Beyond the Data Scarcity Barrier: Unlocking Sentiment Analysis in Hebrew with Data Augmentation

Yonatan Amaru, 316365998, amaruy@post.bgu.ac.il
Amit Baras, 315726828, barasa@post.bgu.ac.il
Amit Finkman, 209122282, amitfi@post.bgu.ac.il,
Or Meiri, 315920462, meirio@post.bgu.ac.il,
Ido Bouhnik, 206586794, idoboh@post.bgu.ac.il,
Noam Munz, 207042292, munz@post.bgu.ac.il

March 5, 2024

Abstract

Data scarcity hinders robust sentiment analysis in under-resourced languages like Hebrew. This paper proposes HebAugment, a novel data augmentation methodology that integrates translation and generation models to address this challenge. HebAugment expands and diversifies the training dataset, improving sentiment classification performance. Our findings demonstrate that HebAugment, leveraging both translation and generation, significantly outperforms traditional approaches, even with limited real data. This research contributes a tailored data augmentation methodology for Hebrew sentiment analysis and valuable insights into tackling data scarcity in under-resourced languages.

Keywords— Sentiment Analysis, Data Augmentation, Generative Models, Hebrew

1 Introduction

Sentiment analysis, the automatic identification and classification of emotions within text, has become an essential tool across numerous domains, including social media monitoring, customer service optimization, and market research. By extracting insights into public opinion and user sentiment, it empowers businesses and organizations to make informed decisions, tailor their strategies, and improve their overall engagement.

However, the effectiveness of sentiment analysis often hinges on the availability of large, high-quality datasets. This poses a significant challenge for under-resourced languages like Hebrew, where data scarcity is a persistent obstacle. Building robust sentiment analysis models in Hebrew requires overcoming this data limitation to achieve accurate and reliable performance.

Existing research has explored diverse strategies to combat data scarcity in NLP. Some, like Karimi et al. (2021)[1], focused on text augmentation through punctuation, while others, like Edwards et al. (2022) [2] and Yoo et al. (2021) [3], investigated the potential of generative models for dataset augmentation in text classification and few-shot learning. Additionally, Wang et al. (2018) [4] Starc et al. (2017) [5] demonstrated the integration of text classification with generation, paving the way for NLP tasks like Natural Language Inference (NLI).

While these studies offer valuable insights, a specific gap remains in addressing data scarcity and improving sentiment analysis tailored for Hebrew. This gap is precisely where we focus our research: investigating the effectiveness of data augmentation techniques specifically designed to empower sentiment analysis in this under-resourced language.

Our research introduces HebAugment, a novel methodology that leverages cutting-edge data augmentation techniques. HebAugment integrates the power of state-of-the-art translation and generative language models to diversify and expand the training dataset for sentiment analysis in Hebrew. This comprehensive approach aims to tackle both the scarcity and quality concerns associated with data limitations in this language.

Through this research, we aim to:

1. Develop a groundbreaking data augmentation methodology specifically crafted for sentiment analysis in Hebrew.
2. Rigorously evaluate the effectiveness of various data augmentation techniques, including translation and generation, in boosting sentiment analysis performance.
3. Offer valuable insights into the challenges and opportunities associated with data augmentation for under-resourced languages.

We hypothesize that by incorporating translation and generation models into HebAugment's data augmentation pipeline, we can significantly enhance sentiment analysis performance in Hebrew, even when real data is limited, compared to traditional approaches.

This exploration delves into the world of sentiment analysis in Hebrew, acknowledging the limitations posed by data scarcity and outlining a novel approach - HebAugment - to overcome this challenge. By leveraging the power of data augmentation techniques, we aim to unlock understanding and analysis of sentiment within the realm of Hebrew language processing.

2 Background

This section provides a brief overview of Bidirectional Encoder Representations from Transformers (BERT), Generative Pre-trained Transformer (GPT), and their relevance to sentiment analysis, particularly in the context of data scarcity.

2.1 BERT for Sentiment Analysis

BERT, introduced by Devlin et al. (2018) [6], is a pre-trained language model based on the Transformer architecture. Its unique ability to process text bidirectionally and capture contextual information from both left and right surroundings has made it a powerful tool for various NLP tasks, including sentiment analysis. Studies by Zhang et al. (2023) [7] demonstrate BERT's effectiveness in understanding complex sentiment variations and achieving high accuracy in sentiment classification tasks.

However, as highlighted by Torge et al. (2023) [8], the success of BERT-based models heavily relies on access to large, high-quality datasets. This poses a significant challenge for under-resourced languages like Hebrew, where data scarcity is prevalent.

2.2 GPT for Data Augmentation and Transfer Learning

GPT, introduced by Radford et al. (2018) [9], utilizes a transformer-based architecture to generate human-quality text. This capability makes it valuable for data augmentation, a technique that addresses data scarcity by expanding training datasets with synthetically generated samples. Furthermore, GPT's proficiency in transfer learning, as shown by Yoo et al. (2021) [10], allows it to adapt its knowledge from a source language to a target language even with limited training data. This characteristic holds immense potential for enhancing NLP tasks in languages like Hebrew, where real data might be limited.

2.3 BERT for Machine Translation

While originally designed for natural language understanding, BERT's architecture has been successfully adapted for machine translation tasks. Wang et al. (2021) [11] showcase how BERT's bidirectional training enables it to capture complex linguistic relationships crucial for accurate translation. This opens up the possibility of leveraging translation models to augment sentiment analysis data in situations where resources in the target language (Hebrew in this case) are limited.

3 Related Work

Our research encompasses strategies for overcoming data scarcity, enhancing data through augmentation, and reviewing the current state-of-the-art in Hebrew natural language processing (NLP).

3.1 Approaches to Addressing Data Scarcity

Karimi et al.[1] introduced a method of text augmentation via punctuation, which, despite its effectiveness in improving model performance, falls short in addressing the complexities of advanced classification tasks. In contrast, our research leverages sophisticated generative models for more contextually relevant text augmentation. Edwards

et al.[2] and Yoo et al.[3] explored the use of generative models for dataset augmentation in few-shot learning and text classification, respectively. Our project extends these methodologies by employing a wider array of generative models, including newer versions like GPT-4, to offer a comprehensive augmentation strategy.

3.2 Integrated Approaches and Novel Applications

Wang et al.[4] and Starc et al.[5] demonstrated the integration of text classification with generation and the construction of datasets for Natural Language Inference (NLI), respectively. Our project draws inspiration from these approaches to enhance text classification through specialized models for both augmentation and classification, aiming for improved outcomes across broader NLP tasks.

3.3 Advancements in Hebrew NLP Models as Baselines for Fine-tuning

In the realm of Hebrew NLP, the development of models such as AlephBERT[12], HeBERT[13], and DictaBERT[14] represents significant progress. These models, each with its unique strengths, provide a solid foundation for our project's aim to fine-tune them for enhanced sentiment analysis. AlephBERT's adept handling of the intricacies of Hebrew syntax and grammar, HeBERT's specialization in sentiment analysis, and DictaBERT's broad training across diverse Hebrew texts, make them ideal baselines. Our project will leverage these models, fine-tuning them with advanced data augmentation techniques to address data scarcity and enhance model robustness, specifically for sentiment analysis in Hebrew.

3.4 Summary and Gap Analysis

This section has outlined the landscape of text classification, data augmentation, and the state-of-the-art in Hebrew NLP, identifying the pivotal role of models like AlephBERT, HeBERT, and DictaBERT. By fine-tuning these models, our project intends to tackle the challenges of data scarcity and model robustness in Hebrew sentiment analysis. Through this approach, we aim to not only build upon the existing advancements but also contribute to the field by optimizing these models for more accurate and nuanced sentiment analysis outcomes.

4 Methodology

Our proposed methodology, HebAugment, enhances sentiment analysis in Hebrew by leveraging advanced data augmentation techniques. We address Hebrew NLP's data scarcity challenge by integrating state-of-the-art translation and generative language models, thus diversifying and expanding the dataset for more effective sentiment classification.

4.1 Sentiment Analysis Using BERT

DictaBERT, a model chosen after comprehensive testing and comparison (details in the evaluation section), serves as our foundation for sentiment classification. The model is fine-tuned to classify tweets into positive (0), negative (1), or off-topic (2) sentiments.

We justify DictaBERT's selection based on its nuanced understanding of Hebrew and superior performance in preliminary tests. The fine-tuning process, including dataset size, epochs, and learning rate, will be elaborated upon in the evaluation section.

4.2 Translation Using BERT

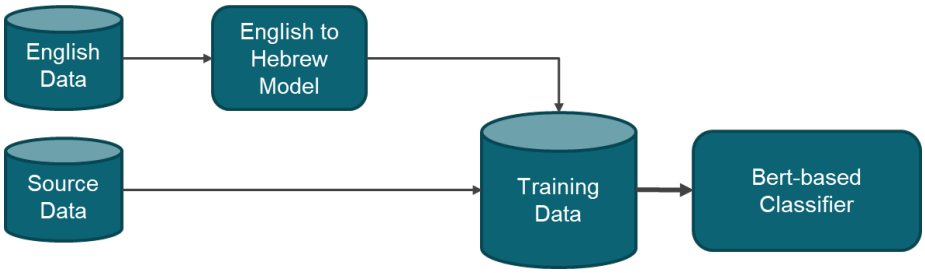

Figure 1: Schematic representation of the stand-alone translation pipeline.

In our study, we used the Helsinki-NLP [15] pre-trained BERT model for translating English to Hebrew. This model is part of the OPUS-MT project and uses neural machine translation (NMT) technology. While it generally performs well, it sometimes struggles with context and names[16]. To maintain the quality of our data, we exclude any tweets that are not accurately translated. This approach is similar to one of the successful models that was pre-trained on Spanish-English data with a bit of indigenous language data. This model performed very well, ranking first in 4 out of 11 language pairs. These results confirm the model's effectiveness across different languages [17], particularly in languages where there are not a lot of resources for training the translation model. Figure 1 reflects the standalone method of integrating the translation model to augment and increase the training set.

4.3 Generating Tweets with Generative Models

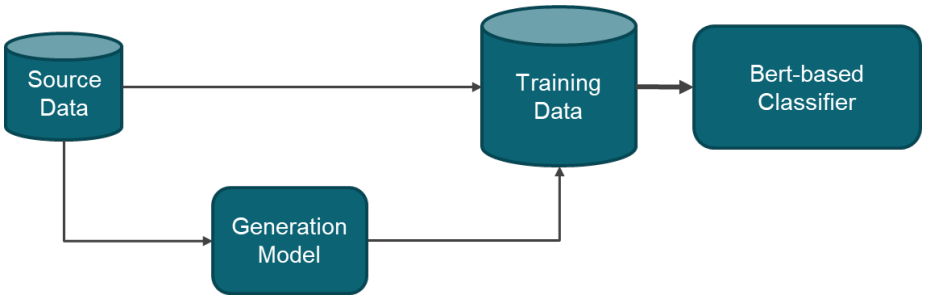

Figure 2: Schematic representation of the stand-alone Generation pipeline.

Another method to augment our dataset involves leveraging generative models, particularly OpenAI's ChatGPT [18].

Recognizing the potential of these models for synthesizing realistic tweets, we employed GPT-4, the most advanced iteration in the GPT series, for our research.
Through precise prompt engineering with GPT-4, we formulated prompts that accurately capture desired emotional tones or sentiments. This technique allowed us to utilize GPT-4's sophisticated language generation capabilities to create a diverse array of authentic, sentiment-specific tweets.
Employing this approach significantly enhanced the quality and sentiment alignment of the synthetic tweets in our dataset, providing a substantial boost to our data augmentation efforts. We can generate the tweets both in English and Hebrew. Figure 2 represents our method of augmenting the data by generating hebrew tweets and adding them to the training set.

4.4 Integrated Pipeline

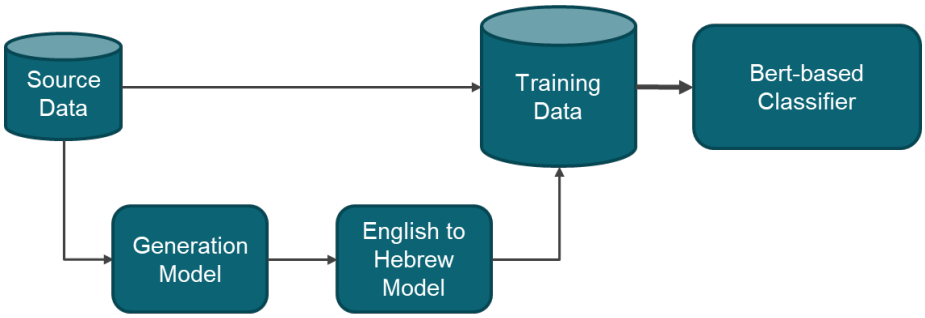

Figure 3: Schematic representation of the integrated pipeline.

Figure 3 represents the pipeline of the method combining both generation and translation. Step-by-step methodology:

1. Collection of an initial source Hebrew dataset for training.
2. Prompting based on the format of the source data.
3. Collection of generated english tweets
4. Translation of tweets using a pretrained translation model
5. concatenating the augmented tweets with the original tweets.
6. Training of the sentiment classification model on the augmented dataset.

This pipeline demonstrates a comprehensive approach to augmenting training data for sentiment analysis in Hebrew, addressing both quantity and quality concerns.

5 Experimental Setup

5.1 Data

Table 5.1 presents the different data sources we collected and prepared for our experimental setup. The baseline dataset, HebrewSentiment[19], will be split into a train set and a test set. All the other data sources will be used strictly for training the model.

Dataset	Num. Tweets	Positive	Negative	Off-Topic
HebrewSentiment[19]	12,804	8,512	3,922	370
Twitter [EN][20]	162,969	72,249	35,509	55,211
Translated (BERT)	4,330	1,765	1,617	948
Generated (GPT4)	3,000	1,794	873	333

Table 1: Datasets characteristics. We use a sample from each dataset in our research.

The HebrewSentiment dataset compromises 12,804 user comments, in Hebrew, from the official Facebook page of Israel's president. Each comment was manually tagged by a expert as either portraying a positive, negative or off-topic sentiment. For translation augmentation, we use a dataset of English tweets, the Twitter Sentiment Dataset[20]. This dataset, focused on tweets generated about the Indian prime minister, is Similarly labeled. We sourced both datasets with consideration to context, despite the distinct cultural and geographical settings. The Translated Dataset was sourced by translating the Twitter Sentiment Dataset using Bert[15]. Tweets in which the translation models failed to translate the sentence completely were removed. Finally the Generated dataset was generated using OPENAI's chatgpt (GPT4)[21]. In order to reduce the effect of imbalanced data, in our experiments we sampled based on the ratio of the HebrewSentiment Dataset.

5.2 Resources

We employ a combination of local and cloud-based computational resources for our experiments, including GPU-accelerated instances for model training and translation tasks. For generation tasks, we used OpenAI's GPT-4 API [21] for access to generative models that require significant resources.

Our experiments utilize the following open-source libraries and platforms:

- The Transformers library by Hugging Face for access to pre-trained models such as DictaBERT[14] and Helsinki-NLP[15] for translation tasks [**?**].
- The OpenAI API for generating synthetic data using GPT-4 [21].
- Scikit-learn for implementing statistical tests [22].

References to specific versions and configurations are included in our code repository [23].

5.3 Research Questions and Experimental Setup

Our evaluation is structured around key research questions, each designed to rigorously test the impact of data augmentation strategies on Hebrew sentiment analysis performance. We outline specific experimental setups to address these questions.

5.3.1 HebrewBERT: Baseline Model Performance

Question: Which baseline model demonstrates the best performance in Hebrew sentiment analysis with limited training data?

Setup: We compare baseline Hebrew models, 'alephBERT'[12], 'DictaBERT' [14], 'alephBERTgimmel' [12], and 'HeRoBERT' [13], using a dataset of 150 training samples and 1,000 test samples. This setup aims to identify the most effective model under data scarcity conditions. The key metric is accuracy for this test.

5.3.2 Effectiveness of the Augmentation

Question: Does Augmenting the training dataset using the proposed methods improve the performance of the classifier?

Setup: We augment the train dataset by adding tweets generated using either translation, generation or the integrated method. Model performance is evaluated as we incrementally increase the training set with the synthesized tweets, focusing on accuracy as the primary metric. This experiment will illustrate the benefits of the augmentations on model efficacy.

5.3.3 Computational Costs of Data Augmentation

Question: What are the computational costs associated with data generation and translation?

Setup: We document the computational time required for translating tweets and generating an equivalent number of synthetic tweets. This measurement will help discuss the practicality and efficiency of implementing translation and generative data augmentation techniques in real-world applications. Time is the primary metric here.

Testing Parameters Our goal is to test the affect of the translation and generation augmentation on the performance of the classification model. The complex nature of Deep Learning models can lead to a number of hyper-parameters, like learning rate or optimizer, affecting the model performance. To minimize the impact of these variables, we optimized the hyper-parameters on the baseline model and standardized them in the rest of our testing. The hyper-parameters we used are available in our code repository [23].

Statistical Validation To validate the impact of our proposed augmentation strategies on model performance we use cross validation and apply paired t-tests to compare the performance of models with and without augmentation, ensuring a significance level of $p \leq 0.05$.

Ablation Study An ablation study is conducted to isolate and understand the contribution of each augmentation strategy—translation and generation—towards the overall performance improvement. This study will compare models trained The original dataset, The dataset augmented with translated data, The dataset augmented with generated data and the dataset augmented with the integrated pipeline.

6 Results

6.1 HebrewBERT

Figure 4 compares the results of the 4 SOTA (State of the art) Hebrew BERT models. These models were fine-tuned using a sample of 1000 tweets. Both alephBERT[12] and

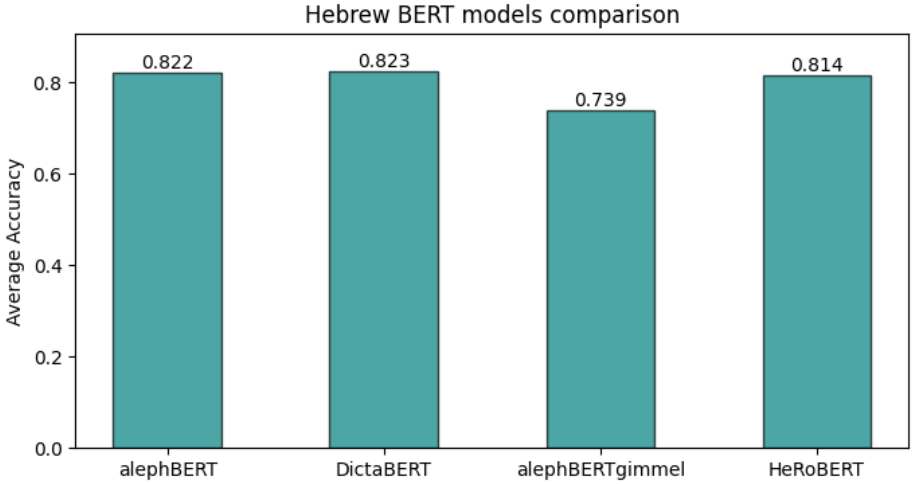

Figure 4: Accuracy of Hebrew Bert models, trained on a sample of 1000 tweets.

DictaBERT[14] achieved a accuracy of ≥ 0.82. We chose DictaBERT as our baseline model, under the assumption that its diverse training will help with its robustness to varying samples from the augmented dataset.

Augmentation	Train size	Accuracy	Training Time (m)
No Augmentation	50	0.51	0.386
Translation (Transformer)	3,050	0.74	7.783
Hebrew Generation (GPT4)	3,050	0.735	47.4
English Generation (GPT4)	3,050	0.73	27.17
Integrated (GPT4 + Transformer)	3,050	0.813	31.55
Combined	12,050	0.51	-

Table 2: Comparison of Model Accuracy and Training Times. *Training time includes data augmentation.*

6.2 Effectiveness

Figure 5 portrays the different methods accuracy with a incremental increase in the number of samples used for training, taken from the augmentation set we created for each method. All the augmentation techniques led to a significant improvement over the baseline of the original 50 training set. The three standalone methods we tested, translation of similar English data, generation in English and generation in Hebrew all achieved a accuracy of 0.73, and do not perform statistically different. The Integrated method, of generating English tweets and then translating them using Bert achieved significantly better accuracy of 0.813, coming close to using actual data.

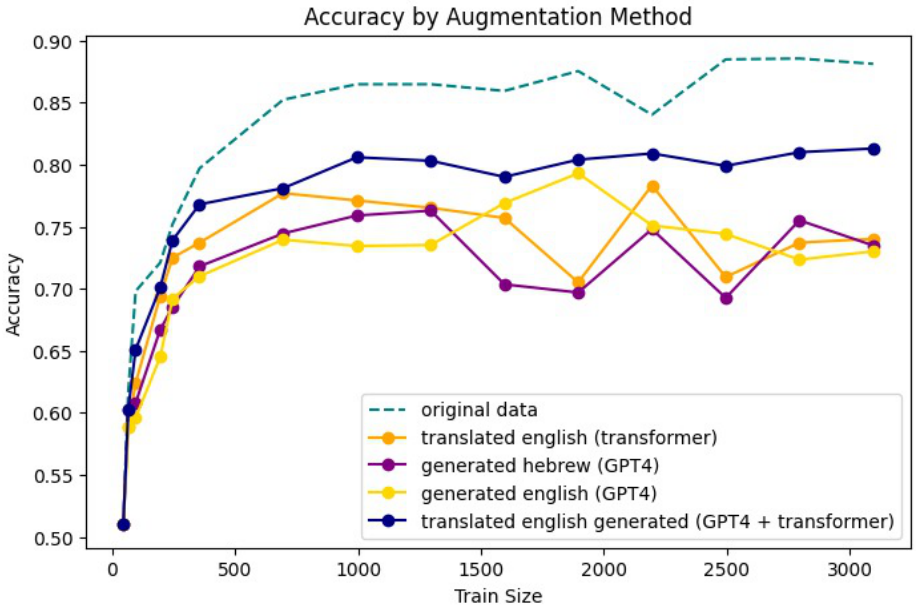

Figure 5: Accuracy of the different proposed augmentation methods, trained using a incrementally increasing sample, starting from 50 tweets and ending at 3000.

6.3 Training Time

Table 2 compares the Training set size, Achieved Model Accuracy and total Training time of the different configurations. The overhead for the training of the classification model is minimal, and most of the training time is spent on augmenting the data. The translation model, considerably lightweight and hosted on our servers, added a little over 7 minutes to the training time. Both English and Hebrew generation models, took a significant amount of time to train, hosted on OPEN-AI's API[21]. The Hebrew generation took almost double the time the English generation took, reflecting the complexity the model faces in Hebrew generation. The duration of the integrated method reflects the expected time of both models combined.

7 Conclusions and Future Work

7.1 Conclusions

Our research investigated the potential of data augmentation techniques to enhance sentiment analysis in Hebrew, specifically addressing the challenge of data scarcity faced by under-resourced languages. We explored the integration of translation and generation models to expand the training dataset and analyzed their impact on sentiment analysis performance.

Augmentation Effectiveness All data augmentation methods significantly improved sentiment analysis performance compared to the baseline model with limited data. The integrated approach of generating English tweets and translating them back to Hebrew achieved the highest accuracy (0.813), approaching the performance of a model trained with real data.

Insights from Challenges

- **Translation Complexity**: While translation using BERT improved accuracy, it highlighted challenges like vocabulary limitations, necessitating workarounds in the future.
- **Generative Model Performance**: Generation in English with GPT-4 yielded better results than Hebrew generation, suggesting potential differences in GPT-4's tokenization for these languages. Both standalone methods achieved similar accuracy, but English generation outperformed Hebrew generation when combined with translation.
- **Training Time Considerations**: Translation introduced minimal training time overhead, but generation, particularly in Hebrew, was computationally expensive.

7.2 Future Work

Our findings encourage further exploration in the following areas:

1. **Improved Hebrew Tokenization for LLMs**: We will investigate and develop improved tokenization methods specifically for Hebrew, tailored to large language models like GPT-4, to enhance their performance in Hebrew generation tasks.
2. **Evaluation on Diverse Use Cases**: We plan to test the effectiveness of our proposed augmentation techniques on a wider range of sentiment analysis tasks beyond the current scope, analyzing their performance across different use cases and datasets. This will provide a more comprehensive understanding of their generalizability.
3. **Modular Data Augmentation Pipeline**: We aim to develop a modular and reusable data augmentation pipeline that can be easily integrated into various NLP applications. This will allow users to customize the pipeline based on specific use cases and resource constraints, fostering broader adoption and adaptability.

By addressing these future directions, we believe our research can contribute significantly to advancements in sentiment analysis for under-resourced languages like Hebrew. This paves the way for more robust and accurate NLP applications in diverse real-world scenarios.

References

[1] A. Karimi, L. Rossi, and A. Prati, "AEDA: An easier data augmentation technique for text classification," in *Findings of the Association for Computational Linguistics: EMNLP 2021* (M.-F. Moens, X. Huang, L. Specia, and S. W.-t.

Yih, eds.), (Punta Cana, Dominican Republic), pp. 2748–2754, Association for Computational Linguistics, Nov. 2021.

[2] A. Edwards, A. Ushio, J. Camacho-collados, H. Ribaupierre, and A. Preece, "Guiding generative language models for data augmentation in few-shot text classification," in *Proceedings of the Fourth Workshop on Data Science with Human-in-the-Loop (Language Advances)* (E. Dragut, Y. Li, L. Popa, S. Vucetic, and S. Srivastava, eds.), (Abu Dhabi, United Arab Emirates (Hybrid)), pp. 51–63, Association for Computational Linguistics, Dec. 2022.

[3] K. M. Yoo, D. Park, J. Kang, S.-W. Lee, and W. Park, "Gpt3mix: Leveraging large-scale language models for text augmentation," *arXiv e-prints*, 2021.

[4] Z. Wang and Q. Wu, "An integrated deep generative model for text classification and generation," *Mathematical Problems in Engineering*, vol. 2018, 2018.

[5] J. Starc and D. Mladenić, "Constructing a natural language inference dataset using generative neural networks," *Computer Speech & Language*, vol. 46, pp. 94–112, 2017.

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[7] X. Zhang, Z. Wu, K. Liu, Z. Zhao, J. Wang, and C. Wu, "Text sentiment classification based on bert embedding and sliced multi-head self-attention bi-gru," *Sensors*, vol. 23, no. 3, p. 1481, 2023.

[8] S. Torge, A. Politov, C. Lehmann, B. Saffar, and Z. Tao, "Named entity recognition for low-resource languages-profiting from language families," in *Proceedings of the 9th Workshop on Slavic Natural Language Processing 2023 (SlavicNLP 2023)*, pp. 1–10, 2023.

[9] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, *et al.*, "Improving language understanding by generative pre-training," 2018.

[10] K. M. Yoo, D. Park, J. Kang, S.-W. Lee, and W. Park, "Gpt3mix: Leveraging large-scale language models for text augmentation," *arXiv preprint arXiv:2104.08826*, 2021.

[11] H. Wang, H. Wu, Z. He, L. Huang, and K. W. Church, "Progress in machine translation," *Engineering*, vol. 18, pp. 143–153, 2022.

[12] A. Seker, E. Bandel, D. Bareket, I. Brusilovsky, R. S. Greenfeld, and R. Tsarfaty, "Alephbert: A hebrew large pre-trained language model to start-off your hebrew nlp application with," *arXiv preprint arXiv:2104.04052*, 2021.

[13] A. Chriqui and I. Yahav, "Hebert and hebemo: A hebrew bert model and a tool for polarity analysis and emotion recognition," *INFORMS Journal on Data Science*, vol. 1, no. 1, pp. 81–95, 2022.

[14] S. Shmidman, A. Shmidman, and M. Koppel, "Dictabert: A state-of-the-art bert suite for modern hebrew," *arXiv preprint arXiv:2308.16687*, 2023.

[15] S. Itkonen, J. Tiedemann, and M. Creutz, "Helsinki-nlp at semeval-2022 task 2: A feature-based approach to multilingual idiomaticity detection," in *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pp. 122–134, 2022.

[16] J. Tiedemann and S. Thottingal, "Opus-mt–building open translation services for the world," in *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, European Association for Machine Translation, 2020.

[17] O. De Gibert, R. Vázquez, M. Aulamo, Y. Scherrer, S. Virpioja, and J. Tiedemann, "Four approaches to low-resource multilingual nmt: The helsinki submission to the americasnlp 2023 shared task," in *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pp. 177–191, 2023.

[18] B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, *et al.*, "Language models are few-shot learners," *arXiv preprint arXiv:2005.14165*, 2020.

[19] "Hebrew sentiment dataset." https://huggingface.co/datasets/hebrew_sentiment. Accessed: 2024-03-05.

[20] "Twitter sentiment dataset [english]." https://www.kaggle.com/datasets/saurabhshahane/twitter-sentiment-dataset. Accessed: 2024-03-05.

[21] "Open-ai api." https://platform.openai.com/docs/api-reference. Accessed: 2024-03-05.

[22] "Scikit-learn library." https://scikit-learn.org/. Accessed: 2024-03-05.

[23] "Hebaugment github repository." https://github.com/AmaruCrunch/HebAugment. Accessed: 2024-03-05.