Intent and Entity Detection with Data Augmentation for a Mental Health Virtual Assistant Chatbot

Ali Zamani, Matthew Reeson, Tyler Marshall, Mohamad Ali Gharaat, Alex Lambe Foster, Jasmine Noble, Osmar R. Zaiane
Department of Computing Science, University of Alberta, Canada
Alberta Machine Intelligence Institute
Mood Disorder Society of Canada

ABSTRACT

We report on implementing MIRA, a mental health resource chatbot to support healthcare workers in finding timely and relevant mental health resources. To generate appropriate queries to our carefully curated resource database, the chatbot must correctly identify the intents of an interlocutor and extract relevant entities from the conversation. With insufficient labelled examples, we employ data augmentation to generate training data automatically. Moreover, instead of detecting intent and extracting entities independently with two different classifiers, we integrate the two tasks by taking advantage of their interdependencies obtaining 99% accuracy for intent detection and 95.4% accuracy in entity extraction.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**; **Natural language generation**; • **Information systems** → *Information retrieval.*

KEYWORDS

Conversational agent, Intent detection, Data Augmentation

ACM Reference Format:
Ali Zamani, Matthew Reeson, Tyler Marshall, Mohamad Ali Gharaat, Alex Lambe Foster, Jasmine Noble, Osmar R. Zaiane. 2023. Intent and Entity Detection with Data Augmentation for a Mental Health Virtual Assistant Chatbot . In *Proceedings of ACM International Conference on Intelligent Virtual Agents (IVA '23).* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3570945.3607324

1 INTRODUCTION

Chatbots, sometimes referred to as conversational agents, are useful for communication with individuals with mental health concerns because they can provide destigmatized and anonymous conversations with users [1, 17, 21].

Chatbots often employ generative models trained on large datasets for response creation. These models utilize deep learning to synthesize outputs and often contain a dynamic lexicon of responses [2]. Implementing rule-based models structures, rather than generative models, may provide security by limiting the conversation to predefined responses; however, this comes at the expense of response creativity. Most models focus specifically on "entity extraction" or "intent detection" as a subset of general sentence modeling, which connects input text with learned special indicators that capture the general meaning of the input (i.e., the intent) in addition to any relevant words or terms (i.e., entities) [11].

Existing mental health chatbots applications have shown to be effective in reducing anxiety, depression, and stress [8, 14]. However, utilizing generative models and large datasets requires careful consideration as these models contain the inherent risk of inappropriate response generation. For instance, there have been documented instances of publicly available chatbots, such as ChatGPT, providing unreliable, inappropriate, or harmful information [5, 9].

One method to mitigate the risk of harmful or inappropriate response generation is to adopt a controlled data source for training purposes. This dataset would include diverse conversation flows and user statements to help ensure appropriate conversational responses. However, generation of controlled datasets requires tedious data curation and manual input of individual examples. Recently, work has been done investigating data augmentation techniques in the context of image classification and generation [19], and some research has been done investigating data augmentation techniques for boosting text classification performance [27]. Manually derived seed data can be expanded by employing a number of text data augmentation techniques that aim to reduce data imbalance or increase data diversity [7].

2 METHODS

This study evaluates the utilization of novel data augmentation techniques in developing a training dataset for a mental health system navigation chatbot. Our research team manually constructed an initial dataset then expanded this dataset using novel data augmentation techniques. The performance of each dataset was assessed via intent detection and entity extraction scores on multiple standard classification models, including Naive Bayes, logistic regression, Support Vector Machine (SVM), feed-forward neural network, recurrent neural network (RNN), and Long Short-Term Memory networks (LSTM). The performance of the initial and augmented datasets for each model was compared to determine the effectiveness of data augmentation in this context.

With the assistance of a multidisciplinary team including computing science, psychiatric experts, and people with lived experience, the initial training dataset was developed by creating a set of 91 unique intents according to the chatbot's anticipated and realized use cases. For example, if a client identified that they wanted a definition of a specific mental health condition, the intent

IVA '23, September 19–22, 2023, Würzburg, Germany

ACM ISBN 978-1-4503-9994-4/23/09.
https://doi.org/10.1145/3570945.3607324

"need_definition" would be detected by the model. One of the responsibilities of the chatbot is to direct individuals at risk of suicide or self-harm toward relevant emergency/crisis lines. As such, we created an intent "suicidal." Each intent requires a list of example strings or hypothetical sentences that inform the chatbot of variable inputs, which could indicate a particular intent.

For ethical purposes, it was essential to the research team that the training data was modelled in a controlled setting. Uncontrolled datasets utilize data from external sources (e.g., forums, conversion logs, blogs, etc.), which may contain misinformation or data that the chatbot can misconstrue (e.g., harmful biases) — potentially producing harmful speech or inaccurate information [25]. To develop the initial training dataset, researchers manually derived approximately 3-4 example strings for each of the 91 intents, creating 357 example strings (257, once 100 test samples had been held out). This initial dataset was used to train the initial version of the Intent classifier. However, as with most AI systems, it was determined that more example strings for each intent (i.e., more data) would enhance the chatbot's functionality. Acknowledging that the initial dataset was insufficient to train a chatbot properly, the research team employed several data augmentation techniques (described below) to enhance the initial dataset. This augmented dataset was used to train the second iteration of the chatbot and contained 2,292 example strings. In addition to sentence labeling, entities were annotated. In this step, important keywords that the chatbot should detect were annotated.

To increase the accuracy of intent detection and entity extraction, we expanded the dataset by employing some data augmentation techniques. The following data augmentation techniques were used:

Easy Data Augmentation (EDA): techniques proposed by [27] to boost text classification task performance. Synonym replacement, random insertion, random swap, and random deletion are four of the EDA's simple but powerful operations.

Synonym Replacement: This common type of data augmentation transforms text into paraphrases by swapping out specific terms with synonyms. The work by [15] introduces one of the earliest uses of this replacement in data augmentation. They used probable synonyms from WordNet to replace words [20].

Embedding Replacement: Comparable to synonym substitution, embedding replacement techniques look for words that best match the text's context while also maintaining the text's core ideas. To do this, words from the examples are translated into a latent representation space, where words from related contexts are placed closer together [12].

Replacement by Language Models: By anticipating subsequent or missing words based on the prior or surrounding context, language models represent language (classical and respectively masked language modelling) [3]. Language models provide a more localized replacement instead of embedding replacements by word embeddings that consider a global context [18].

With the labeled training data we can train a predictive model to detect intents from sentences and another predictive model to extract entities from the same sentences. To achieve this, we used and compared models using Naïve Bayesian, Logistic Regression, SVM and different neural network approaches. Neural Network approaches include a feed-forward neural network, a two-layer, fully connected neural network with 20 and 30 neurons in the first and second layers and an output layer; an RNN, an embedding layer followed by an RNN with a dense layer with 10 neurons and an output layer; and an LSTM, an embedding layer and an LSTM with the layer of a 10-dense neuron followed by an output layer. The feed-forward neural network, RNN, and LSTM are trained for 50 epochs.

In addition since intents and entities have inter-dependencies, we can take advantage of this interplay between them during detection and consider a simultaneous detection of the intents and entities. To do so we experiment with DIET. The Dual Intent and Entity Transformer (DIET) model [4] is a transformer based architecture which can outperform fine-tuned BERT and is six times faster to train. It has four main components: featurization, transformer, named entity recognition, and intent classification.

In the featurization component, the input text is transformed to a series of tokens. Token level one-hot encoding or multi-hot encoding can be used to get sparse features. Pre-trained word embedding like BERT [6], ConveRT [13], or GloVe [23] can be used to generate dense features. It adds a special classification token *_CLS_* to the end of each sentence which specifies the intent class of that sentence. In the transformer component, a two layer transform with relative position attention is used to retrieve context from the input. Features from the previous component are concatenated and passed to another fully connected layer with shared weights across all sequence steps. The named entity recognition component consumes the output of the transformer component and utilizes a Conditional Random Field (CRF) layer [16]. Finally, the intent classification feature uses transformer output for a *_CLS_* token to find the similarity between it and the intent embedding.

The DIET model was trained on the training data for 100 epochs. 28% of the data was selected randomly as test data (at least one example for each intent), and the remaining was used as training data. Also, because of class imbalance in both intent detection and entity extraction tasks, a macro-averaged F1 score was utilized. An acceptable model should perform well on intent detection and entity extraction tasks.

3 RESULTS

The intent detection F1, entity extraction F1 and the Average Intent and Entity (AIE) scores were the primary outcome variables used to determine intent detection and entity extraction in each dataset.

The AIE score, a macro-averaged F1-score of intent detection and entity extraction, is defined as follows:

$AIE = \frac{IntentDetectionF1 + EntityExtractionF1}{2}$

A higher F1 score is usually preferable, which can be achieved through increasing recall and precision [24, 28]. F1 scores for intent detection and entity extraction were calculated for each model with the original and augmented datasets. The AIE score combines these two F1 scores to give an overall performance metric for the model.

We observed a mean (SD) increase of 8.7% (6.8%) from pre-augmentation to post-augmentation on F1 intent detection scores across all models. The mean (SD) F1 intent detection score was 36.0% (4.48%) across all models trained on the original dataset, and the mean (SD) across all models trained on the augmented dataset was 44.7% (32.0%). The DIET architecture model improved the most of all models, with an F1 score increase of 23.1% from pre-augmentation (70.8%) to post-augmentation (97.2%).

We observed a mean (SD) increase of 12.6% (7.20%) from pre-augmentation to post-augmentation on entity F1 extraction scores across all models. The mean (SD) F1 entity extraction score of all models trained on the original dataset was 75.7% (4.48%), and the mean (SD) of all models trained on the augmented dataset was 88.4% (2.82%). The Diet Architecture model exhibited the most improvement, with an increase of 28.9% from pre-augmentation (65.6%) to post-augmentation (94.5%).

We observed a mean (SD) increase of 10.9% (7.0%) from pre to post-augmentation on AIE scores across all models. The mean (SD) AIE score was 55.8% (12.7%) across all models trained on the original dataset, and the mean (SD) across all models trained on the augmented dataset was 66.7% (17.2%). The DIET architecture model exhibited the most improvement in AIE scores, with an increase of 26.4% from pre-augmentation (70.8%) to post-augmentation (97.2%). Figure 1 shows all models' AIE scores from pre-augmentation to post-augmentation.

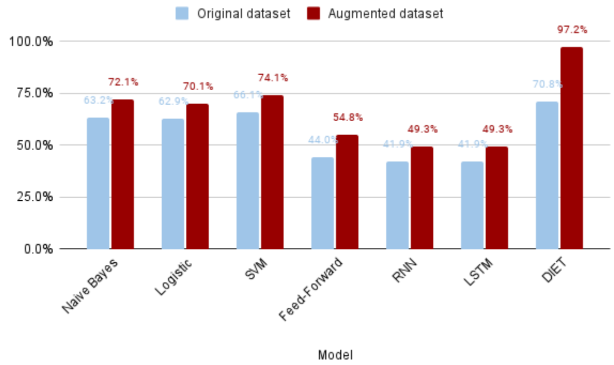

Figure 1: AIE Scores Pre/Post Data Augmentation

4 DISCUSSION

To our knowledge this is the first study to evaluate text-based data augmentation on baseline machine learning models for intent detection and entity extraction. We also assessed the performance of a Dual Intent and Entity Transformer (DIET) model for an artificially intelligent-based mental health chatbot.

Our data suggest that text-related data augmentation may improve the functionality of a mental health navigation chatbot without the use of large human annotated data. The performance of baseline machine learning models was evaluated before and after data augmentation. The results showed improved performance in all models.

Out of all the models tested, the DIET architecture achieved the highest level of accuracy on both intent detection and entity extraction and had the largest improvement in both metrics after augmentation. Utilizing a transformer-based architecture allowed the chatbot to detect intent and extract features simultaneously. Although all models improved in performance, the large improvement in AIE score in the DIET architecture after augmentation may suggest that transformer-based models benefit the most from the data augmentation techniques, particularly when the inter-dependencies between intent and entities is taken into account.

Data augmentation techniques have been utilized in computer vision and image processing [29, 30], but the field of text-based data augmentation is still in its infancy. The emergence of text-based augmentation provides the ability to build a dataset from controlled seed data that assists in avoiding the harms and risk of using LMMs [7, 22].

The findings of this study support the use of data augmentation for text-based datasets used to train chatbots in the context of mental health. This process, which relied on seed data developed by relevant experts, allowed for a cleaner and more reliable dataset than what would have been available through open-source LMMs. This is especially significant when considering response generation in chatbots designed for vulnerable populations. Employing these techniques can assist in developing safe and informed chatbots. To this end, data augmentation may provide a solution to the problem of misinformation and harmful speech without drastically compromising performance.

5 CONCLUSION AND FUTURE DIRECTION

The findings of this study supports that text augmentation, on the scale of an order of magnitude increase in synthetic data, can improve the predictive capabilities of models performing intent identification and entity extraction across various architectures, especially in mental health chatbots. Further, the DIET architecture performs exceptionally well relative to other tested architectures at that task, both with the pre-augmented and augmented datasets. Further, the outlined data augmentation techniques broadly apply and suggest a pathway for text augmentation across domains and tasks.

The potential for data augmentation to effectively generate synthetic training on a vast scale is a particularly interesting implication of this finding. We continued to see an increase in functionality without the augmented data overfitting the original training dataset. Recent speculation suggests that natural language datasets will run out of quality training data before 2026[26], and synthetic data will become increasingly important if that trend holds.

Further, these results show that datasets with relatively few samples may produce results comparable to that of relatively larger datasets when using these augmentation techniques. We see opportunities for organizations/individuals to create small datasets that, once augmented, can be used to fine-tune existing large language models to more narrow tasks, increasing the availability of machine learning beyond those with the resources to create expansive datasets; and still saving money for those that do.

Although the results indicate that these data augmentation techniques are effective, further investigation into text-related data augmentation in mental health chatbots is needed. It is also important to study the impact of effective data augmentation on the wider context of large language models and with broader kinds of datasets. Specific focus on ethical AI practices must consider the impact of maintaining smaller, controlled datasets that may prevent the generation of harmful speech or misinformation [10, 25].

REFERENCES

[1] Alaa A. Abd-Alrazaq, Mohannad Alajlani, Nashva Ali, Kerstin Denecke, Bridgette M. Bewick, and Mowafa Househ. 2021. Perceptions and Opinions of Patients about Mental Health Chatbots: Scoping Review. Issue 1. https://doi.org/10.2196/17828

[2] Eleni Adamopoulou and Lefteris Moussiades. 2020. An Overview of Chatbot Technology. In *Artificial Intelligence Applications and Innovations*, Ilias Maglogiannis, Lazaros Iliadis, and Elias Pimenidis (Eds.). Springer International Publishing, Cham, 373–383.

[3] Markus Bayer, Marc-André Kaufhold, and Christian Reuter. 2022. A Survey on Data Augmentation for Text Classification. *ACM Comput. Surv.* 55, 7, Article 146 (dec 2022), 39 pages. https://doi.org/10.1145/3544558

[4] Tanja Bunk, Daksh Varshneya, Vladimir Vlasov, and Alan Nichol. 2020. Diet: Lightweight language understanding for dialogue systems. *arXiv preprint arXiv:2004.09936* (2020).

[5] Ameet Deshpande, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, and Karthik Narasimhan. 2023. Toxicity in ChatGPT: Analyzing Persona-assigned Language Models. *arXiv preprint arXiv:2304.05335* (2023).

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[7] Steven Y Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A Survey of Data Augmentation Approaches for NLP. *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 968–988.

[8] Kathleen Kara Fitzpatrick, Alison Darcy, and Molly Vierhile. 2017. Delivering Cognitive Behavior Therapy to Young Adults With Symptoms of Depression and Anxiety Using a Fully Automated Conversational Agent (Woebot): A Randomized Controlled Trial. *JMIR Mental Health* 4 (2017), e19–e19. https://doi.org/10.2196/mental.7785

[9] Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models. *Findings of the Association for Computational Linguistics: EMNLP 2020*, 3356–3369. https://doi.org/10.18653/v1/2020.findings-emnlp.301

[10] Trystan S. Goetze and Darren Abramson. 2021. Bigger isn t better: The ethical and scientific vices of extra-large datasets in language models. *ACM International Conference Proceeding Series* (2021), 69–75. https://doi.org/10.1145/3462741.3466809

[11] Chih-Wen Goo, Guang Gao, Yun-Kai Hsu, Chih-Li Huo, Tsung-Chieh Chen, Keng-Wei Hsu, and Yun-Nung Chen. 2018. Slot-gated modeling for joint slot filling and intent prediction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. 753–757.

[12] Anne Harris, Stacy Holman Jones, Anne Harris, and Stacy Holman Jones. 2016. Words. *Writing for Performance* (2016), 19–35.

[13] Matthew Henderson, Iñigo Casanueva, Nikola Mrkšić, Pei-Hao Su, Tsung-Hsien Wen, and Ivan Vulić. 2019. ConveRT: Efficient and accurate conversational representations from transformers. *arXiv preprint arXiv:1911.03688* (2019).

[14] Jing Huang, Qi Li, Yuanyuan Xue, Taoran Cheng, Shuangqing Xu, Jia Jia, and Ling Feng. 2015. *TeenChat: A Chatterbot System for Sensing and Releasing Adolescents' Stress*. 133–145 pages. https://doi.org/10.1007/978-3-319-19156-0_14

[15] Oleksandr Kolomiyets, Steven Bethard, and Marie Francine Moens. 2011. Model-portability experiments for textual temporal analysis. *ACL-HLT 2011 - Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* 2 (2011), 271–276.

[16] John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. (2001).

[17] Yi-Chieh Lee, Yichao Cui, Jack Jamieson, Wayne Fu, and Naomi Yamashita. 2023. Exploring Effects of Chatbot-based Social Contact on Reducing Mental Illness Stigma. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–16.

[18] Vukosi Marivate and Tshephisho Sefara. 2020. Improving Short Text Classification Through Global Augmentation Methods. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 12279 LNCS (2020), 385–399. https://doi.org/10.1007/978-3-030-57321-8_21

[19] Agnieszka Mikołajczyk and Michał Grochowski. 2018. Data augmentation for improving deep learning in image classification problem. In *2018 International Interdisciplinary PhD Workshop (IIPhDW)*. 117–122. https://doi.org/10.1109/IIPHDW.2018.8388338

[20] George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. 1990. Introduction to wordnet: An on-line lexical database. *International Journal of Lexicography* 3 (1990), 235–244. Issue 4. https://doi.org/10.1093/ijl/3.4.235

[21] Inez Myin-Germeys. 2020. Digital technology in psychiatry: towards the implementation of a true person-centered care in psychiatry? , 401-402 pages. Issue 4. https://doi.org/10.1007/s00406-020-01130-1

[22] Curtis Northcutt, Lu Jiang, and Isaac Chuang. 2021. Confident learning: Estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research* 70 (2021), 1373–1411.

[23] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

[24] Aleksandr Perevalov, Daniil Kurushin, Rustam Faizrakhmanov, and Farida Khabibrakhmanova. 2019. Question Embeddings Based on Shannon Entropy Solving intent classification task in goal-oriented dialogue system. (2019), 73–78. Issue March.

[25] Nicole M. Thomasian, Carsten Eickhoff, and Eli Y. Adashi. 2021. Advancing health equity with artificial intelligence. *Journal of Public Health Policy* 42 (2021), 602–611. Issue 4. https://doi.org/10.1057/s41271-021-00319-5

[26] Pablo Villalobos, Jaime Sevilla, Lennart Heim, Tamay Besiroglu, Marius Hobbhahn, and Anson Ho. 2022. Will we run out of data? An analysis of the limits of scaling datasets in Machine Learning. (10 2022).

[27] Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference* (2019), 6382–6388. https://doi.org/10.18653/v1/d19-1670

[28] Zhen Xu, Chengjie Sun, Yinong Long, Bingquan Liu, Baoxun Wang, Mingjiang Wang, Min Zhang, and Xiaolong Wang. 2019. Dynamic Working Memory for Context-Aware Response Generation. *IEEE/ACM Transactions on Audio Speech and Language Processing* 27 (2019), 1419–1431. Issue 9. https://doi.org/10.1109/TASLP.2019.2915922

[29] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. 2019. CutMix: Regularization Strategy to Train Strong Classifiers With Localizable Features. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

[30] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. 2020. Random erasing data augmentation. *Proceedings of the AAAI conference on artificial intelligence* 34, 13001–13008. Issue 07.