Reducing cohort bias in natural language understanding systems with targeted self-training scheme

Dieu-Thu Le
Amazon Alexa AI
deule@amazon.com

Gabriela Cortes
Amazon Alexa AI
cortgab@amazon.com

Bei Chen
Amazon Alexa AI
chenbe@amazon.com

Melanie Bradford
Amazon Alexa AI
neunerm@amazon.com

Abstract

Bias in machine learning models can be an issue when the models are trained on particular types of data that do not generalize well, causing under performance in certain groups of users. In this work, we focus on reducing the bias related to new customers in a digital voice assistant system. It is observed that natural language understanding models often have lower performance when dealing with requests coming from new users rather than experienced users. To mitigate this problem, we propose a framework that consists of two phases (1) a fixing phase with four active learning strategies used to identify important samples coming from new users, and (2) a self training phase where a teacher model trained from the first phase is used to annotate semi-supervised samples to expand the training data with relevant cohort utterances. We explain practical strategies that involve an identification of representative cohort-based samples through density clustering as well as employing implicit customer feedbacks to improve new customers' experience. We demonstrate the effectiveness of our approach in a real world large scale voice assistant system for two languages, German and French through a number of experiments.

1 Introduction

Deep machine learning models tend to inherit the bias existing in the datasets used for training (Manzini et al., 2019) (Zhao et al., 2017). For example, GPT-3 a state of the art in contextual language model, showed bias regarding religion, race and gender (Brown et al., 2020). Even though deep learning models are trained on large amounts of data, it is hard to capture all the variations of the language that different users can use. Even within the same language people talk differently, depending on the age group, part of the country, background, etc (Kern et al., 2016) (Eisenstein et al., 2010) (Hovy and Søgaard, 2015). If the training data is skewed towards a certain demographic group, this can cause models to pick up on patterns that do not generalize and underperform on certain user groups. Bias on predictive models is an issue that has been studied for some time. Most of the related literature is focused on social bias, specially gender and race (Zhao et al., 2017) (Manzini et al., 2019) and centered on measuring an specific type of bias and providing contra measures for it, which usually do not generalize to other types of bias (Zhao et al., 2018) (Goldfarb-Tarrant et al., 2020) (Garrido-Muñoz et al., 2021) (Dixon et al., 2018) (Shah et al., 2020). For example, on digital assistants, we identify other types of group bias, like customer tenure. Everyday, new customers join services like Amazon Alexa, Siri or Google Home. These new customers experience digital assistants for the first time and interact with it differently than mature cohort. New customers tend to try out more different functionalities, while mature customers often use utterances that work for them and settle down in daily-related domains. The experience of new customers is a closer reflection of how natural communication looks like as they are not yet "taught" how to communicate with the devices. Learning from new customers therefore might be one of the best ways to learn natural interactions with digital assistants. Contrary to most studies that focus on using semi-supervised learning for general accuracy (Chapelle et al., 2009a) (Clark et al., 2018) (Ding et al., 2018) (Hinton et al., 2015), we focus on improving the accuracy of the new customer (early cohort) natural language understanding task (McClosky et al., 2006) and show that our framework could target a strategic customer cohort to improve their experiences, thus improve the overall accuracy in all customers in an industry scale experiment. Even so, our approach can be easily applied to any customer cohort to mitigate other types of bias. Our proposal consists in a method to identify important utterances coming from the early cohort that need to be fixed, then em-

ploy self training techniques to mitigate them. The main idea is to automatically expand the training data to increase the representativeness of utterances that characterize early cohort customers.

2 Related work

Detecting and mitigating bias in model predictions have attracted a lot of studies recently. For example, (Zhao and Chang, 2020) proposed a bias detection technique based on clustering. Their approach focuses in local bias detection, which refers to bias exhibited in a neighborhood of instances rather than on the entire data. (Garrido-Muñoz et al., 2021) did a survey on bias in deep NLP, where they present a review of the state-of-the-art in bias detection, evaluation and correction, where they used vector space manipulation (Bolukbasi et al., 2016), data augmentation, data manipulation or attribute protection for dealing with the bias. (Shah et al., 2020) proposed a predictive bias framework for NLP and identified four potential origins of biases: label bias, selection bias, model over-amplification, and semantic bias. To mitigate model bias, common methods such as adversarial learning (Li et al., 2018; Le et al., 2022b), data augmentation with synthetic data generation using back translation (Sennrich et al., 2016), pretrained language model (Sahu et al., 2022; Wang et al., 2021; Kobayashi, 2018; Kumar et al., 2019; Le et al., 2022a) and semi-supervised learning (Cho et al., 2019; Zhu, 2005; Zhu and Goldberg, 2009; Chapelle et al., 2009b) have shown to be effective, especially when there is a lack of labeled data. While most of these studies focus on general biases in training models, we specifically aim at new customer cohort in a real world large scale voice assistant system. We employ both active and semi-supervised learning approaches that take customer feedback into consideration to improve the model prediction on this specific cohort.

3 Natural language understanding task in early cohort

Early cohort is defined as a group of new customers who have started to use the voice assistant device within the last 7 days. In contrast, **mature cohort** refers to the group of customers that have used the device for at least more than 30 days. Typically, a voice assistant consists of different components, starting from WakeWord detection, to Automatic Speech Recognition (ASR) that converts voice signals to texts, which will be used by the Natural Language Understand (NLU) component. In this work, we focus on how improving the NLU part could help to improve the end to end experience of new customers. In this study, we use devices' response results and weak signals as a way to improve the system over time. In particular, **friction** is defined as commands from customers that the system failed to provide an answer to (e.g., when the system gives responses such that "sorry I do not understand"). We also consider **negative feedback** from customers as a signal that the system did not response well to their previous requests. Finally, in order to measure the impact of our approach, we carried out offline NLU experiments (testing on annotated data). The only change is the implemented early cohort self training scheme.

4 Our approach

We propose an end to end framework (Figure 1) to identify cohort representativeness and effective data selection and augmentation to improve the model performance on a specific cohort without degrading the overall performance. It is composed of two phases, with the first phase looks for utterances from early cohort that need to be fixed using active learning using different strategies. After these utterances are annotated with human annotators, they are included in the training data to train a new NLU teacher model. We then employ a self learning phase to further extend similar utterances using semi-supervised learning to have more representatives of samples coming from early cohort.

4.1 Active Learning strategies

We define phase I with active learning strategies to fix important utterances from early cohort that the model might struggle with. The aim is to select all utterances with the highest values to be annotated to improve the performance of the NLU system with a given budget of ζ annotated utterances.

Let $\mathcal{L} = \{(x_i, y_i)\}_{i=1}^{|L|}$ be a set of labeled training data that is currently used for the NLU model F with $x_i \in X^L$, a set of all labeled utterances including customer and/or synthesis utterances. We have $\mathcal{U} = \{(x_i, y_i')\}_{i=1}^{|U|}$ as a set of unlabelled data, which contains $x_i \in X^U$, a set of all unlabelled utterances. y_i denotes labels from human annotators while y_i' denotes the annotation coming from NLU model F. y_i' contains the first hypothesis from F model and additional information about the

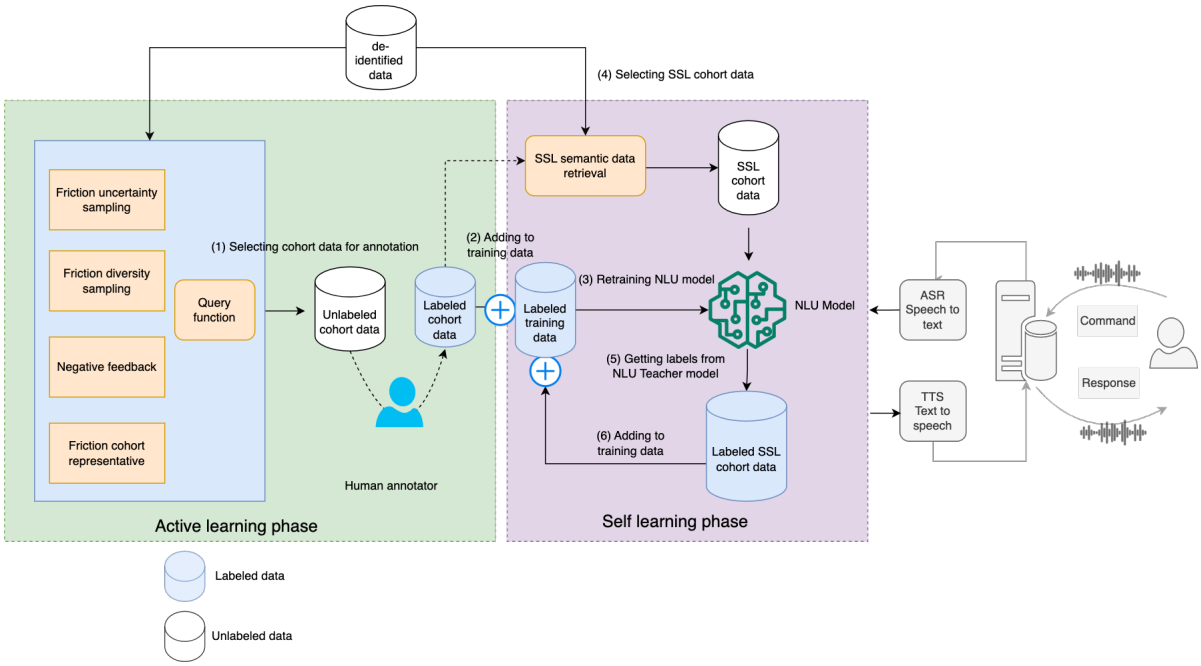

Figure 1: Our framework combines active and semi-supervised learning (self learning phase) to minimize labeling cost while improving the accuracy of early cohort: (1) selecting cohort-based data to be sent to human annotation, (2) adding the annotated data to training data, (3) retraining the NLU model based on the added dataset, (4) self learning scheme, extending the annotated cohort-based data with semi-supervised learning (SSL), (5) using the trained NLU teacher model to get labels for the SSL data.

corresponding utterance such as whether it belongs to the early cohort EARLY (i.e., utterances that occur during the 30 first days of the customers), the frictional group FRICTION or the LOW bin (i.e., utterances that have low confidence scores during NLU prediction).

To select utterances that are relevant and important for new customers, we employ four different sampling strategies with a set of acquisition functions $a = \langle a^{(1)}, a^{(2)}, a^{(3)}, a^{(4)} \rangle$ to select a set of $\mathcal{T}$ utterances to be annotated with $\mathcal{T} = \{(x_i, y_i)|x_i \in X\}_{i=1}^{|\zeta|}$. The goal is to find the set of all $X = \{x_i|x_i \in \mathcal{U}\}$ that provides the best model's performance $P(F')$ of model F' that is trained on $\mathcal{L}' = \mathcal{L} \cup \mathcal{T}$.

We cover (1) difficult utterances (uncertainty sampling), (2) wide coverage (diversity sampling) as well as (3) utterances that are representative of new customers (cohort-representativeness sampling) and finally (4) using customer feedbacks as an additional signal to trace back problematic utterances. The selection function gives us a set of classified utterances focusing on early cohort. The final set is the union of all four strategies (Algorithm 1).

4.1.1 Uncertainty and diversity sampling

As a common approach in active learning, the first sampling strategy is to query for utterances that have low NLU confidence scores and utterances where texts are similar, but NLU hypotheses are different. Those are utterances that the model are unsure about its predictions. For diversity sampling, we select representative frictional utterances using k-means clustering, extracting the centroid of each cluster to get a set of representative broken utterances from early cohort.

4.1.2 Identification of cohort representativeness

To get a visualization on a target cohort friction data, we propose the following approach that takes into account contextual information embedded in BERT representations together with hidden topic modeling (Blei et al., 2003). While BERT embedding provides contextual information about how words are interacting and accompanying each other, topics project utterances to a hidden topical space that is easy for interpretation. We then compare the density area for each cohort with topic guidance to identify the areas that are representative of friction utterances from early cohort. The process is composed of two main steps (1) Step 1: inspired from

BERT topic combination (Bianchi et al., 2021), we perform parameter estimation and data fitting, where the LDA (Latent Dirichlet Allocation) (Blei et al., 2003) topic estimation, Auto Encoder (Liou et al., 2014) and Uniform Manifold Approximation and Projection (UMAP) (McInnes et al., 2018) are learned from generic frictional data. Note that this training phase is completely unsupervised, only the first NLU hypothesis domain labels are integrated into the training data for a better domain focus representation. (2) Step 2: Topic inference with BERT representation and transforming friction data from different cohorts (e.g. early and mature cohort) separately through Autoencoder and UMAP, we use density clustering with topic guidance to identify the areas that are representative of friction utterances from the early cohort.

Training with the original data gave rather poor results since the friction data is very unbalanced with main focus on the bigger domains (e.g Music and Knowledge). LDA is not able to capture correctly other domains and classes when training on original data, but gave a much better results after upsampling minority classes. Furthermore, integrating NLU domain label hypotheses gives another dimension of information, hence improve domain focus and give a better labelling for interpreting topics. In the inference phase (Figure 2), friction data is included in its original distribution (e.g., no resampling is used). Early and mature cohort friction data are fed separately into the models. Before doing UMAP transformation, we employ density clustering with topic guidance to extract utterances that most characterize early cohort (i.e., are often asked by the early cohort and gave them frictions in compared to mature cohort). The visualization of early and mature cohort give insights into which topics are mostly asked, identifying domains that are usually confused to each other, top words that are used in each domain/topic that lead to friction. This helps to understand which types of requests from new customers need to be fixed.

4.1.3 Using customer feedback inputs

In this sampling approach, we look at utterances from new customers that might contain negative feedbacks (NF). To this end, we employ a binary classifier that predicts whether an utterance contains a negative feedback (e.g., "this is not what I meant", "you did not understand it"). If it is likely that an utterance contains a negative feedback, we trace back to the previous de-identified utterance that might have led to the negative feedback. This is the fourth sampling strategy used for querying utterances for active learning.

4.2 Data augmentation with self-training scheme

Since the budget of ζ annotated utterances is limited, we want to combine SSL together with data augmentation as the second step for enriching the training data with utterances that best solve the problems of young cohort. Many recent studies have shown that augmented data with semi-supervised learning (Chapelle et al., 2006) can boost the performance of text classification tasks with reduced number of annotated data. We integrate them together with the utterances selected for annotation in Phase I in a self-training scheme to select best utterances that can be augmented into the training set. In particular, the process consist of the following steps:

1. Take all data coming from $\mathcal{T}$ and with the output NLP model F', retrained in Phase I.

2. F' runs on a new unlabelled set of utterances to achieve H1 and their scores.

3. Construct the set $\mathcal{T}_{ssl}$ that contains all selected utterances that are most similar to those coming from $\mathcal{T}$ with the highest confidence to be added to the training data with a data retrieval module based on similarity search.

Figure 3 shows how the SSL data selection works, where we search for early cohort most relevant utterances from the live traffic. Due to the large scale of the data, it is prohibitively expensive to search for relevant utterances from the de-identified live traffic data using pair wise similarity search. Therefore, we encoded and indexed all utterances once, clustering the data where each cluster is represented by their centroids, which are used as inverted file and indices (Johnson et al., 2019). For each of the selected early cohort utterances (that were annotated in Phase I), we find those that are most semantically similar (but are not identical). When a query vector comes in, a most suitable cluster found based on its similarity with the centroids is returned together with the top K-nearest utterances coming from the live traffic data.

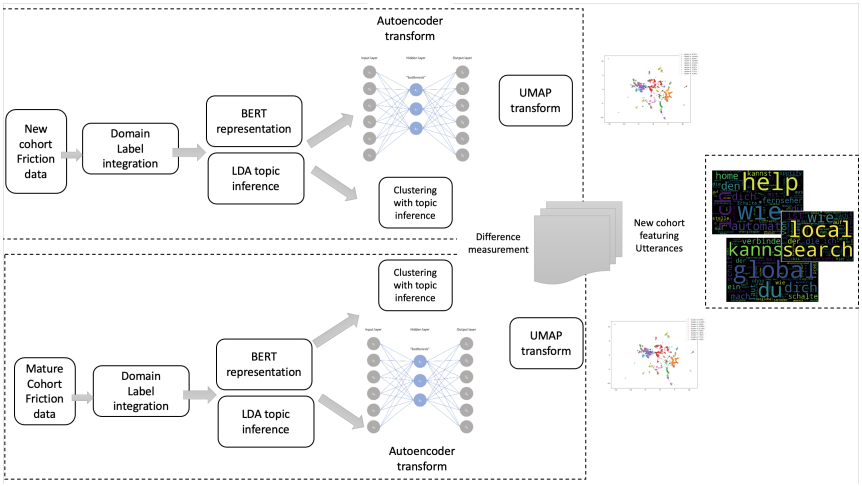

Figure 2: Identification of cohort representative utterances

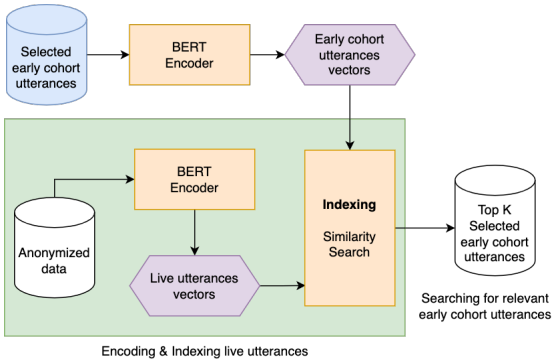

Figure 3: Indexing and searching module

5 Experiment setup

5.1 Models and dataset

We took aggregated and de-identified data for evaluating our framework for both German (DE) and French (FR) languages. The offline results are evaluated on human annotated test set that comes from the live traffic distribution. For the first phase we got 8K annotated utterances. For the second phase, we further enriched with ∼13K utterances using semantic retrieval for SSL. The offline results are reported with a sample test set containing 1M samples for DE and 800K samples for FR.

5.2 Metrics

We report our offline results in semantic error rate (SEMER), which is calculated by the number of errors (at slot and intent level) divided by the total number of reference slots and intent classification error rate (ICER), which takes only intent classification error into consideration (see A.2 for more information).

6 Results

We report offline results testing on annotated test data.

Table 1 shows the relative changes of each phase in compared to a baseline model for both ICER and SEMER metrics. We observe a constant improvement across domains for both phases in German (DE) and French (FR) languages. In particular, the biggest gain (6.99% intent error and 6.1% semantic error reduction) is observed in German second phase, where we include semi-supervised learning with focus on early cohort. Among all domains, we see especially good improvements in Help, Notifications and Knowledge domains. These domains are also popular domains among new customers, who tend to try out different functionalities and require support (Help) from the devices to understand how to use them. We see also some small degradation (0.29% in ICER for French phase II), but no degradation with SEMER metric, when we take also slot information into consideration. Overall, the offline results show that both active and semi-supervised learning are effective in improving the performance of the model.

7 Conclusions

In this work, we provide an end to end framework for bias mitigation with a focus on early cohort. This framework is also general enough to apply to other customer cohorts and other types of bias. Our approach uses a combination of active and semi-supervised learning techniques in a self-learning scheme for effective data selection and augmentation. Our main contribution is the identification

	DE Phase I		FR Phase I		DE Phase II		FR Phase II	
Domain	ICER	SEMER	ICER	SEMER	ICER	SEMER	ICER	SEMER
Music	-1.48%	-0.99%	1.29%	1.36%	-2.74%	-1.82%	-0.59%	+2.13%
Global	+2.31%	+1.63%	-1.08%	-1.78%	-2.23%	-1.95%	0.96%	+0.15%
HomeAutomation	-0.61%	-0.20%	-1.13%	-0.4%	-0.79%	-1.27%	0.97%	+1.64%
Knowledge	-0.13%	-0.56%	-1.49%	-3.41%	-5.44%	-5.16%	2.89%	-1.99%
Notifications	-1.16%	-0.20%	-1.8%	-1.99%	-41.11%	-32.79%	-1.23%	-0.8%
Communication	+0.00%	-2.44%	-3.5%	-4.01%	-3.51%	-1.48%	0%	-3.69%
LocalSearch	+1.59%	+1.52%	1.46%	-1.65%	-1.17%	-0.71%	-0.35%	+0.69%
Help	-1.46%	-1.93%	-5.21%	-5.31%	+1.08%	-3.62%	-0.33%	+1.12%
Overall	**-0.12%**	**-0.40%**	**-1.43%**	**-2.15%**	**-6.99%**	**-6.10%**	**0.29%**	**-0.09%**

Table 1: ICER and SEMER relative changes (%) (negative shows an improvement, while positive indicates a degradation)

of the cohort representativeness where we use a combination of BERT topic embeddings with Autoencoder and density clustering to create a better representation of each cohort data and identify the contrastive area, where the new customers' data is missing. Furthermore, we applied SSL using a data retrieval module based on similarity search to augment the training data relevant to the early cohort. We compared a model that was trained on a random set of data with a model that was selected based on the active semi supervised learning approach. The proposed approach improves overall semantic and intent error rate for both German and French languages during offline testing.

Limitations

In this work, we have employed different strategies to identify the important utterances from early cohort. However, since a voice assistant system consists of many components, such as Wakeword, automatic speech recognition, NLU, dialogue manager, where errors occurring in one step might result to the final overall incorrect response. We have not discussed or considered the interaction among these components in this study. Last but not least, weak signal learning using users' feedbacks has shown to be beneficial in many studies, it is important to classify and identify the types of feedbacks that are relevant and those that are not relevant to NLU improvement (e.g., a negative feedback might not be caused by an immediate previous request, but be caused by other factors such as unsupported features, ASR incorrect recognition, device technical problems).

Ethics Statement

In the self learning phase, we have increased the representativeness of early cohort utterances in the training data. While it helps to improve the end to end experience of new customers, the method described in this work focuses on improving common customers and potentially introduces bias into the training data as well as the model. For example, minor customers that are not well represented in the live traffic will have lower chances of having their types of requests fulfilled through the active semi-supervised learning phased. Similarly, certain types of customers (e.g., those who use the device frequently) may have better chances of having correct NLU predictions overtime, while the system might still struggle dealing with rare requests in some specific domains.

References

Federico Bianchi, Silvia Terragni, Dirk Hovy, Debora Nozza, and Elisabetta Fersini. 2021. Cross-lingual contextualized topic models with zero-shot learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1676–1683, Online. Association for Computational Linguistics.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

O. Chapelle, B. Scholkopf, and A. Zien, Eds. 2009a. Semi-supervised learning (chapelle, o. et al., eds.; 2006) [book reviews]. *IEEE Transactions on Neural Networks*, 20(3):542–542.

Olivier Chapelle, Bernhard Scholkopf, and Alexander Zien. 2009b. Semi-supervised learning. *IEEE Transactions on Neural Networks*, 20(3):542–542.

Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien, editors. 2006. *Semi-Supervised Learning*. The MIT Press.

Eunah Cho, He Xie, and William M Campbell. 2019. Paraphrase generation for semi-supervised learning in nlu. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pages 45–54.

Kevin Clark, Minh-Thang Luong, Christopher D. Manning, and Quoc Le. 2018. Semi-supervised sequence modeling with cross-view training. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1914–1925, Brussels, Belgium. Association for Computational Linguistics.

Yifan Ding, Liqiang Wang, Deliang Fan, and Boqing Gong. 2018. A semi-supervised two-stage approach to learning from noisy labels.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, page 67–73, New York, NY, USA. Association for Computing Machinery.

Jacob Eisenstein, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. 2010. A latent variable model for geographic lexical variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1277–1287, Cambridge, MA. Association for Computational Linguistics.

Ismael Garrido-Muñoz, Arturo Montejo-Ráez, Fernando Martínez-Santiago, and L. Alfonso Ureña-López. 2021. A survey on bias in deep nlp. *Applied Sciences (Switzerland)*, 11.

Seraphina Goldfarb-Tarrant, Rebecca Marchant, Ricardo Muñoz Sanchez, Mugdha Pandya, and Adam Lopez. 2020. Intrinsic bias metrics do not correlate with application bias.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network.

Dirk Hovy and Anders Søgaard. 2015. Tagging performance correlates with author age. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 483–488, Beijing, China. Association for Computational Linguistics.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

Margaret L. Kern, Gregory J. Park, Johannes C. Eichstaedt, H. A. Schwartz, Maarten Sap, Laura K Smith, and Lyle H. Ungar. 2016. Gaining insights from social media language: Methodologies and challenges. *Psychological methods*, 21 4:507–525.

Sosuke Kobayashi. 2018. Contextual augmentation: Data augmentation by words with paradigmatic relations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 452–457.

Ashutosh Kumar, Satwik Bhattamishra, Manik Bhandari, and Partha Talukdar. 2019. Submodular optimization-based diverse paraphrasing and its effectiveness in data augmentation. 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 3609–3619.

Dieu-Thu Le, Jose Garrido Ramas, Yulia Grishina, and Kay Rottmann. 2022a. De-biasing training data distribution using targeted data enrichment techniques. In *KDD 2022 Workshop on Deep Learning Practice and Theory for High-Dimensional Sparse and Imbalanced Data (DLP)*.

Hieu Le, Dieu-Thu Le, Verena Weber, Chris Church, Melanie Bradford Kay Rottmann, and Peter Chin. 2022b. Semi-supervised adversarial text generation based on seq2seq models. In *EMNLP 2022*.

Yitong Li, Timothy Baldwin, and Trevor Cohn. 2018. Towards robust and privacy-preserving text representations.

Cheng-Yuan Liou, Wei-Chen Cheng, Jiun-Wei Liou, and Daw-Ran Liou. 2014. Autoencoder for words. *Neurocomputing*, 139:84–96.

Thomas Manzini, Yao Chong Lim, Yulia Tsvetkov, and Alan W Black. 2019. Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings.

David McClosky, Eugene Charniak, and Mark Johnson. 2006. Effective self-training for parsing. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 152–159, New York City, USA. Association for Computational Linguistics.

Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction.

Gaurav Sahu, Pau Rodriguez, Issam Laradji, Parmida Atighehchian, David Vazquez, and Dzmitry Bahdanau. 2022. Data augmentation for intent classification with off-the-shelf large language models. In

Proceedings of the 4th Workshop on NLP for Conversational AI, pages 47–57, Dublin, Ireland. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96.

Deven Santosh Shah, H. Andrew Schwartz, and Dirk Hovy. 2020. Predictive biases in natural language processing models: A conceptual framework and overview. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Zirui Wang, Adams Wei Yu, Orhan Firat, and Yuan Cao. 2021. Towards zero-label language learning.

Jieyu Zhao and Kai-Wei Chang. 2020. Logan: Local group bias detection by clustering.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods.

Xiaojin Zhu and Andrew B Goldberg. 2009. Introduction to semi-supervised learning. *Synthesis lectures on artificial intelligence and machine learning*, 3(1):1–130.

Xiaojin Jerry Zhu. 2005. Semi-supervised learning literature survey.

A Appendix

A.1 Algorithm for sampling strategies

Algorithm 1 Sampling strategies for early cohort

Input: current NLU model release F
set of recent live traffic data $\mathcal{U} = \{(x_i, y_i')\}_{i=1}^{|U|}$
K: the number of clusters used for diversity sampling strategy
Output: A boosting model F' built on top of F

1) Initialize $\mathcal{T} = \emptyset$
2) From a set of live traffic data $\mathcal{U}$, select utterances to be annotated with acquisition functions $a = \langle a^{(1)}, a^{(2)}, a^{(3)}, a^{(4)} \rangle$
3) **Uncertainty sampling**
Select $X^{(1)}$ using $a^{(1)}$ that selects utterances with (1) low confidence from early cohort that causes friction or (2) have different annotations while the utterance texts are the same.

$$a^{(1)}(x_i) = \begin{cases} 1, & (x_i, y_i') \in \mathsf{EARLY}, \mathsf{FRICTION}, \\ & (x_i, y_i') \in \mathsf{LOW} \\ & \text{or } \exists (x_k, y_k') \text{ where } x_i = x_k, y_i' \neq y_k' \\ 0, & otherwise \end{cases}$$

Send $X^{(1)}$ for human annotation to get $\mathcal{T}^{(1)} = \{(x_i, y_i) | x_i \in X^{(1)}\}_{i=1}^{|X^{(1)}|}$
4) **Diversity sampling**
Using k-means algorithm on a set of $\mathcal{U}$
YF = {(x_i, y_i') | (x_i, y_i') $\in$ EARLY, $(x_i, y_i') \in$ FRICTION$\}_{i=1}^{|\mathcal{U}}$
YF|K$is the number of clusters and \zeta^{(2)}$ is the utterance budget for annotation
Assign initial values for $\mathcal{E}(x_1), \mathcal{E}(x_2), ..., \mathcal{E}(x_{\mathcal{U}_{YF}})$
repeat
assign each item $\mathcal{E}(x_i)$ to the cluster with the closest centroid; calculate new centroid for each cluster
until converge
For each cluster, select $\zeta^{(2)}/K$ representative utterances to be added to $X^{(2)}$
Send $X^{(2)}$ for human annotation to get $\mathcal{T}^{(2)} = \{(x_i, y_i) | x_i \in X^{(2)}\}_{i=1}^{|X^{(2)}|}$
5) **Using customer feedback inputs**
Let $\mathcal{U}_{\mathsf{NF}} = \{(x_i, y_i') | (x_i, y_i') \in \mathsf{EARLY}, (x_i, y_i') \in \mathsf{PNF\_}i = 1^{|\mathcal{U}_{\mathsf{PNF}}|}$ where
$\mathsf{PNF} = \{(x_i, y_i') | (x_i^{next}, y_i'^{next}) \in \mathsf{EARLY}, (x_i^{next}, y_i'^{next}) \in \mathsf{NF}\}_{i=1}^{|\mathcal{U}_{\mathsf{NF}}|}$ defines a group of all previous utterances of those that are classified as containing negative feedbacks.
Send $X^{(3)}$ for human annotation to get $\mathcal{T}^{(\ni)} = \{(x_i, y_i) | x_i \in X^{(3)}\}_{i=1}^{|X^{(3)}|}$
6) **Using cohort representative data with density clustering**
Using BERT and LDA to define an embedding function of each utterances coming from both early and mature cohort
Use density clustering to define clusters where early cohort out-populates mature cohort in density to get $X^{(4)}$
Send $X^{(4)}$ for human annotation to get $\mathcal{T}^{(4)} = \{(x_i, y_i) | x_i \in X^{(4)}\}_{i=1}^{|X^{(4)}|}$
7) **Train a new model $\mathcal{F}'$ on $\mathcal{L}' = \mathcal{L} \cup \mathcal{T}$ on top of F**

The algorithm for sampling strategies in the first phase is given in Table 1, where the aim is to select

utterances with highest values for early cohort. The final set of the utterances is the union of the four sampling strategies.

A.2 Metrics

Semantic error rate (SEMER), is a metric used in offline evaluation where model prediction on domain/intent/slots is compared to human annotations. SEMER considers substitution error (S), insertion error (I), deletion error (D) at intent and slot level, and number of correct intents/slots classification (C) (see Equation 1).

$$\begin{aligned} SEMER &= \frac{\#errors}{\#referenceslots} \\ &= \frac{(S+I+D)}{(C+I+S+D)} \end{aligned} \tag{1}$$

ICER stands for Intent classification error, a metric calculated as the percentage of utterances containing an error intent classification error divided by total number of samples in this intent. BPS represent the percent increase/decrease from the current value. Friction refers to instances where a model does not understand the user or can not action the user's request.