT}). Here, during every epoch, we initially train the model with the mini-batches sampled from the translated data followed by sampling mini-batches from strong data. This clearly has an advantage over the earlier approach as it helps model adapt to the target domain and avoid domain shift arising from the translation engine employed.

We also explore an approach leveraging curriculum learning that is agnostic of the distinction between translated and strong data for training the M-BERT model. Curriculum learning(Hacohen and Weinshall, 2019) involves using the prior knowledge of the difficulty of the training samples to sample training mini-batch. To rank the difficulty of the training sample (x_i, y_i) we need a scoring function. Scoring function $f : X \rightarrow R$ is any function which scores the difficulty of a given training sample. If $f(x_i, y_i) > f(x_j, y_j)$ then (x_i, y_i) is more difficult than (x_j, y_j). We also use a pacing function (Hacohen and Weinshall, 2019) which determines the sequence of subsets $X_1, .., X_m \subseteq X$ of size g_i from which mini-batches $\{B_i\}_{i=1}^{M}$ are sampled. These are generally monotonically increasing functions so the likelihood of the easier samples decrease over time.

In our case, we use BOE_LIN (See Section 2.3) as our scoring function- a proxy for hardness of the sample. Samples with confident predictions by BOE_LIN for positive and negative classes are considered easy while hardness increases as the samples are closer to boundary of separation. We initially pick the easier samples for the first x iterations. We augment the training samples with difficult samples progressively for every x iterations till all the data is seen by the model. In our case, we consider $x = 2$ and split the data into 5 sets of increasing difficulty. Iterations 1–2 are trained using the set having the most easy samples defined by the scoring function f. In iterations 3–4, we take the initial two sets of easy samples. In such a progression, the model sees the entire dataset in iterations 9–10. We use early stopping to choose the model at iteration i.

3.3 Multi-Lingual training of M-BERT

In Section 3.1 - 3.2, we explored methods of augmenting data from external sources for the same language i.e they were trained on monolingual data. However, in weak supervision, the quality of weak data is contingent on sourcing technique used. Using translated data from source domains risks introducing semantic drift due to inaccuracies in the translation engine used. Advertisers create ads for different markets and we have limited data in French(FR), Spanish(ES), Italian(IT) apart from English(EN) and German(DE). To mitigate these challenges, we explore multilingual training of M-BERT leveraging data from different languages to train a classifier for the target DE language thus avoiding sourcing technique to augment data.

Pires et al. (2019) show that M-BERT is good at zero shot cross lingual transfer where task specific text in one language is used for fine-tuning the model for a different target language. They further show that the transfer is more pronounced when there is more lexical overlap between the languages. They also show that transfer works with zero lexical overlap when the two languages are typologically similar i.e the ordering of subject, object and verbs among other parts of speech in a sentence. In our experiments we mainly rely on the lexical similarity between languages for training M-BERT. Table 4 (Wikipedia contributors, 2004) provides the lexical similarity between the languages for which we have labeled data. Lexical similarity score of 1 would mean total overlap between vocabularies and 0 would mean no overlap between vocabularies.

From entries for lexical similarity in Table 4, we observe that DE is lexically most similar to EN followed by FR. In case of missing values, we consider the corresponding languages as lexically farthest to the target language. Since M-BERT is trained on monolingual corpora and the above-mentioned 5 languages are among them, the vocabulary of M-BERT would have all the alphabets from these languages. On the basis of results evidenced in Pires et al. (2019), we hypothesise that

Table 3: Precision and TNR improvements at our high recall threshold for all the explored models for DE and FR languages using different training strategies and for ablation studies in Section 4.1 over BOE_LIN. Here B refers to M-BERT finetuned with limited labeled data.

		Model Type							Ablation Type			
		B	B_SS	B_SWS	B_TL_{RS}	B_TL_{FT}	B_TL_{CL}	B_ML_{LEX}	B_TL_{ACL}	B_TL_{RCL}	$B\_ML_{LEX}^{REV}$	$B\_ML_{LEX}^{RAND}$
DE	TNR	+14.35	+23.15	+24.79	+13.84	+23.76	+26.40	**+29.08**	+25.7	+21.0	+14.17	+26.90
	Prec.	+0.76	+1.69	+1.95	+0.72	+1.78	+2.24	**+2.83**	+2.11	+1.4	+0.75	+2.34
FR	TNR	+15.29	+19.16	+20.05	+15.10	+20.41	**+22.11**	+19.57	+21.02	+20.68	+12.71	+18.80
	Prec.	+0.77	+1.10	+1.19	+0.75	+1.23	**+1.43**	+1.14	+1.30	+1.26	+0.59	+1.07

Table 4: Lexical Similarity scores between languages of interest taken from Wikipedia.

Language	EN	DE	FR	ES	IT
EN	1	0.6	0.27	-	-
DE	0.6	1	0.29	-	-
FR	0.27	0.29	1	0.75	0.89
ES	-	-	0.75	1	0.82
IT	-	-	0.89	0.82	1

the zero shot transfer is more likely among similar lexical languages and devise our multi-language training of M-BERT in the following manner. We take the labeled data available in 5 languages and sort them based on increasing lexical similarity with the target language. For target language DE, the ordering would be ES, IT, FR, EN, DE. We feed all the data in the aforementioned ordering and progressively drop the lexically farthest language every x iterations until we are only left with the target language. In our case we set $x = 2$ and train the M-BERT. We generally stop training the model after 10 iterations since we do not observe significant gains beyond this.

4 Results

In all experiments, we track model performance using precision and recall. Precision indicates the fraction of ads correctly rejected by model. Recall indicates the fraction of true defective products rejected by the model for a particular category.

SS and SWS for EN Table 2 shows the improvement in precision for all the models built using the semi-supervision and semi-weak-supervised approaches. We see semi-supervision(B_SS) consistently perform better than the baseline, BERT finetuned with strong data, across all categories. For CAT1–2, we observe a substantial lift in precision over baseline compared to other categories. This is attributed to strong sourcing characteristics for these categories observed in Table 1. We observe significant gains by SWS(B_SWS) models especially in low resource categories like CAT3–5. For CAT3, CAT4 and CAT5 we see 6–8% better precision respectively.

Results for low resource languages In case of low resource languages the amount of defective ads is much lesser and is of order 0.02-0.15 as called out earlier. Since the quantity of positive class is drastically low, precision does not always indicate the true gains seen by our models. Hence we also report true negative rate(TNR) which is the % of non-defective ads rightly approved by our models.

Table 3 provides the relative improvements in metrics of all the models in comparison to baseline BOE_LIN. The complex and heavily parameterised M-BERT(B) model achieves a significant increase in TNR despite dearth of training data. From performance numbers in Table 3, we see that fine-tuning(B_TL_{FT}) the model with target domain after pre-training with translated data is better than random sampling(B_TL_{RS}) of mini-batches across strong and translated data. Plain augmentation of data through translation without any curriculum during training the model might not always show gains as indicated by M-BERT's performance. However, introducing a curriculum(B_TL_{CL}) based on the difficulty of the training samples outperforms the initial two approaches.

Table 3 also shows performance of weak supervision techniques (see Section 4.1). Models trained using both SS(B_SS) and SWS(B_SWS) approaches outperform the model which was trained only using the strong data.

We observe the best performance for the model (B_ML_{LEX}) leveraging data from multiple languages and trained in lexical order fashion. Since DE is lexically similar to EN, the larger training data in EN aided the model performance in this setting. We also rerun the experiments with FR with same setting and results are provided in Table 3. If we observe the lexical similarity in Table 4, FR is most similar to IT and ES and farther away from EN which has the most amount of labeled data. Hence, we do not see the similar kind of gains for FR as seen in DE which is lexically closer to EN. For FR the model trained using curriculum (B_TL_{CL}) based on the hardness of the sample

performs the best. We observe a similar trend in FR for rest of the approaches.

4.1 Ablations

We ablate the effects of curriculum learning based on increasing difficulty using models trained in two control conditions. (a) Anti-curriculum learning (B_TL_{ACL}) using scoring function $f' = -f$ where harder samples are fed first and (b) random curriculum (B_TL_{RCL}) where scoring function randomly scores the training samples. As seen from the Table 3 anti-curriculum and random curriculum are not as effective as the curriculum of increasing hardness. Further, random scoring function results in significant degradation of performance when compared to approaches employing a curriculum. Similar trends are observed for respective models trained in FR as well.

We further conduct ablations to rule out any other factors contributing to the gain in recall from curriculum based on lexical similarity. We perform two other experiments where we train the model in similar manner but feed the languages in reverse lexical similarity order($B\_ML_{LEX}^{REV}$) and random order ($B\_ML_{LEX}^{RAND}$). However, in both the experiments we feed the target language at the end to minimise domain shift. We see that the model trained in the lexical similarity order beats the performance of the other two models in Table 3. We validate statistical significance of gains from both lexical and hardness curricula using the McNemar's Test (Dietterich, 1998; McNemar, 1947) (Raschka, 2018). The gains through both curriculum are statistically significant as p-value is < 0.05 for both DE and FR.

5 Conclusion

We have explored multiple ways of training a GLM and it's multilingual variant in low resource settings. When large in-domain monolingual corpora is present but labeled data is limited, sourcing weak data applied in semi and semi-weak supervision training improves model performance consistently. Curricula are useful in resource constrained settings. Multilingual training on a lexical similarity based curriculum is useful when target language is lexically closer to resource rich languages. Alternate curriculum like sample hardness is useful in low resource languages which are lexically distant to resource rich language such as EN.

6 Related Work

Lately, there has been exponential progress in generating efficient embeddings for various natural language processing(NLP) tasks using language models (Radford et al. (2019); Liu et al. (2019)). BERT (Devlin et al., 2019) based embeddings achieved SOTA results in eleven NLP tasks at the time of its release. Devlin et al. (2019) also release a multilingual version of BERT(M-BERT), pre-trained using monolingual corpora of 104 different languages. M-BERT is also surprisingly good at zero shot transfer between languages as shown by Pires et al. (2019). Prior to and in parallel to M-BERT multiple works have been done for multilingual NLP tasks (Ruder et al., 2019). LASER described in Artetxe and Schwenk (2019) achieve language independent representation by having a single encoder and decoder which are shared by all language pairs for the translation task. Conneau and Lample (2019) propose using parallel data to train translation language model as an extension to M-BERT. Conneau et al. (2019) release XLM-R which is pretrained using 100 languages using much larger corpus compared to M-BERT.

Most of the recently launched language models have millions of parameters which demands huge amount of labelled data for training robust models. However, obtaining large amount of labeled data is a laborious and expensive process. Semi-supervised approaches involve efficiently incorporating huge quantity of unlabelled data along with limited labelled data. There has been a lot of work in this area in image and text domain. Yalniz et al. (2019) propose a teacher-student paradigm for incorporating both unlabelled and weakly labelled data for training a image classifier. Karamanolakis et al. (2019) also make use of teacher-student approach for leveraging weak signals for aspect detection in text. Variational auto encoders (Yang et al. (2017); Gururangan et al. (2019)) and virtual adversarial training(Miyato et al., 2016) have been extensively used in semi-supervised setting. Recently interpolations in textual hidden space(Chen et al., 2020) have been used for semi-supervised learning as well.

Multiple prior works (Sculley et al. (2011); Sanzgiri et al. (2018)) detect adversarial ads in online advertising platforms. While Sculley et al. (2011) provide a holistic view of creating an adversarial ad detection system, Sanzgiri et al. (2018) look at techniques for detecting sensitive content in images.

Our work focuses on techniques we leverage to train state of the art language models for detecting adversarial advertising content in text. However, the uncommon nature of these violations pose a challenge, often compounded in low resource languages. We leverage related work in semi-weak supervision and curriculum learning to overcome these challenges. We also show how data available in multiple languages can be used for training classifiers for a given target language.

References

Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2017. A simple but tough-to-beat baseline for sentence embeddings.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Jiaao Chen, Zichao Yang, and Diyi Yang. 2020. Mixtext: Linguistically-informed interpolation of hidden space for semi-supervised text classification. *arXiv preprint arXiv:2004.12239*.

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. *arXiv preprint arXiv:1911.02116*.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, pages 7059–7069.

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

Thomas G. Dietterich. 1998. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Comput.*, 10(7):1895–1923.

Suchin Gururangan, Tam Dang, Dallas Card, and Noah A Smith. 2019. Variational pretraining for semi-supervised text classification. *arXiv preprint arXiv:1906.02242*.

Guy Hacohen and Daphna Weinshall. 2019. On the power of curriculum learning in training deep networks. *CoRR*, abs/1904.03626.

Giannis Karamanolakis, Daniel Hsu, and Luis Gravano. 2019. Leveraging just a few keywords for fine-grained aspect detection through weakly supervised co-training. *arXiv preprint arXiv:1909.00415*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

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

Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

Takeru Miyato, Andrew M Dai, and Ian Goodfellow. 2016. Adversarial training methods for semi-supervised text classification. *arXiv preprint arXiv:1605.07725*.

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. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Sebastian Raschka. 2018. Mlxtend: Providing machine learning and data science utilities and extensions to python's scientific computing stack. *Journal of Open Source Software*, 3(24):638.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2019. A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research*, 65:569–631.

Ashutosh Sanzgiri, Daniel Austin, Kannan Sankaran, Ryan Woodard, Amit Lissack, and Sam Seljan. 2018. Classifying sensitive content in online advertisements with deep learning. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 434–441. IEEE.

D Sculley, Matthew Eric Otey, Michael Pohl, Bridget Spitznagel, John Hainsworth, and Yunkai Zhou. 2011. Detecting adversarial advertisements in the wild. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 274–282.

Konstantinos Sechidis, Grigorios Tsoumakas, and Ioannis Vlahavas. 2011. On the stratification of multi-label data. *Machine Learning and Knowledge Discovery in Databases*, pages 145–158.

Dinghan Shen, Guoyin Wang, Wenlin Wang, Martin Renqiang Min, Qinliang Su, Yizhe Zhang, Chunyuan Li, Ricardo Henao, and Lawrence Carin. 2018. Baseline needs more love: On simple word-embedding-based models and associated pooling mechanisms. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 440–450, Melbourne, Australia. Association for Computational Linguistics.

Piotr Szymański and Tomasz Kajdanowicz. 2017. A network perspective on stratification of multi-label data. In *Proceedings of the First International Workshop on Learning with Imbalanced Domains: Theory and Applications*, volume 74 of *Proceedings of Machine Learning Research*, pages 22–35, ECML-PKDD, Skopje, Macedonia. PMLR.

Wikipedia contributors. 2004. Plagiarism — Wikipedia, the free encyclopedia. [Online; accessed Feb-2020].

I. Zeki Yalniz, Hervé Jégou, Kan Chen, Manohar Paluri, and Dhruv Mahajan. 2019. Billion-scale semi-supervised learning for image classification. *CoRR*, abs/1905.00546.

Zichao Yang, Zhiting Hu, Ruslan Salakhutdinov, and Taylor Berg-Kirkpatrick. 2017. Improved variational autoencoders for text modeling using dilated convolutions. *arXiv preprint arXiv:1702.08139*.