The Effect of Round-Trip Translation on Fairness in Sentiment Analysis

Jonathan Gabel Christiansen* and **Mathias Lykke Gammelgaard*** and **Anders Søgaard**
Dpt. of Computer Science, University of Copenhagen

Abstract

Sentiment analysis systems have been shown to exhibit sensitivity to protected attributes. Round-trip translation, on the other hand, has been shown to normalize text. We explore the impact of round-trip translation on the demographic parity of sentiment classifiers and show how round-trip translation consistently improves classification fairness at test time (reducing up to 47% of between-group gaps). We also explore the idea of retraining sentiment classifiers on round-trip-translated data.

1 Introduction

It is both unethical and potentially illegal for document classification algorithms to perform significantly better for some groups in society than for others (Mehrabi et al., 2019). Many document classification technologies have, however, been shown to be sensitive to protected attributes such as gender and age (Mehrabi et al., 2019; Delobelle et al., 2020; Ferrer et al., 2020; Koh et al., 2021). This also holds for sentiment analysis (Johannsen et al., 2015; Hovy, 2015; Kiritchenko and Mohammad, 2018; Bhaskaran and Bhallamudi, 2019; Touileb et al., 2020). At the same time it is known that round-trip machine translation (Huang, 1990; Federmann et al., 2019) can be used to normalize text (Ling et al., 2013; Rabinovich et al., 2017; Prabhumoye et al., 2018). This could potentially remove group specific deviations from normal language. However, Stanovsky et al. (2019) found that machine translation is also prone to potentially introducing gender bias. Combined, this leaves it an open question whether round-trip translation can be used to reduce the sensitivity of document classifiers to protected attributes of the authors.

In this paper, we evaluate the effect of round-trip translation on fairness using a representative corpus of Danish Trustpilot reviews, in which reviews are associated with self-reported protected attributes (gender and age) (Hovy et al., 2015). We evaluate this effect across nine different document classification architectures, both in the setting in which round-trip translation happens at test time only, and in the setting in which both training and test data are translated to a foreign language and back.

Text
Sally is a whiz at math.
Sally es una experta en matemáticas.
Sally is a math expert.

Table 1: The normalizing effect of round-trip translation: translating English to Spanish and back.

Contributions We evaluate round-trip translation as a technique for mitigating sensitivity to protected attributes across two attributes and three document classification architectures. We find that round-trip translation at test time consistently reduces the fairness gap (with up to 47%), but that for our best models (SVMs stacked on BERT representations), the effect disappears when both training and test data are translated into a foreign language and back.

2 Round-trip Translation

Round-trip machine translation (Huang, 1990; Federmann et al., 2019) is the process of machine translating a document to another language and then back to the original language. Table 1 shows a toy example of this process. Ling et al. (2013) found that machine translations of human translations of English tweets into Chinese, back into English, had a tendency to normalize the original text. Rabinovich et al. (2017) observed a similar normalization effect in machine translation systems, and based on their observation, Prabhumoye

* Equal contributions. Corresponding author: soegaard@di.ku.dk.

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 4423–4428
November 7–11, 2021.

Group	Characteristics	# of reviews
1	Male, Age ≤ 42	7628
2	Female, Age ≤ 42	5020
3	Male, $42 <$ Age ≤ 55	7081
4	Female, $42 <$ Age ≤ 55	5571
5	Male, $55 <$ Age	7535
6	Female, $55 <$ Age	4489

Table 2: Test groups

et al. (2018) subsequently proposed to use round-trip translation for style transfer. In this paper, we investigate whether this observed normalization effect can be used to mitigate fairness problems in document classification. Specifically, we will use Google cloud translation[1] to translate Danish product reviews from the Trustpilot corpus (into English and back), to obtain a (machine) normalized version of this corpus, which we make publicly available.[2]

3 Experiments

Fairness metric The fairness literature is rich with definitions of fairness (Mehrabi et al., 2019), most of which are interpretations of the Rawlsian notion of fairness as *equal opportunity* (Rawls, 1971). In this work, we adopt the following definition of fairness:

Definition 3.1 (Fairness as equal risk)**.** A model θ is fair for a set of groups $\mathcal{G}$, if $\mathbb{E}[\ell(\theta(\mathcal{X}_g), \mathcal{Y}_g)] = \mathbb{E}[\ell(\theta(\mathcal{X}_h, \mathcal{Y}_h)]$ for any $g, h \in \mathcal{G}$.

If the maximal difference in empirical risk between any two groups in $\mathcal{G}$ is ϵ, we say θ is ϵ-fair. Below we use the F1-score for the negative class as our $\ell(\cdot)$.[3] Note how fairness as equal risk is a generalization of approximately equal conditional risk (Donini et al., 2018).

Data The Danish section of the Trustpilot Corpus consists of 149,240 reviews annotated for (self-reported) sentiment, gender and age. The sentiment ratings are provided on a scale from 1 to 5, which we binarize by mapping low ratings to negative class, i.e., $\{1, 2, 3\} \mapsto 0$, and high ratings to positive class, i.e., $\{4, 5\} \mapsto 1$. This leads to a highly skewed distribution of 8,257 negative reviews and 140,983 positive ones. This also motivates the use of F1-score for the negative class as our performance metric.

We randomly split the data set into training and test leaving 75% of the reviews as training data, and 25% of the reviews as our test data. The test set is further split into six demographic groups according to self-reported gender[4] and age (binned in three equal groups), as presented in Table 2. We use these six roughly equal-sized groups to evaluate the fairness of our models.

Impact of round-trip translation We use KL-divergence[5] to get a first impression of the extent to which round-trip translation normalizes our data. This is done across the most frequent 1,000 words in the Trustpilot corpus. For each group, we calculate the probability distribution for these words and compute its KL-divergence to the overall distribution. We do this before and after round-trip translating. Table 4 lists the KL-divergencies before and after round-trip translation. As expected, we observe a significant *decrease* in KL-divergence for all groups after round-trip translating. This indicates that the reviews were indeed normalized during the process of round-trip translation. We also see that the number of unique words dropped by 36% after round-trip translation.

Document classifiers Our document classifiers all rely on vector representations from pretrained language models. We use two different pretrained language models, namely the multilingual LASER model (Artetxe and Schwenk, 2019)[6] and a monolingual BERT (Devlin et al., 2019) trained for Danish.[7] On top of these we train several classifiers, including nearest neighbor, logistic regression, and (Gaussian kernel) support vector machines (SVMs). We set regularization parameters through grid-search and cross-validation over the training data, but also report results for unregularized logistic regression and SVMs. See Table 3 for hyper-parameters and results.

4 Results

In Table 3, we evaluate three document classification architectures across three scenarios: (a) a

[1] https://cloud.google.com/translate/

[2] https://github.com/anonymous/

[3] This is motivated by the observation that in practice, most end users of sentiment analysis systems are interested in identifying *negative* reviews.

[4] The Trustpilot corpus contains only two values for gender.

[5] https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.entropy.html

[6] https://pypi.org/project/laserembeddings/

[7] https://pypi.org/project/danish-bert-embeddings/

Model	Params	Min	Max	Fairness Gap	Reduction?
REVIEWS → REVIEWS (baseline condition)					
BERT-KNN	$k = 6$	0.428	0.628	0.200	Undefined
BERT-LR	ℓ_2-regularized	0.628	0.748	0.120	
BERT-LR	unregularized	0.633	0.735	0.102	
BERT-SVM	$C = 10, rbf$	0.642	0.750	0.108	
BERT-SVM	$C = 1000, rbf$	0.625	0.744	0.119	
LASER-LR	ℓ_2-regularized	0.009	0.045	0.036	
LASER-LR	unregularized	0.519	0.693	0.174	
LASER-SVM	$C = 10, rbf$	0.570	0.737	0.167	
LASER-SVM	$C = 1000, rbf$	0.573	0.737	0.164	
REVIEWS → ROUND-TRIP					
BERT-KNN	$k = 6$	0.420	0.571	0.151	⇓
BERT-LR	ℓ_2-regularized	0.608	0.712	0.104	⇓
BERT-LR	unregularized	0.608	0.706	0.098	⇓
BERT-SVM	$C = 10, rbf$	0.612	0.710	0.098	⇓
BERT-SVM	$C = 1000, rbf$	0.615	0.710	0.095	⇓
LASER-LR	ℓ_2-regularized	0.009	0.048	0.039	⇑
LASER-LR	unregularized	0.529	0.682	0.153	⇓
LASER-SVM	$C = 10, rbf$	0.580	0.705	0.125	⇓
LASER-SVM	$C = 1000, rbf$	0.586	0.705	0.119	⇓
ROUND-TRIP → ROUND-TRIP					
BERT-KNN	$k = 6$	0.468	0.575	0.107	⇓
BERT-LR	ℓ_2-regularized	0.605	0.725	0.120	⇒
BERT-LR	unregularized	0.617	0.726	0.109	⇑
BERT-SVM	$C = 10, rbf$	0.592	0.742	0.150	⇑
BERT-SVM	$C = 1000, rbf$	0.600	0.731	0.131	⇑
LASER-LR	ℓ_2-regularized	0.610	0.730	0.120	⇑
LASER-LR	unregularized	0.545	0.689	0.144	⇓
LASER-SVM	$C = 10, rbf$	0.589	0.713	0.124	⇓
LASER-SVM	$C = 1000, rbf$	0.588	0.715	0.127	⇓

Table 3: We use F_1 score of positive class as our performance metric. We make several observations: (a) Round-trip translation at test time consistently reduces the fairness gap, and with up to 47%. (b) Round-trip translation of training and test data reduces the fairness gap for LASER models, but widens it for BERT models. (c) Generally, LASER models seem less fair than BERT models, and unregularized models seem more fair than regularized ones. The latter observation aligns with previous work indicating that sparseness is at odds with robustness and fairness (Globerson and Roweis, 2006; Søgaard, 2013; Khani and Liang, 2021).

Group	KLD REVIEWS	KLD ROUND-TRIP
1	0.027	0.021
2	0.028	0.023
3	0.011	0.009
4	0.023	0.021
5	0.028	0.022
6	0.027	0.020

Table 4: The KL-divergence between the probability distribution of the 1000 most frequent words in each group and the general distribution, before and after round-trip translation. **Round-trip translation reduces group-level divergences.**

baseline condition in which classifiers are trained and evaluated on Trustpilot reviews; (b) a scenario in which reviews are round-trip-translated at test time for normalization; and (c) a condition in which the classifiers are retrained on round-trip-translated reviews and evaluated on round-trip-translated reviews.

Test time normalization with round trip translation (b) has an overall positive effect on cross-group generalization, reducing the fairness gap with up to $\sim 27\%$. The third scenario (c) – i.e., the idea of using round-trip translation for normalizing *both* the training and the test data – yields mixed results, with fairness gap increases up to $\sim 39\%$ (for BERT models) and decreases up to $\sim 47\%$ (for LASER models). Machine translation introduces its own biases, and some representations may be more sensitive to such biases. Note also that the process of round-trip translating the data consistently reduces the overall accuracy of our document classifiers, suggesting a trade-off between fairness and accuracy.

5 Discussion

Round-trip translation is a simple technique for test-time input normalization, and we have shown that it can significantly reduce sensitivity to protected attributes at a low performance cost. One advantage of round-trip translation is that it does not require annotation of protected attributes. Such datasets are generally only available for English at this point. High-quality machine translation, in contrast, is available for hundreds of languages. In this paper, we experiment with using round-trip translation to reduce group disparity of sentiment classifiers for Danish.

It is important to note, however, that the overall performance drop that results from round-trip translation, while relatively small, means that the absolute performance on minority group *drops*. In other words, all users experience worse performance with the more fair sentiment classifiers. This, of course, is unfortunate and potentially introduces an ethical dilemma. In fact, it is only with our LASER models that minority group performance improves *and* fairness is reduced.

Round-trip translation is orthogonal to other approaches to improving fairness, such as distributionally robust optimization (Sagawa et al., 2020), invariant risk minimization (Arjovsky et al., 2020), and adversarial training (Dayanik and Padó, 2021). Round-trip translation can thus easily be combined with any of these approaches, but note that these approaches require annotation of protected attributes. Round-trip translation does not and can thus be considered an *unsupervised* approach to reducing group disparities.

The fairness gap was most consistently reduced by test-time round-trip translation, but doing round-trip translation may be more effective for other machine translation systems. In our experiments, Google Translate *introduced* new biases when relying on BERT representations, but the approach was successful for document classifiers based on LASER representations: Here, we saw *both* reductions of the fairness gap and improvements for minority groups. For 2/4 classifiers, we even saw improvements for the majority groups.

6 Conclusion

Sentiment classifiers perform better on reviews written by some demographic groups rather than others, with groups defined by protected attributes such as gender and age. We present a first experiment with round-trip translation as a means of reducing this fairness gap in sentiment classification. Specifically, we show that translating Danish product reviews into English and back, reduces group disparity across three different classification architectures. While the performance cost may in our case be prohibitive for some architectures, in practice, we believe that round-trip translation can be an important technique for improving the fairness of document classifiers in the future, which is easier to scale to new tasks and languages than approaches that require annotation of protected attributes.

Ethics Statement

The gender and age information in the Trustpilot Corpus is self-reported, and all reviewers were free to *not* report this information. All reviewers that supplied gender information, identified as either male or female, but were free to report other genders.

Acknowledgement

This work was supported by the Innovation Fund Denmark (Grants no. 0175-00011A and 0175-00014B).

References

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. 2020. Invariant risk minimization.

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.

Jayadev Bhaskaran and Isha Bhallamudi. 2019. Good secretaries, bad truck drivers? occupational gender stereotypes in sentiment analysis. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 62–68, Florence, Italy. Association for Computational Linguistics.

Erenay Dayanik and Sebastian Padó. 2021. Disentangling document topic and author gender in multiple languages: Lessons for adversarial debiasing. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 50–61, Online. Association for Computational Linguistics.

Pieter Delobelle, Paul Temple, Gilles Perrouin, Benoît Frénay, Patrick Heymans, and Bettina Berendt. 2020. Ethical adversaries: Towards mitigating unfairness with adversarial machine learning.

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.

Michele Donini, Luca Oneto, Shai Ben-David, John S Shawe-Taylor, and Massimiliano Pontil. 2018. Empirical risk minimization under fairness constraints. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Christian Federmann, Oussama Elachqar, and Chris Quirk. 2019. Multilingual whispers: Generating paraphrases with translation. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 17–26, Hong Kong, China. Association for Computational Linguistics.

Xavier Ferrer, Tom van Nuenen, Jose M. Such, Mark Coté, and Natalia Criado. 2020. Bias and discrimination in ai: a cross-disciplinary perspective.

Amir Globerson and Sam T. Roweis. 2006. Nightmare at test time: robust learning by feature deletion. In *ICML*, volume 148 of *ACM International Conference Proceeding Series*, pages 353–360. ACM.

Dirk Hovy. 2015. Demographic factors improve classification performance. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 752–762, Beijing, China. Association for Computational Linguistics.

Dirk Hovy, Anders Johannsen, and Anders Søgaard. 2015. User review sites as a resource for large-scale sociolinguistic studies. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15, page 452–461, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Xiuming Huang. 1990. A machine translation system for the target language inexpert. In *COLNG 1990 Volume 3: Papers presented to the 13th International Conference on Computational Linguistics*.

Anders Johannsen, Dirk Hovy, and Anders Søgaard. 2015. Cross-lingual syntactic variation over age and gender. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 103–112, Beijing, China. Association for Computational Linguistics.

Fereshte Khani and Percy Liang. 2021. Removing spurious features can hurt accuracy and affect groups disproportionately. In *FAccT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual Event / Toronto, Canada, March 3-10, 2021*, pages 196–205. ACM.

Svetlana Kiritchenko and Saif Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 43–53, New Orleans, Louisiana. Association for Computational Linguistics.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton A. Earnshaw, Imran S. Haque, Sara Beery, Jure Leskovec,

Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. 2021. Wilds: A benchmark of in-the-wild distribution shifts.

Wang Ling, Chris Dyer, Alan W Black, and Isabel Trancoso. 2013. Paraphrasing 4 microblog normalization. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 73–84, Seattle, Washington, USA. Association for Computational Linguistics.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2019. A survey on bias and fairness in machine learning.

Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W Black. 2018. Style transfer through back-translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 866–876, Melbourne, Australia. Association for Computational Linguistics.

Ella Rabinovich, Raj Nath Patel, Shachar Mirkin, Lucia Specia, and Shuly Wintner. 2017. Personalized machine translation: Preserving original author traits. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1074–1084, Valencia, Spain. Association for Computational Linguistics.

John Rawls. 1971. *A Theory of Justice*, 1 edition. Belknap Press of Harvard University Press, Cambridge, Massachussets.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. 2020. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization.

Anders Søgaard. 2013. Zipfian corruptions for robust POS tagging. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 668–672, Atlanta, Georgia. Association for Computational Linguistics.

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy. Association for Computational Linguistics.

Samia Touileb, Lilja Øvrelid, and Erik Velldal. 2020. Gender and sentiment, critics and authors: a dataset of Norwegian book reviews. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 125–138, Barcelona, Spain (Online). Association for Computational Linguistics.