2024

# Comparative Analysis of Data Augmentation on Sentiment Analysis in Three Distinct Languages

Hyesu Lee
*Minnesota State University, Mankato*

# Comparative Analysis of Data Augmentation on Sentiment Analysis in Three Distinct Languages

By

Hyesu Lee

A Thesis Submitted in Partial Fulfillment of the

Requirements for the Degree of

Masters

In

Data Science

Minnesota State University, Mankato

Mankato, Minnesota

May 2024

Date: 12-01-2023

Title: Comparative Analysis of Data Augmentation on Sentiment Analysis in Three Distinct Languages

Student’s Name: Hyesu Lee

This Design Project has been examined and approved by the following members of the student’s committee.

Dr. Suboh Alkhushayni

Dr. Naseef Mansoor

Dr. John Burke

# Acknowledgement

I would like to express my gratitude to my supervisor Dr. Alkhushayni for their invaluable guidance, continuous support, and remarkable patience throughout my master’s degree. Dr. Alkhushayni’s immense knowledge and plentiful experience inspired and guided me all the time of my academic research journey.

I also would like to express my thankfulness to the committee members, Dr. Naseef Mansoor, and Dr. John Burke for their precious feedback and guidance. Their insightful feedback and guidance enriched the quality of my thesis work.

I appreciate all the support I received from the rest of my family, especially my parents, Jong Il Lee, and Jong Rye Park, and my brother, Taesu Lee. Without their tremendous understanding and encouragement over the past few years, it would be impossible for me to complete my studies.

I am also grateful to my husband, Zhong Zhang, without whom this achievement would not have been possible. It is his unwavering assistance and support that have made both my academic pursuits and life in the United States a wonderful time.

I also appreciate all the support I received from the family, father-in-law, Qing Zhang, and mother-in-law, Wei Yan Zhong. Their consistent support and belief have been supported my academic journey and fostered an environment for success.

Lastly, I would like to express my gratefulness to all my friends, with a special mention to Esther Hiyoung Son, for the invaluable feedback and encouragement that kept me motivated to complete this thesis.

# Abstract

Machine learning in natural language processing analyzes datasets to make future predictions for various filed in the real world. By training machine algorithms on the datasets of text, the model can learn patterns and structure of the text in many different languages. Then the model enables to perform the text classification, sentiment analysis, and other tasks. A large and balanced dataset is required to develop an accurate machine learning model. However, the collection of a reliable, large, and equally distributed dataset is a challenging and requires significant resources and time. As a solution to this challenge, a data augmentation technique can be used to increase the size of a dataset by generating new data from the original dataset. This study investigates the impact of data augmentation on the performance of a machine learning models using small datasets in three diverse languages: French, German, and Japanese. After the data augmentation inflates the three diverse languages training datasets, three models are trained by each augmented training dataset. The three models' performance were compared with other three models' performance that are trained by each three original training datasets. This not only addresses the issue of a lack of large and balanced datasets but also the issue of dataset scarcity in various areas. Towards this, the generalization of each model trained by an augmented dataset is evaluated on each test dataset in different languages. A machine learning's capability of generalization can contribute situations where cross-lingual capabilities are needed, such as, international market research, multilingual customer support, obtaining cultural insights, etc. The models' performances and generalization are measured through evaluation metrics: accuracy, precision, recall, and f1-scores. Our results show that data augmentation improved the performance of the model's sentiment analysis with the languages French and Japanese. The results also showed that a model trained with a Japanese dataset showed improved performance in sentiment analysis when tested using German test data and vice versa. Similarly, a model trained with the German dataset showed marked improvement in its performance in sentimental analysis when tested with the French test dataset and vice versa.

Key words – Machine Learning, Natural Language Process, Text Classification, Sentiment Analysis, Data Augmentation, Translations

# Table of Contents

# List of Figures

# List of Tables

# Chapter 1

## Introduction

In the field of machine learning, the process of collecting a large and balanced dataset is crucial step for advancing to follow steps, the training and implementing an effective and successful machine learning model. The dataset plays an essential role to affect the model's performance and its ability to generalize. A dataset, especially that are small and have imbalanced distribution of data, can negatively influence the performance and reliability of a machine learning model by introducing a high risk of complex generalization challenges and unwanted bias toward specific classes within the models. However, due to the high cost of the time-consuming data collection process, getting a sizeable and balanced dataset for classification is difficult.

To address these challenges, data augmentation technique has been proposed as a solution. In data augmentation, the training data is artificially increased [1] or alternatively, the proportion of a minority classes in an imbalanced dataset is increased [2]. However, the data must be inflated with a reasonable modification. For instance, in an image-based dataset, if an image recognition model were to be trained with a dataset that only includes images of a human facing right, the model would only recognize humans who were facing right. To solve this issue, the dataset can be augmented to include several images of humans facing different directions by cropping and rotation of the existing dataset, and the model can then be trained by these new images to recognize the image of a human facing many directions, which is a reasonable way to inflate the dataset. For text-based datasets, data can be augmented in several methods, for example, changing the order of the words in the texts, replacing the words with synonyms, translating the texts, or summarizing the texts.

The importance of having a large training dataset cannot be underestimated in the field of machine learning. Using a larger training dataset can lead to a reliable machine learning model, which can make better predictions or classifications. On the other hand, using a small training dataset makes it difficult to train a reliable machine learning model [3], because small data sample are insufficient enough to represent all the possible data values for regression or classification tasks. Therefore, a decision made by a model trained by a smaller dataset are unreliable due to the significant uncertainties [4]. Moreover, by using the small dataset, there is a high risk of overfitting the model which in turn leads to poor generalization capabilities of a model. The overfitting machine learning will have high accuracy and better performance when it tested against the training data but will have a low accuracy against test dataset or new dataset [5]. Therefore, it is important to use the large training datasets to create reliable and accurate machine learning models to predict the future performances.

In a classification task, obtaining a balanced distribution in the dataset is vital to ensuring that each class has a roughly equal amount of data. This in turn leads to a model that is not biased toward any particular class. However, getting a balanced dataset can be challenging since it is difficult to find samples for specific classes, which can lead to imbalanced datasets.  Imbalanced datasets pose a challenge for classification algorithms since minority classes in imbalanced datasets tends to have higher misclassification costs than the majority classes [6]. This means that models trained using imbalanced dataset frequently show bias towards the majority classes while being unbiased towards the minority classes [7]. This results in a lower performing model.

Although finding a dataset that have large and equal distribution of different sentiment classes is a challenging problem in sentiment analysis, it is crucial for accurate and unbiased analysis. For example, in the business market, customer sentiment can significantly affect other customers' intentions [8]. Therefore, it is essential to understand the customer's opinion to develop the market. However, with either a small or unbalanced dataset, it can be difficult to understand the customer due to a lack of information. As a result, it is crucial to that a dataset used to train a sentiment analysis model is large and balanced [9].

Data augmentation can be used to enhance the quality of training datasets to address any problem caused by poor datasets and to improve the model's performance. For a small dataset, by expanding the small dataset to large dataset, the machine learning model can capture more patterns from the dataset and generalize its prediction on the new dataset more effectively. For an imbalanced dataset, by expanding the minority class with data augmentation, a balanced dataset can be created, and this balanced dataset can effectively solve the issue of misclassification bias toward the majority class [10]. Consequentially, there will be a positive effect on the text sentiment classification [2]. However, research on the impact of data augmentation to and its effectiveness on the machine learning model is still lacking and needs to be investigated.

In this study, the effect of data augmentation in natural language process is investigated to determine its effects on the performance of sentiment analysis models on three distinct languages datasets: French, German, and Japanese. Previous scholars [11] [2] [12] [13] in the data augmentation field primarily focused on sentiment analysis within single languages, mostly in English language, and overlooked sentiment analysis across several languages. Sentiment analysis across various languages should be considered since every language has its own structures, alphabet, grammar, and others.

Translation augmentation techniques are used to translate the data from original languages to a target language. Kryscinski et al. [14] used intermediate language datasets such as French, German, Chinese, Spanish, and Russian for a back translation augmentation technique in multilingual sentiment communication. This approach increased the diversity of the translation. Based on that, in this study, different intermediate languages are used to translate French, German, and Japanese dataset to a target language in translation augmentation process. Furthermore, two machine translations are used in translation: Google Translate and DeepL Translator APIs. The quality of these machine translation model varies depending on the translation algorithm for specific language pairs. This study examines the optimal machine translation system for each distinct languages during translation augmentation process.

Additionally, it is relatively easy to find text-based datasets in English since English is one of the most widely used languages. On the other hand, it can be difficult to obtain datasets in uncommonly used languages. Data augmentation on various languages is therefore crucial to solving this issue. Therefore, in this study, a model that is created based on one language will be tested on the test dataset of another language to examine the generalization capability of the model. This generalization is essential to explore a model’s adaptability, since it demonstrates that the patterns from the model trained on one language can be applied to different languages. In cases where data is limited for specific languages, a model trained on large and balanced language data can be used to test on small and unbalanced language datasets. As an example, it is relatively easy to find text-based datasets in English or Spanish since They are one of the most widely used languages across the world. On the other hand, it can be difficult to obtain datasets in uncommonly used languages. Through investigation on the generalization, a model created by commonly used languages can be used on the uncommonly used languages. In practical situations where required models work with multiple languages, improved generalization performance with data augmentation supports the development of cross-lingual applications. For example, in international business and market research, the generalization can lead to more accurate insights and better decision-making in a global context.

In the following section titled, ‘Related Work,’ a summary of the existing research on data augmentation is provided. It includes a discussion of the topic, key findings, methodologies, and any gaps in the field. Subsequently, in the ‘Methodology’ section, the dataset description of the data preparations and algorithms is described. The findings and evaluations in this study are then described in the ‘Result’ section. Lastly, the main conclusions and limitations are summarized to conclude this investigation, and suggestions are proposed for future research.

# Chapter 2

## Related Work

This section explores any key findings and related insights from the related research. Any existing knowledge of the data augmentation techniques and its performance in the natural language process are highlighted.

Data augmentation is a machine learning technique designed for the synthetic generation of training datasets, which means that the original training dataset is artificially inflated by reasonable modification with label-preserving transformation [15] [1]. Bottou et al. [16] developed the first application of the data augmentation technique on handwritten digits classification in 1998. This technique increased the accuracy of a model from 68% to 82%, and it was the first evidence that showed that a larger training dataset led to an improvement in the machine learning algorithms, and lead to better performance for a handwritten recognition system.

Following the research of data augmentation on handwritten digits classification, the data augmentation technique has been used in various areas, including image recognition, text mining, and numerous others. The data augmentation technique has been used successfully in image processing tasks for several years now [17] [18] [19], for object recognition, object detection, medical image analysis, and others. Conversely, data augmentation techniques in natural language process (NLP) have yet to be successfully adopted [15]. However, in recent years, studies on data augmentation techniques in natural language processes have increased. The use of the technique has increased for sentiment analysis models in market research, product development, political analysis, and other areas.

## 2.1 Sentiment Analysis

Sentiment analysis is a text and opinion mining technique which extracting information from the opinions, and expressions in text format [20]. It is used in various applications, such as market research, product development, political analysis, and diverse other area for understanding and evaluating the private states of individuals towards multiple aspects of the subject [21]. The main goal of the sentiment analysis is to predict sentiment polarity by analyzing the words in the sentences or documents [21].

Sentences or documents can be classified into two principal classes: subjective and objective. The subjective classes contain personal information, which include opinions, beliefs, judgments, and views about specific objects, and the objective classes contain factual information, including facts and evidence on particular objects [22]. Sentiment can be divided into two types: positive and negative emotions. In the text, positive emotion can be expressed as ‘happy,’ ‘joy,’ ‘beautiful,’ etc., while negative emotion can be expressed as ‘hate,’ ‘anger,’ ‘sad,’ and others [23]. However, in the real world, it is important to note that human expressions are complex, and it is challenging to categorize them as positive and negative sentiments. In recent years, several studies [20] [21] [24] [25] have studied sentiment analysis to extract valuable perceptions from human expressions and understand them from, for example, such as customer reviews.

In sentiment analysis, the dataset undergoes data preprocessing and data labeling steps before training and evaluating the sentiment analysis model. Data preprocessing involves three stages: removing any special characters or punctuations, removing null data, and transforming words to normal form or converting words to lowercase [23], depending on the language’s characteristics. These stages are essential steps in natural language processing, as they can affect the performance of the analysis. Data labeling is a crucial step in sentiment analysis to categorize the dataset into positive and negative polarities. Nguyen et al. [23] used a review dataset that was rated on a scale of 1 to 10, in which the reviews rated below five were labeled negative, and the reviews rated above five were labeled positive. After preprocessing, the dataset was prepared to train and evaluate the model.

Various machine learning models can be used to analyze sentiment. These include Decision Trees (DT), Naïve Bayes, Logistics Regression, Support Vector Machine (SVM), and Long Short-Term Memory Networks (LSTM). Gondhi et al. [26] used a Long Short Term Memory Network (LSTM) with the word2vec embedding technique to analyze product reviews on the Amazon shopping website, and LSTM showed f1-scores of 93%; as the dataset is unbalanced, an f1-score was used to measure the performance of the model. Another research, Nguyen et al. [23], explored four different machine learning models: Decision Tree (DT), Naïve Bayes, Logistic Regression, and Support Vector Machine (SVM) to analyze customer reviews on online food service. Compared to other models, SVM showed the highest accuracy of 91.5%, which was higher than Naïve Bayes by 9% [23]. Based on these previous research study, the SVM machine learning model was used in this research for the sentiment analysis.

Even with high accuracy in sentiment analysis, one of the challenges several researchers faced was a machine learning model's dependence on the dataset quality. The datasets were either an imbalanced dataset or a small dataset. Thus, a researcher had to spend most of the time collecting the dataset and find a technique to qualify the dataset. Few researchers have used data augmentation to address the limitations by improving the dataset quality.

## 2.2 Data Augmentation Techniques in Natural Language Process

In natural language processing, data augmentation has been a used to increase the size of training datasets or balanced the training dataset to address any problem caused by small or unbalanced datasets to improve the model's performance. Over the years, different ways of augmenting the dataset have been proposed: paraphrasing, transformation, and generation [27]. Paraphrasing and transformation techniques are commonly used data augmentation technique.

Transformation augmentation technique is a common data augmentation technique adopted by researchers in natural languages process. It transforms a sentence by simple substitution operation, synonym replacement is one of the approaches used to do so [27]. This approach generates new sentences or text by replacing words with their synonym [28]. In processing, the words are replaced with a randomly selected synonym from the list of synonyms, excluding non-stop words [29]. Feng et al. [28] proposed a tailor text augmentation algorithm to improve the synonym replacement augmentation technique. It consists of probabilistic synonym replacement and irrelevant words zero masking. The probabilistic synonym replacement replaces a word based on a probability that considers the relevance and discriminative of the word to emotion. Irrelevant words zero masking uses zero masking which replaces the irrelevant words with zero vector instead of synonyms. For instance, in a sentence of "the food is tasty," the word "food," which is an irrelevant word, will be replaced with a zero vector. The researchers' goal for these considerations was to generate an effective synthetic sentence and balance irrelevant words distributions. Another researcher, Peng et al. [2] proposed a new strategy in the synonym replacement augmentation technique. It is a balancing strategy based on word replacement for text sentiment classification. It involves two stages: oversampling stages and noise modification stages.

In oversampling stages, new samples are generated by replacing words of the representative data, which generate new samples in minority class by replacing the feature words from majority class with their antonyms and the feature words from minority class with their synonyms. And noise modification stage is to detect any noisy data whose sentiment polarity is incorrect and replace with correct words, relatively cleaning the dataset.

Paraphrasing technique is an augmentation technique to manipulate the sentence by rephrasing and translation is one of the approaches [27]. There are different ways of translation, for example, back-translation, forward translation, multilingual translation, random language translation, and others. Back translation is a commonly used translation augmentation technique [15] [30] [27]. This technique generates translations from the original language to another language and then back to the original language. In this study, instead of, back translation, a commonly used technique, another translation augmentation technique is used, which generates new data by translating from the original language to the target languages through several intermediate languages. While augmenting datasets with the translation technique, it is crucial to increase the dataset with the diversity of the translation, as an example, the text can be translated into several intermediate languages and back to the original language [14]. Referring to the latter, the texts are translated to the target languages through several different languages to increase the diversity of the datasets in this study.

Although there are several data augmentation techniques, they don’t always show positive effect on every dataset. Body et al. [11] investigated several data augmentation technique on different sample size datasets, and proved that as an original dataset size gets larger, the difference in total error rate between the augmented and original models decreased, which means that the error increases on the augmented data model as the original dataset size gets larger. This result showed that the data augmentation technique does not work on every dataset. Sometimes, it might give poor results on a large and balanced dataset model. Therefore, data augmentation techniques showed higher effect when the greater the scarcity of training dataset [15] [29].

## 2.3 Data Augmentation Technique in Sentiment Analysis

The data augmentation technique was adopted with low usage in natural language processing [15]. However, the method has recently been used in sentiment analysis that deals with small or unbalanced datasets [27] [11] [2] [12] [13] [28]. The increase in usage of the data augmentation indicates that the benefit of the data augmentation technique has been recognized in improving a machine learning model's performance and its contribution to dealing with the challenge of data sparsity. The research has been done through different data augmentation techniques used in the area of the natural language process.

EDA, BT, PREDATOR, and BART generally improved the machine learning model's performance with imbalanced and small datasets to predict the sentiments, especially in back translation boosted LSTM, GRU, CNN, RF, ERINIE, and BERT machine learning models for both imbalanced and small datasets [27]. Significantly, PREDATOR boosted the result of LSTM. The tailor text augmentation technique proposed by Feng et al. [28] was applied to COVID-19 sentiment analysis, and the technique showed effectiveness on the model's performance and generalization capability. In addition, the balancing strategy approach of the word replacement augmentation technique improved the performance of the model by 3% compared to a raw imbalanced dataset.

Abonizio et al. [27] explored several data augmentation techniques: easy data augmentation (EDA), back translation (BT), pretrained data augmenter (PREDATOR), and BART, to improve the sentiment analysis model with several machine learning models on two kinds of datasets: imbalanced and small datasets. These four techniques had different advantages and disadvantages. For advantages, PREDATOR and BART had a high possibility to improve the model's performances, EDA is not an expensive technique to use, and BT increase the dataset stably. However, EDA might drop the model's performance, and the other three techniques are expensive methods or require expensive technology. For instance, BT requires expensive translation APIs to translate the text with quality. Despite these disadvantages, these techniques are useful in improving the model's performances.

Enhancing the small or unbalanced dataset with data augmentation techniques improved the prediction accuracy of the models. However, it is essential to consider how and when to apply data augmentation. Applying data augmentation to every dataset may not improve the model's performance. Body et al. [11] used back-and-forth translation augmentation techniques to artificially increase sample sizes from varying numbers of original datasets to evaluate the performance between small and large datasets. Overall, the technique improved the movie review sentiment analysis performances.

## 2.4 Limitation: Uncertainty of Machine Translator on Translation Augmentation Technique

Although the translation technique showed significant improvement in the data augmentation model compared to the original datasets, previous researchers faced uncertainty of machine translation in the translating process. Machine Translation is a convenient tool to translate one language into another [15]. Two of the most used machine translators are Google Translate and DeepL Translate. Google Translate is one of the most used and easily accessed machine translations. It changed its translation algorithm in 2017 from a statistical machine translation algorithm to a neural machine translation algorithm [31]. DeepL Translate was launched in 2017 with neural machine translation, and DeepL Translate showed better performance of translation than the expectation [32]. The neural machine translation system excels in keeping the word order, accurately inserting function words, enhancing morphological agreement, making better lexical choices, and improving translation fluency [32].

Since Google Translate changed their translation algorithm, Google Translate performed 92% and 81% accurately in translating the discharge instructions into Spanish and Chinese for Spanish and Chinese-speaking patients; also, the new Google Translate algorithm had fewer harmful inaccuracies than an old algorithm [31]. DeepL Translate also showed effective performance in translation from German literature to English [33]. Google Translate and DeepL Translate APIs are effective, and they can translate not only short phrases but even large passages [33]. However, DeepL is a powerful translator in European languages [33]. In this study, both machine translations are used to assess the model's performance in each language. Although they may not translate the text with perfect grammar [33], they are capable of translating sentences reliably with high quality and safety [33] [34].

# Chapter 3
## Methodology

The methodology process demonstrates the progression from dataset preparation to the analysis and generalization of the models as shown in Figure 1.

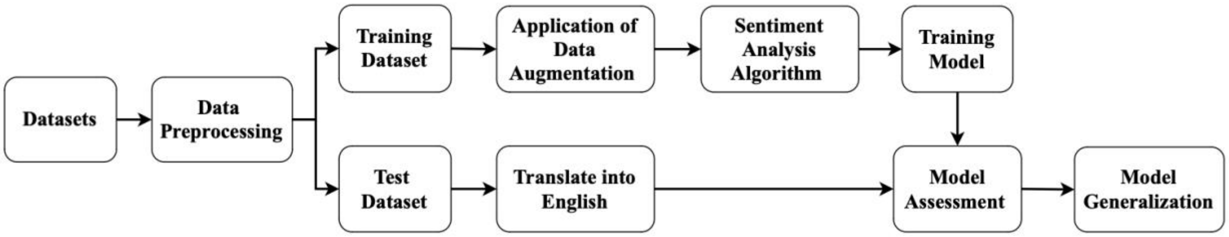


*Figure 1. Methodology Process*

After the dataset preparation, datasets underwent a preprocessing step, including data cleaning, feature selection, stop words removal, data labeling, and others. Following this process, the sentiment analysis model was trained using SVM machine learning model, and then its performance was measured using an evaluation matric. Lastly, the model's generalization capabilities were evaluated.

### 3.1 Dataset Description

In this study, online shopping review datasets were chosen to implement sentiment analysis models with analysis that deals. The datasets were collected from the Multilingual Amazon Reviews Corpus, Amazon AWS. They include Amazon product reviews from the US, Japan, Germany, France, Spain, and China in English, Japanese, German, French, Spanish, and Chinese languages between November 1, 2015, and November 1, 2019. The dataset includes several features: reviewer ID, review text, review title, rating from 1 to 5, product ID, and product category. There are numerous reviews in several product categories. For this study, reviews in the 'beauty' product category were selected. It is essential to choose the specific product category, as customer reviews can vary depending on the product type. For instance, in the technology category, a review can be "This is very fast and easy to set up," whereas, in the shoe category, a review can be "They are very comfortable when worn." Thus, focusing on the specific product category allows us to gain the unique aspects and expression from the customer reviews.

Several language datasets: French, German, and Japanese, were used to implement sentiment analysis model with data augmentation techniques. These languages were chosen due to their different language systems. French is from the Romance language family and Latin directly influenced it. And German is from Germanic language family, which includes English [6]. Unlike other language, Japanese is an isolated language that does not belong to any language family [6]. Japanese uses three writing scripts, Kanji, Hiragana, and Katakana, while German and French use the Latin alphabet like English [6]. These three languages use different word orders. Japanese uses subject-object-verb word order, and German and French uses subject-verb-object word orders more like English [6]. Furthermore, these languages have differences in various aspects such as cultures, nouns, grammar, etc.

All reviews were rated on a scale of 1 to 5. To see the distribution of positive and negative reviews, a binary variable was created. A review with a rating above 3 was grouped as a positive review, and review with a rating below 3 was grouped as a negative review. The rating of 3 is a neutral rate, and it was removed because the neutral point is neither good nor bad, it does not affect any positive or negative points.

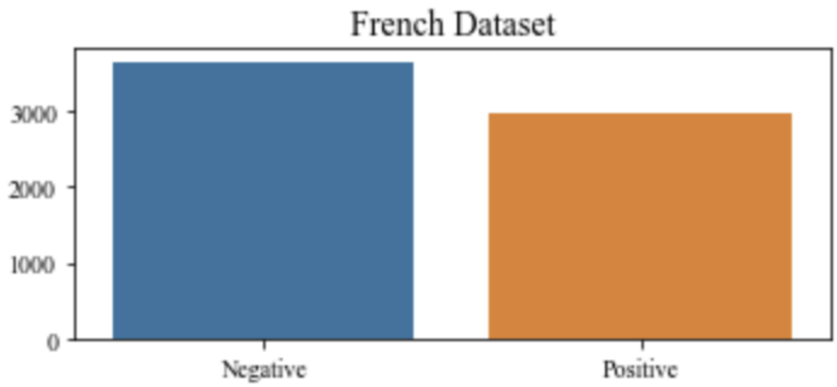


*Figure 2. Distribution of Negative and Positive Review in the French Datasets*

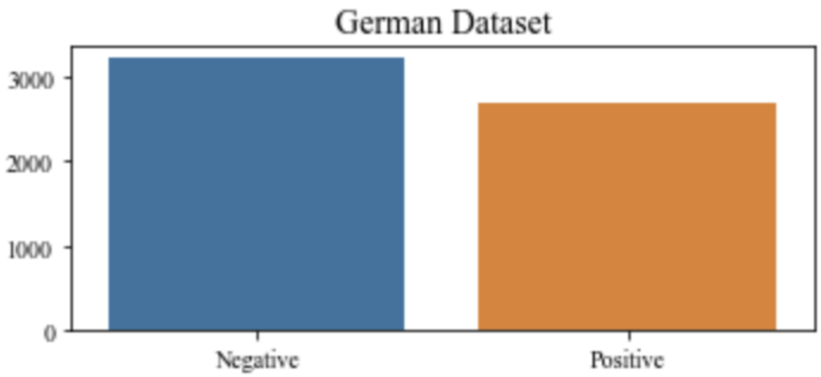


*Figure 3. Distribution of Negative and Positive Review in the German Datasets*

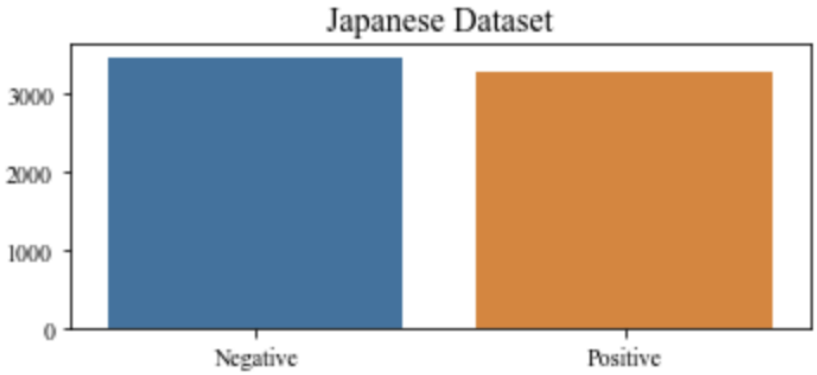


*Figure 4. Distribution of Negative and Positive Review in the Japanese Datasets*

Figure 2, Figure 3, and Figure 4 provides visual representation of the distribution of the positive and negative reviews within three datasets. Although there is a slight difference between positive and negatives distribution in three figures, they are in a balanced distribution condition.

*Table 1. The Number of Negative and Positive Reviews in the Three Datasets*

| Dataset | Number of Negative | Number of Positive | Total Number |
|---|---|---|---|
| French | 2,961 | 3,650 | 6,611 |
| German | 2,679 | 3,205 | 5,884 |
| Japanese | 3,281 | 3,462 | 6,743 |

To support this visual representation, **Error! Reference source not found.** provides the number of each positive and negative reviews in three datasets. Total number of reviews in each dataset are around six thousand. Since the datasets are small size, there will be a challenge for the sentiment analysis model to capture the critical characteristics from these datasets to make the optimal and accurate performance on the prediction.

## 3.2 Translation Augmentation Technique

In this study, a translation augmentation technique was used to increase the training dataset by translating it into another language. While inflating the dataset, it was essential to increase diversity of the translation. Thus, instead of translating French, German, and Japanese directly into English, they were translated through distinct intermediate languages into English, as shown in Figure 5.

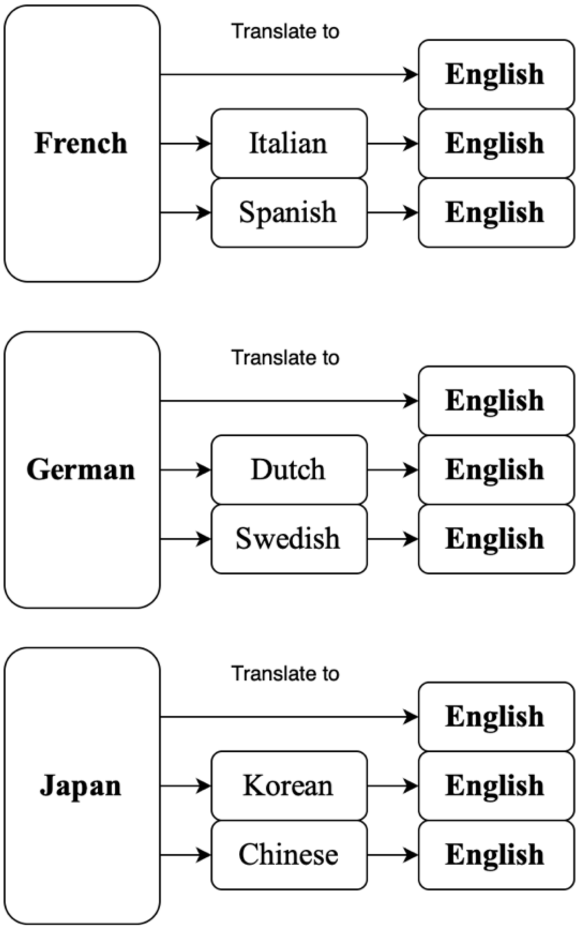


*Figure 5. Translation Process*

The selection of the intermediate languages was chosen based on the similarities in the language in terms of as grammar, alphabet, culture, and other aspects, between the original language and intermediate language. This approach generated diverse sentences in the target language. For instance, as shown in Figure 6, the original sentence was a Korean product review.

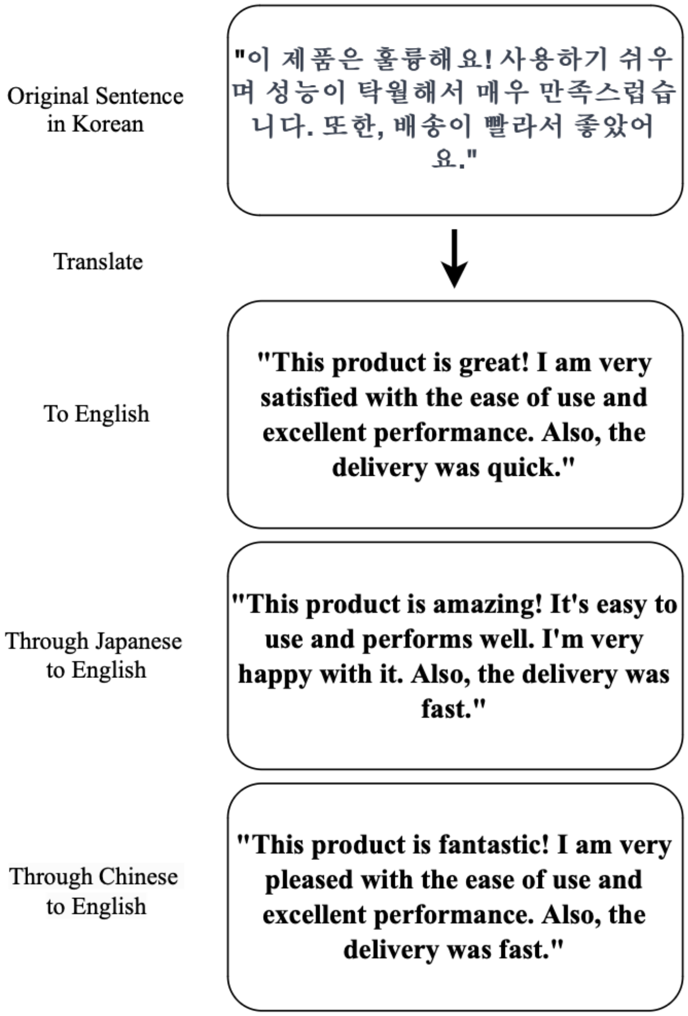

Original Sentence in Korean
"이 제품은 훌륭해요! 사용하기 쉬우며 성능이 탁월해서 매우 만족스럽습니다. 또한, 배송이 빨라서 좋았어요."
Translate
To English
"This product is great! I am very satisfied with the ease of use and excellent performance. Also, the delivery was quick."
Through Japanese to English
"This product is amazing! It's easy to use and performs well. I'm very happy with it. Also, the delivery was fast."
Through Chinese to English
"This product is fantastic! I am very pleased with the ease of use and excellent performance. Also, the delivery was fast."

*Figure 6. Translation Example*

The review passed the translation process through three distinct ways: a direct translation to English, a translation from Korean to Japanese and then to English, and translation from Korean to Chinese and then to English. These translation ways generated sentences that deliver the same meaning in English. However, the generated sentences have a different vocabulary and sentence structure, resulting in a diversity of languages expressions in target language.

For the translation process, Google Translate and DeepL Translation APIs machines were selected. These machine translations used the neural machine translation system which is a learning algorithm that reuses patterns stored in text corpora, resulting an exceptional performance in word representation and word prediction [32].

## 3.3 Implementation

After applying translation augmentation using Google Translate and DeepL Translate APIs, six augmented training datasets were created: three datasets using Google Translate API and three datasets using DeepL Translate API. These augmented training datasets were utilized to train the sentiment analysis model for predicting the polarity of the product reviews. For the machine learning algorithm, the Support Vector Machine, SVM, was chosen, which is commonly used in sentiment analysis. An evaluation was then conducted on the test dataset to assess the model’s performance by measuring accuracy, precision, recall, and f1-scores. Subsequently, the model’s generalization was tested on different test datasets. Figure 7 provides a visual representation of the assignment of specific models to the particular test datasets.

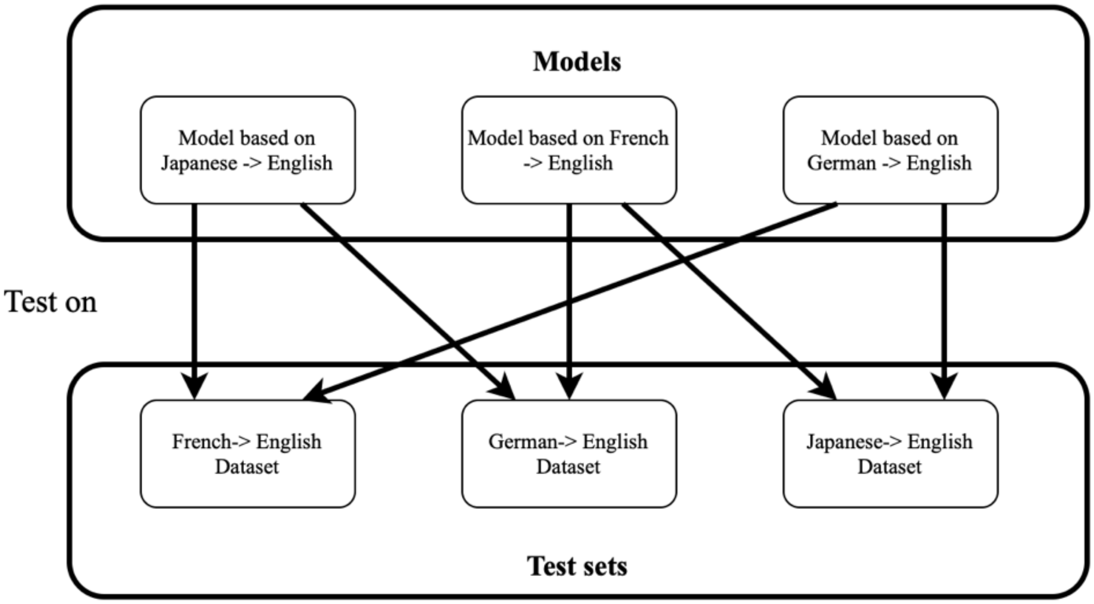

Models
Model based on Japanese -> English
Model based on French -> English
Model based on German -> English
Test on
French-> English Dataset
German-> English Dataset
Japanese-> English Dataset
Test sets

*Figure 7. Generalization Process*

The sentiment analysis model created using dataset in specific languages was tested on the test datasets in different languages to access the models' capacity for generalization. For example, a model trained on Japanese to English training datasets was evaluated against the test datasets in German to English and French to English. This generalization performance suggests potential application across multilingual contexts.

## 3.4 Evaluation

In the NLP field, the evaluation and comparison of machine learning models are essential to determine the effectiveness of the model and evaluation metric is commonly used, and it consists of accuracy, f1-score, precision, and recall [35] [12]. Accuracy provides the measurement of the correct classification ability of the model [12]. For each language, the performance of the machine learning model with data augmentation was assessed by the evaluation metrics. A confusion matrix is a table to measure the classification by predicted and actual classes, as shown in Table 2, and it is the basis for measuring accuracy, precision, recall, and f1-score.

*Table 2. Confusion Matrix*

| | Predicted: No | Predicted: Yes |
|---|---|---|
| Actual: No | True Negative (TN) | False Positive (FP) |
| Actual: Yes | False Negative (FN) | True Positive (TP) |

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$F - Score = \frac{2 * Precision * Recall}{Precision + Recall} \tag{4}$$

Accuracy in Equation 1 measures how often the classifier is correct overall. Precision in Equation 2 measures how often the classifier is accurate when it predicts ‘yes.’ Recall in Equation 3 measures how often the classifier predicts ‘yes’ when it’s actually ‘yes.’ The f1-score in Equation 4 measures the accuracy of a binary classification by using precision and recall values.

# Chapter 4

## Results

In this section, the findings of the study are presented from the investigation of the performance of the sentiment analysis with the translation augmentation across three distinct languages.

### 4.1 Performance of Data Augmentation Technique

The results of the performance evaluation of the sentiment analysis across datasets in three distinct languages, with original dataset, with only translation, and with the application of translation augmentation technique were shown in Table 3,

Table 4, and

Table 5.

*Table 3. Performance of Sentiment Analysis Model by Original Languages Dataset*

| Training Dataset | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|
| French | 87.66% | 0.88 | 0.88 | 0.88 |
| German | 79.59% | 0.8 | 0.8 | 0.79 |
| Japanese | 62.5% | 0.67 | 0.62 | 0.58 |

*Table 4. Performance of Sentiment Analysis Model with only Translation and without Data Augmentation Technique by Languages*

| Machine Translation | Training Dataset | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|---|

| | | | | | |
|---|---|---|---|---|---|
| Google Translation | French | 89.61% | 0.9 | 0.9 | 0.9 |
| | German | 82.31% | 0.83 | 0.82 | 0.82 |
| | Japan | 76.97% | 0.77 | 0.77 | 0.77 |
| DeepL Translation | French | 88.31% | 0.89 | 0.88 | 0.88 |
| | German | 78.91% | 0.8 | 0.79 | 0.79 |
| | Japan | 80.26% | 0.81 | 0.8 | 0.8 |

*Table 5. Performance of Sentiment Analysis Model with Translation Augmentation Technique by Languages*

| **Machine Translation** | **Training Dataset** | **Accuracy** | **Precision** | **Recall** | **F1-score** |
|---|---|---|---|---|---|
| Google Translation | French | 90.26% | 0.91 | 0.9 | 0.9 |
| | German | 80.95% | 0.81 | 0.81 | 0.81 |
| | Japan | 83.55% | 0.84 | 0.84 | 0.84 |
| DeepL Translation | French | 88.96% | 0.89 | 0.89 | 0.89 |
| | German | 74.83% | 0.75 | 0.75 | 0.75 |
| | Japan | 79.61% | 0.8 | 0.8 | 0.8 |

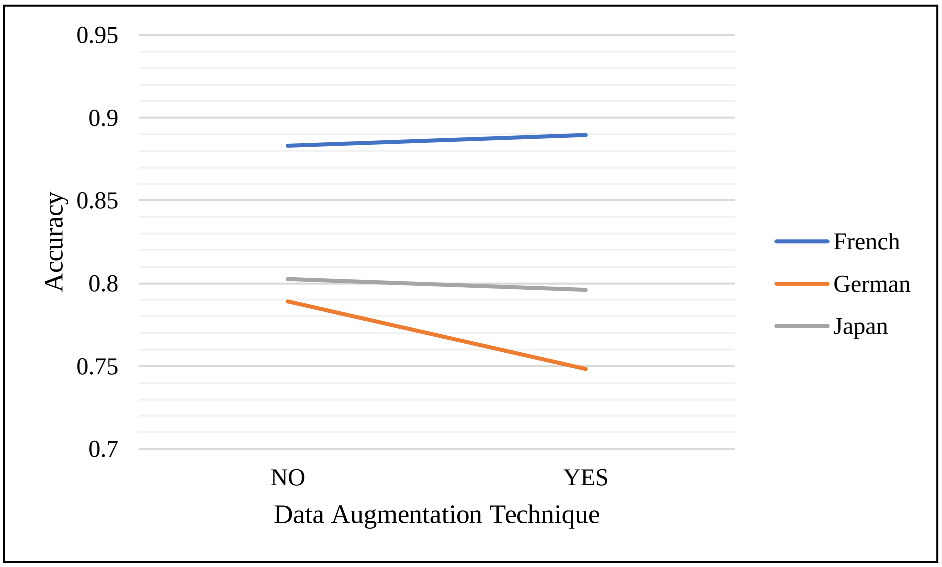


*Figure 8. Graph of Performance of Translation Augmentation with DeepL Translator on Sentiment Analysis*

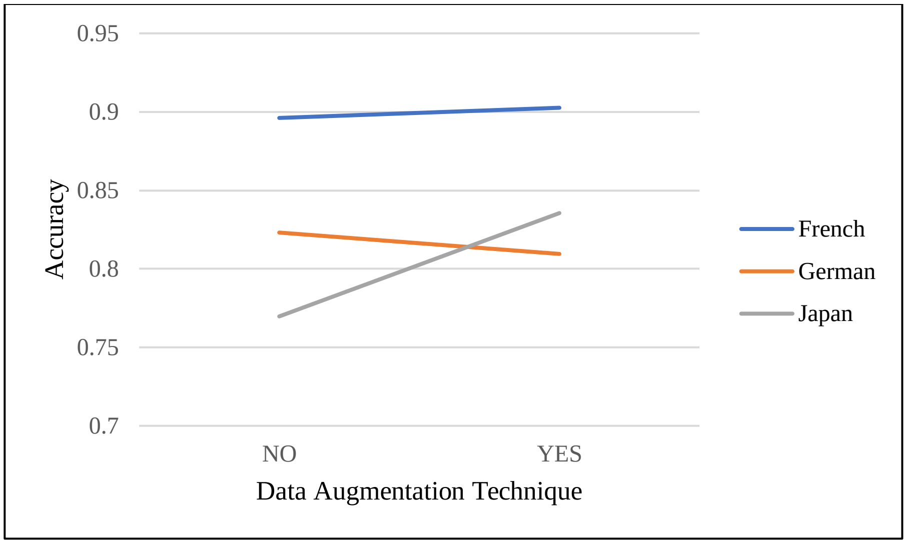


*Figure 9. Graph of Performance of Translation Augmentation with Google Translator on Sentiment Analysis*

As shown in  Figure 8 and Figure 9, the translation augmentation technique showed slight improvement on the model's performance for the sentiment analysis in ’French’ where both Google Translate and DeepL Translate APIs were used in the translation process. In ‘Japanese’ dataset, the translation augmentation technique showed significant improvement by increasing 6.58% where Google Translate API were used. However, with DeepL Translate API, there was no improvement.

The data augmentation technique did not improve the performance of model across all the languages. It also failed to improve the performance of the sentiment analysis using dataset in the German languages. This result is evident that the translation augmentation technique does not enhance the performance of the sentiment analysis across all languages in the world when dealing with relatively small datasets.

## 4.2 Generalization Performance

The result of the generalization performance with only translation and with the application of translation augmentation technique are shown in Table 6 and Table 7.

*Table 6. Performance of Generalization with only Translation and without Data Augmentation Technique by Languages*

| Machine Translation | Training Dataset | Test Dataset | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|---|---|
| Google Translation | French | German | 73.47% | 0.75 | 0.73 | 0.73 |
| | | Japanese | 77.63% | 0.78 | 0.78 | 0.78 |
| | German | French | 81.17% | 0.81 | 0.81 | 0.81 |
| | | Japanese | 73.03% | 0.73 | 0.73 | 0.73 |
| | Japanese | French | 76.62% | 0.77 | 0.77 | 0.77 |
| | | German | 72.11% | 0.73 | 0.72 | 0.72 |
| DeepL Translation | French | German | 74.83% | 0.76 | 0.75 | 0.74 |
| | | Japanese | 78.29% | 0.78 | 0.78 | 0.78 |
| | German | French | 81.82% | 0.82 | 0.82 | 0.82 |
| | | Japanese | 71.71% | 0.72 | 0.72 | 0.71 |
| | Japanese | French | 81.17% | 0.81 | 0.81 | 0.81 |
| | | German | 72.79% | 0.73 | 0.73 | 0.72 |

*Table 7. Performance of Generalization with Translation Augmentation Technique by Languages*

| Machine Translation | Training Dataset | Test Dataset | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|---|---|
| Google Translation | French | German | 77.55% | 0.79 | 0.78 | 0.77 |
| | | Japanese | 75.66% | 0.76 | 0.76 | 0.76 |
| | German | French | 85.71% | 0.86 | 0.86 | 0.86 |
| | | Japanese | 73.68% | 0.74 | 0.74 | 0.74 |

| | | | | | | |
|---|---|---|---|---|---|---|
| | Japanese | French | 75.32% | 0.75 | 0.75 | 0.75 |
| | | German | 75.51% | 0.77 | 0.76 | 0.75 |
| DeepL Translation | French | German | 74.15% | 0.75 | 0.74 | 0.74 |
| | | Japanese | 76.97% | 0.77 | 0.77 | 0.77 |
| | German | French | 79.87% | 0.8 | 0.8 | 0.8 |
| | | Japanese | 79.61% | 0.8 | 0.8 | 0.8 |
| | Japanese | French | 79.87% | 0.8 | 0.8 | 0.8 |
| | | German | 74.83% | 0.75 | 0.75 | 0.75 |

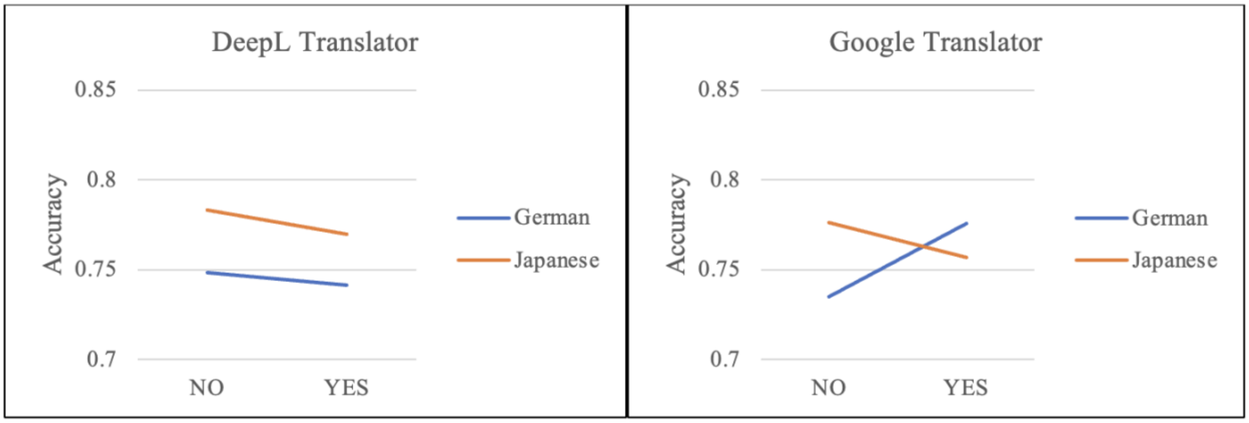


*Figure 10. Graph of Performance of Generalization of Translation Augmentation on French Training Dataset*

As shown in Figure 10, when the model is trained by French training dataset, there was no improvement of generalization on German and Japanese test datasets when DeepL Translator was used in translation augmentation process. However, there was improvement of generalization on German test dataset when Google Translator was used in the augmentation process. The accuracy was increase by 4.08% compared to the baseline.

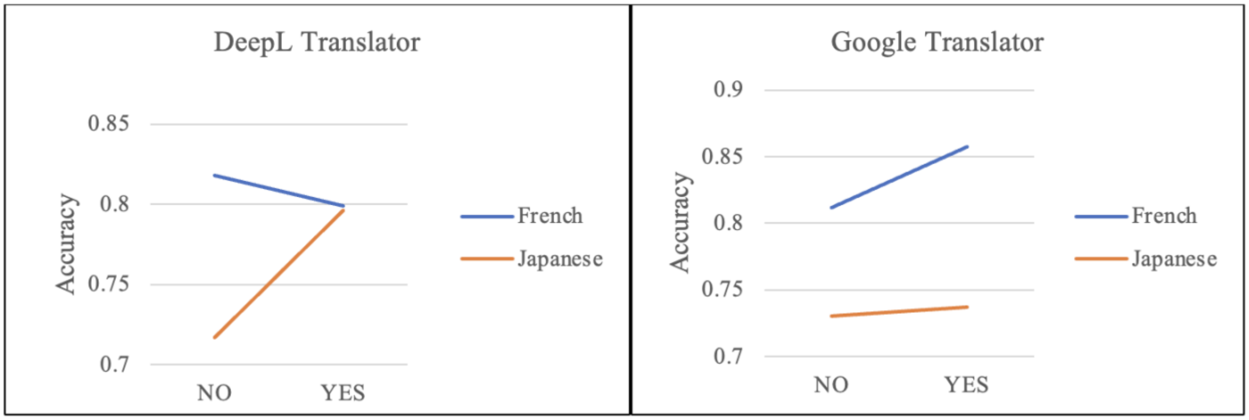


*Figure 11. Graph of Performance of Generalization of Translation Augmentation on German Training Dataset*

The performances of generalization of German training dataset on the French and Japanese test dataset were shown in Figure 11. There was generally improvement except one condition; generalization on French test dataset when DeepL Translator was used in the augmentation process. On Japanese test dataset, the accuracy was improved when both DeepL and Google Translators were used however, significantly increased by 7.9% when DeepL Translator was used. and on French test dataset, the accuracy was increased by 4.54% when only Google Translator was used.

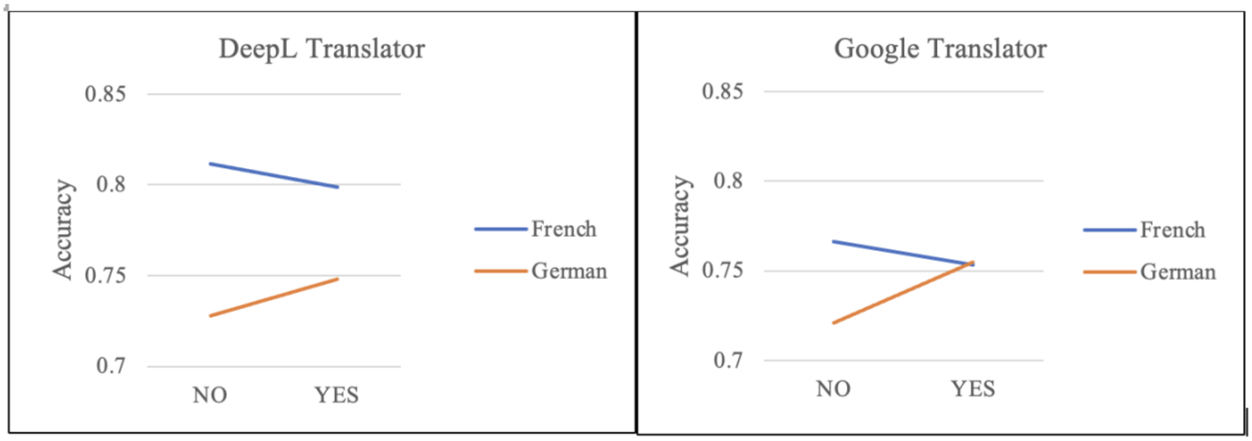


*Figure 12. Graph of Performance of Generalization of Translation Augmentation on Japanese Training Dataset*

As shown in Figure 12, the performances of generalization of Japanese training dataset on French and German test dataset were shown. The model trained by Japanese training dataset showed the improvement on German test dataset when both DeepL and Google Translators were used, but no improvement on French dataset.

The translation augmentation technique improved the generalization performance on 'Japanese' and 'German.' To specific, the models trained by augmented 'Japanese' and 'German' datasets showed better performance than non-augmented datasets when tested on another's test dataset when utilizing either the Google Translate API or DeepL Translate API. Furthermore, models trained by datasets in 'German' and 'French' showed the similar patterns, showing the improvement of the generalization of each model when tested each other's test dataset when utilizing only Google Translate API. These findings prove the importance of intersection between the languages in influencing the generalization capabilities of cross lingual sentiment analysis models.

# Chapter 5

## Conclusion

Sentiment analysis is one of the most complicated models in natural language processes when dealing with unbalanced dataset or limited dataset. To address this challenge, the data augmentation technique is introduced to increase the dataset. In this study, the effect of data augmentation technique in the sentiment analysis was investigated across three distinct languages: French, German, and Japanese. The sentiment analysis was conducted both with and without the translation augmentation technique, and the generalization was tested on test datasets in different languages.

Translation augmentation was utilized with two machine translations: Google Translate API and DeepL Translate API. The process of the translation augmentation is to inflate the training dataset by translating the dataset to the target languages through different intermediate languages depends on the original languages. The process of translating through different intermediate languages increased the diversity of the dataset.

By utilizing this technique, the performance of the sentiment analysis models across French and Japanese was improved effectively compared to the performance of the model with using augmentation technique. However, there was no improvement for German dataset. This suggests that the technique's enhanced model accuracy depends on the languages. Also, the reason that all the model’s performances had not been enhanced can be that dataset size is not small enough. According to [11] [15] [29], the data augmentation techniques shows the significant effect when the dataset is smaller. And the smallest dataset used by Body et al. [11] has the sample of 350 and showed highest improvement.

Then, the generalization of sentiment analysis models was investigated across distinct languages test datasets. The sentiment models with the data augmentation technique using both Google and DeepL Translate APIs on Japanese and German showed marked improvement in generalization on each other's test datasets. Additionally, the sentiment models trained by French and German showed improvement in generalization on each other's test dataset using Google Translate API. These results suggest the adaptability of sentiment analysis models trained on one language to others if both languages are translated into the same language. This means there is potential for cross-lingual. Furthermore, this adaptability can help to address the challenges of working with limited data for specific languages in the real-world scenario.

# Chapter 6

## Limitation and Future Work

This study has tested the effectiveness of data augmentation on the sentiment analysis of Amazon online shopping reviews and contributed to the field of data augmentation in natural language processing by offering insights into translation augmentation techniques across three distinct languages. While this research provided valuable insights, there were certain limitation that may affect the interpretation of the findings.

In the process of augmenting the training datasets through different intermediate languages, there were uncertainty regarding to the translated text whether the sentences, and words are translated differently. This leads to a limitation in this study as there is no guarantee that all the augmented training datasets consist of identical words, even they went through different intermediate language translation during data augmentation implementation. To address this limitation, future work could implement to check and verify the identical augmented dataset.

For more future research, selecting different translation augmentation techniques based on the dataset language's characteristics is encouraged as the translation augmentation did not improve the model's performance trained by 'German' in this study. Also, exploring more diverse language pairs should be done to see if there is any improvement in performance and generalization of different various language pairs.

Additionally, the selection of different machine learning models for the sentiment analysis and the exploration of more areas of the online shopping categories or in other fields, such as hotel reviews, are suggested to evaluate the impacts there. Additionally, this study used small and balanced datasets. However, the translation augmentation technique can be used for datasets that are unbalanced. For future research, using an unbalanced dataset is suggested to see whether the technique improves the performance of the sentiment analysis with such datasets.

## References

[1] L. Talyor and G. Nitschke, "Improving Deep Learning with Generic Data Augmentation," *2018 IEEE symposium series on computational intelligence (SSCI),* pp. 1542-1547, 2018.

[2] Z. Peng, H. Li, C. Wang, J. Shi and J. Zhou, " A two-stage balancing strategy based on data augmentation for imbalanced text sentiment classification," *Journal of intelligent & fuzzy systems,* vol. 40, pp. 10073-10086, 2021.

[3] J. Zhang and M. D. Shields, "The effect of prior probabilities on quantification and propagation of imprecise probabilities resulting from small datasets," *Computer methods in applied mechanics and engineering,* vol. 334, pp. 483-506, 2018.

[4] J. Zhang and M. D. Shields, " On the quantification and efficient propagation of imprecise probabilities resulting from small datasets," *Mechanical systems and signal processing,* vol. 98, pp. 465-483, 2018.

[5] J. Lever, M. Krzywinski and N. Altman, "Points of Significance: Model selection and overfitting," *Nature methods,* vol. 13, p. 703, 2016.

[6] P. Vuttipittayamongkol, E. Elyan and A. Petrovski, " On the class overlap problem in imbalanced data classification," *Knowledge-based systems,* vol. 212, p. 106631, 2021.

[7] S. Das, S. Datta and B. B. Chaudhuri, "Handling data irregularities in classification: Foundations, trends, and future challenges," *Pattern recognition,* vol. 81, pp. .674-693, 2018.

[8] J. Guo, X. Wang and Y. Wu, " Positive emotion bias: Role of emotional content from online customer reviews in purchase decisions," *Journal of retailing and consumer services,* vol. 52, p. 101891, 2020.

[9] B. (. Chae, " Insights from hashtag #supplychain and Twitter Analytics: Considering Twitter and Twitter data for supply chain practice and research," *International journal of production economics,* vol. 165, pp. 247-259, 2015.

[10] K. Sakai, K. Oishi, M. Miwa, H. Kumagai and H. Hirooka, "Behavior classification of goats using 9-axis multi sensors: The effect of imbalanced datasets on classification performance," *Computer and Electronics in Agriculture,* vol. 166, p. 105027, 2019.

[11] T. Body, X. Tao, Y. Li, L. Li and N. Zhong, "Using back-and-forth translation to create artificial augmented textual data for sentiment analysis models," *Expert systems with applications,* vol. 178, p. 115033, 2021.

[12] L. Wang, X. Xu, C. Lui and Z. Chen, " M-DA: A Multifeature Text Data-Augmentation Model for Improving Accuracy of Chinese Sentiment Analysis," *Scientific programming,* vol. 2022, pp. 1-13, 2022.

[13] M. Bayer, M.-A. Kaufhold, . B. Buchhold, M. Keller, J. Dallmeyer and C. Reuter, "Data augmentation in natural language processing: a novel text generation approach for long and short text classifiers," *International journal of machine learning and cybernetics,* vol. 14, pp. 135-150, 2023.

[14] W. Kryscinski, B. McCann, C. Xiong and R. Socher, "Evaluating the Factual Consistency of Abstractive Text Summarization," *Association for Computational Linguistics,* vol. EMNNP, p. 9332–9346, 202.

[15] L. F. A. O. Pellicer, T. M. Ferreira and A. H. R. Costa, " Data augmentation techniques in natural language processing," *Applied soft computing,* vol. 132, p. 109803, 2023.

[16] L. Bottou, Y. Lecun, Y. BENGIO and P. HAFFNER, "Gradient-Based Learning Applied to Document Recognition," *Proceedings of the IEEE,* vol. 86, pp. 2278-2324, 1998.

[17] . O. . O. Abayomi-Alli, R. Damaševičius, . S. Misra and R. Maskeliūnas, "Cassava disease recognition from low-quality images using enhanced data augmentation model and deep learning," *Expert Systems,* vol. 38, p. e12746, 2021.

[18] K. Maeda, S. Takada, T. Haruyama, R. Togo, T. Ogawa and M. Haseyama, " Distress Detection in Subway Tunnel Images via Data Augmentation Based on Selective Image Cropping and Patching," *Sensors (Basel, Switzerland),* vol. 22, p. 8932, 2022.

[19] L. Nanni, M. Paci, S. Brahnam and A. Lumini, "Comparison of Different Image Data Augmentation Approaches," *Journal of imaging,* vol. 7, p. 254, 2021.

[20] J. Sadhasivam and R. B. Kalivaradhan, "Sentiment Analysis of Amazon Products Using Ensemble Machine Learning Algorithm," *International Journal of Mathematical, Engineering and Management Sciences,* vol. 4, p. 508–520, 2019.

[21] T. T. Thet, J.-C. Na and C. S. Khoo, " Aspect-based sentiment analysis of movie reviews on discussion boards," *Journal of Information Science,* vol. 36, pp. 823-848, 2010.

[22] R. Feldman, " Techniques and applications for sentiment analysis," *Communications of the ACM,* vol. 56, pp. 82-89, 2013.

[23] B. Nguyen, V.-H. Nguyen and T. Ho, " Sentiment Analysis of Customer Feedback in Online Food Ordering Services," *Business Systems Research,* vol. 12, pp. 46-59, 2021.

[24] M. Al-Smadi, O. Oawasmeh, M. Al-Ayyoub, Y. Jararweh and B. Gupta, "Deep Recurrent neural network vs. support vector machine for aspect-based sentiment analysis of Arabic hotels' reviews," *Journal of computational scienc3,* vol. 27, pp. 386-393, 2018.

[25] Y. Wen, Y. Liang and X. Zhu, "Sentiment analysis of hotel online reviews using the BERT model and ERNIE model-Data from China," *PloS one,* vol. 18, pp. e0275382-e0275382, 2023.

[26] N. K. Gondhi, Chaahat, E. Sharma, A. H. Alharbi, R. Verma and M. A. Shah, "Efficient Long Short-Term Memory-Based Sentiment Analysis of E-Commerce Reviews," *Computational intelligence and neuroscience,* vol. 2022, pp. 1-9, 2022.

[27] H. Q. Abonizio, E. C. Paraiso and S. Barbon, "Toward Text Data Augmentation for Sentiment Analysis," *IEEE Transactions on Artificial Intelligence,* vol. 3, pp. 657-668, 2021.

[28] Z. Feng, H. Zhou, Z. Zhu and K. Mao, " Tailored text augmentation for sentiment analysis," *Expert systems with applications,* vol. 205, p. 117605, 2022.

[29] J. Wei and K. Zou, "EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks," *arXiv.org,* 2019.

[30] A. Sugiyama and N. Yoshinaga, "Data augmentation using back-translation for context-aware neural machine translation," *Association for Computational Linguistics,* vol. Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019), p. 35–44, 2019.

[31] E. C. Khoong, E. Steinbrook, C. Brown and A. Fernandez, " Assessing the Use of Google Translate for Spanish and Chinese Translations of Emergency Department Discharge Instructions," *JAMA internal medicine,* vol. 179, pp. 580-582, 2019.

[32] M.-J. Varela Salinas and R. Burbat, "Google Translate and DeepL: Breaking taboos in translator training: Observational study and analysis," *Ibérica (Castellón de la Plana, Spain),* pp. 243-266, 2023.

[33] S. Zulfiqar, M. Wahab, M. I. Sarwar and I. Lieberwirth, "Is Machine Translation a Reliable Tool for Reading German Scientific Databases and Research Articles?," *Journal of chemical information and modeling,* vol. 58, pp. 2214-2223, 2018.

[34] J. M. Miller, E. M. Harvey, S. Bedrick, P. Mohan and E. Calhoun, "Simple Patient Care Instructions Translate Best: Safety Guidelines for Physician Use of Google Translate," *Journal of clinical outcomes management,* vol. 25, 2018.

[35] F. Mehraliyav, I. C. C. Chan and A. P. Kirilenko, " Sentiment analysis in hospitality and tourism: a thematic and methodological review," *International journal of contemporary hospitality management,* vol. 34, pp. 46-77, 2022.

[36] B. R. Taira, V. Kreger, A. Orue and L. C. Diamond, "A Pragmatic Assessment of Google Translate for Emergency Department Instructions," *Journal of general internal medicine : JGIM*, vol. 36, pp. 3361-3365, 2021.

# Appendix A: Supplementary Tables

*Table A.1: The Number of Negative and Positive Reviews in the Three Datasets*

| Dataset | Number of Negative | Number of Positive | Total Number |
|---|---|---|---|
| French | 2,961 | 3,650 | 6,611 |
| German | 2,679 | 3,205 | 5,884 |
| Japanese | 3,281 | 3,462 | 6,743 |

*Table A.2: Performance of Original Sentiment Analysis Model by Languages*

| Languages that Model based on | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|
| French | 87.66% | 0.88 | 0.88 | 0.88 |
| German | 79.59% | 0.8 | 0.8 | 0.79 |
| Japanese | 62.5% | 0.67 | 0.62 | 0.58 |

*Table A.3: Performance of Sentiment Analysis Model with only Translation and without Data Augmentation Technique by Languages*

| Machine Translation | Languages that Model based on | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|---|
| Google Translation | French | 89.61% | 0.9 | 0.9 | 0.9 |
| | German | 82.31% | 0.83 | 0.82 | 0.82 |
| | Japan | 76.97% | 0.77 | 0.77 | 0.77 |
| DeepL Translation | French | 88.31% | 0.89 | 0.88 | 0.88 |
| | German | 78.91% | 0.8 | 0.79 | 0.79 |
| | Japan | 80.26% | 0.81 | 0.8 | 0.8 |

*Table A.4: Performance of Sentiment Analysis Model with Translation Augmentation Technique by Languages*

| Machine Translation | Languages that Model based on | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|---|
| Google Translation | French | 90.26% | 0.91 | 0.9 | 0.9 |
| | German | 80.95% | 0.81 | 0.81 | 0.81 |
| | Japan | 83.55% | 0.84 | 0.84 | 0.84 |
| DeepL Translation | French | 88.96% | 0.89 | 0.89 | 0.89 |
| | German | 74.83% | 0.75 | 0.75 | 0.75 |
| | Japan | 79.61 | 0.8 | 0.8 | 0.8 |

*Table A.5: Performance of Generalization with only Translation and without Data Augmentation Technique by Languages*

| Machine Translation | Languages that Model based on | Test Dataset | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|---|---|
| Google Translation | French | German | 73.47% | 0.75 | 0.73 | 0.73 |
| | | Japanese | 77.63% | 0.78 | 0.78 | 0.78 |
| | German | French | 81.17% | 0.81 | 0.81 | 0.81 |
| | | Japanese | 73.03% | 0.73 | 0.73 | 0.73 |
| | Japanese | French | 76.62% | 0.77 | 0.77 | 0.77 |
| | | German | 72.11% | 0.73 | 0.72 | 0.72 |
| DeepL Translation | French | German | 74.83% | 0.76 | 0.75 | 0.74 |
| | | Japanese | 78.29% | 0.78 | 0.78 | 0.78 |
| | German | French | 81.82% | 0.82 | 0.82 | 0.82 |
| | | Japanese | 71.71% | 0.72 | 0.72 | 0.71 |
| | Japanese | French | 81.17% | 0.81 | 0.81 | 0.81 |
| | | German | 72.79% | 0.73 | 0.73 | 0.72 |

*Table A.6: Performance of Generalization with Translation Augmentation Technique by Languages*

| Machine Translation | Languages that Model based on | Test Dataset | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|---|---|
| Google Translation | French | German | 77.55% | 0.79 | 0.78 | 0.77 |
| | | Japanese | 75.66% | 0.76 | 0.76 | 0.76 |
| | German | French | 85.71% | 0.86 | 0.86 | 0.86 |
| | | Japanese | 73.68% | 0.74 | 0.74 | 0.74 |
| | Japanese | French | 75.32% | 0.75 | 0.75 | 0.75 |
| | | German | 75.51% | 0.77 | 0.76 | 0.75 |
| DeepL Translation | French | German | 74.15% | 0.75 | 0.74 | 0.74 |
| | | Japanese | 76.97% | 0.77 | 0.77 | 0.77 |
| | German | French | 79.87% | 0.8 | 0.8 | 0.8 |
| | | Japanese | 79.61% | 0.8 | 0.8 | 0.8 |
| | Japanese | French | 79.87% | 0.8 | 0.8 | 0.8 |
| | | German | 74.83% | 0.75 | 0.75 | 0.75 |

## Appendix B: Supplementary Figures

*Figure B.1: Methodology Process*

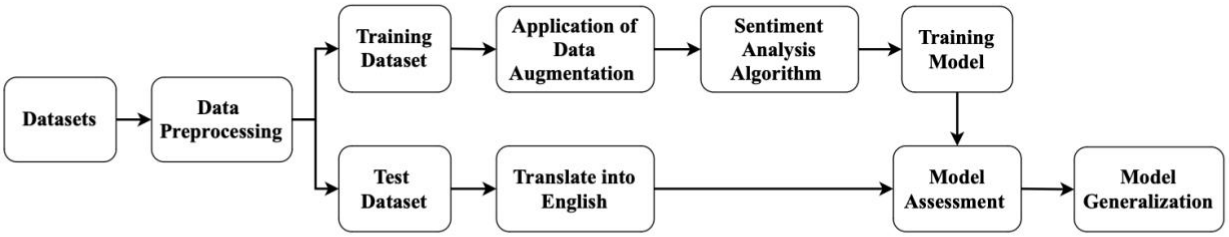


*Figure B.2: Distribution of Negative and Positive Review in the French Datasets*

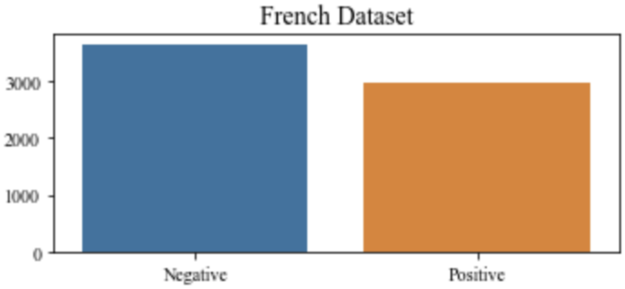


*Figure B.3: Distribution of Negative and Positive Review in the German Datasets*

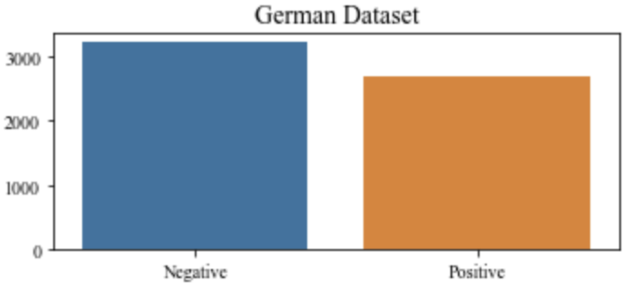


*Figure B.4: Distribution of Negative and Positive Review in the Japanese Datasets*

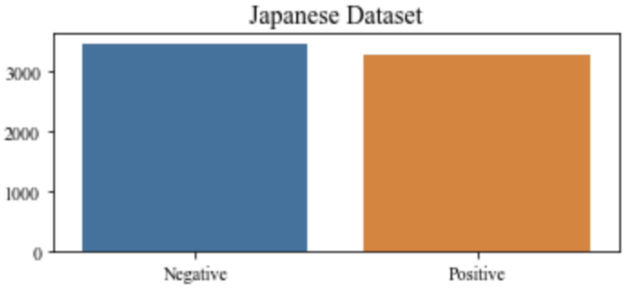


*Figure B.5: Translation Process*

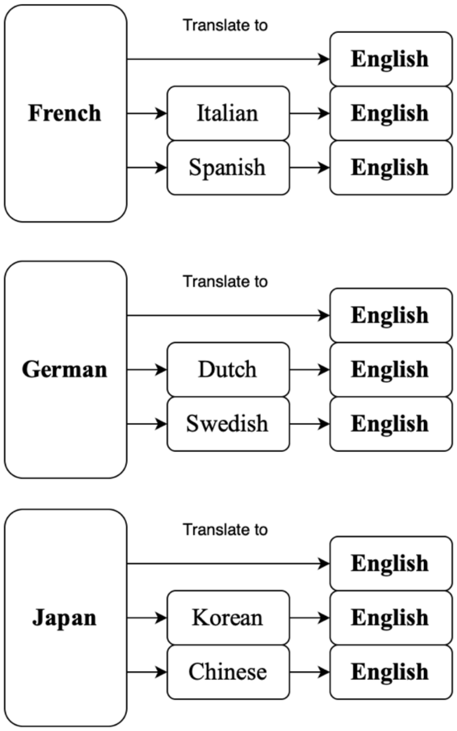
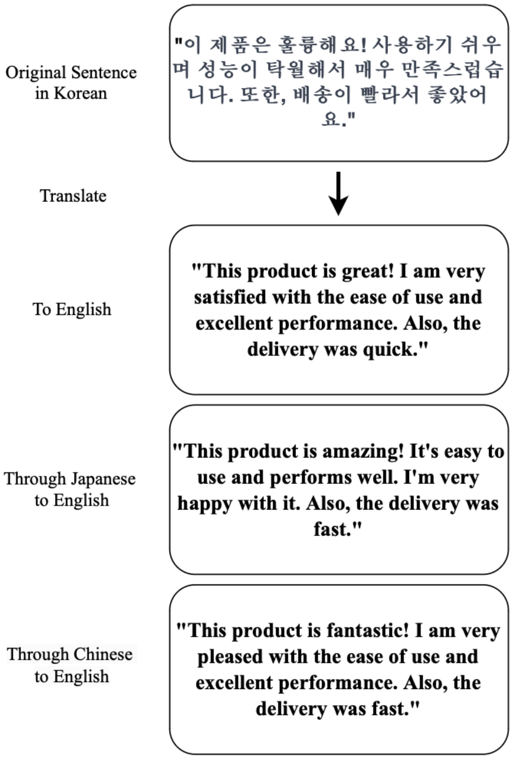


*Figure B.6: Translation Process Example*



Original Sentence in Korean

"이 제품은 훌륭해요! 사용하기 쉬우며 성능이 탁월해서 매우 만족스럽습니다. 또한, 배송이 빨라서 좋았어요."

Translate

To English

**"This product is great! I am very satisfied with the ease of use and excellent performance. Also, the delivery was quick."**

Through Japanese to English

**"This product is amazing! It's easy to use and performs well. I'm very happy with it. Also, the delivery was fast."**

Through Chinese to English

**"This product is fantastic! I am very pleased with the ease of use and excellent performance. Also, the delivery was fast."**

*Figure B.7: Generalization Process*

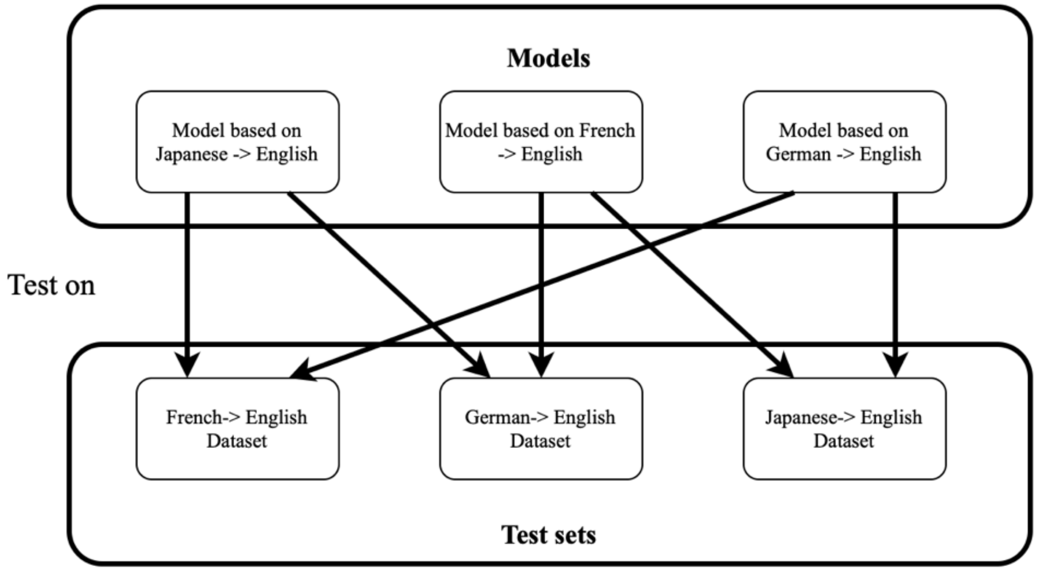

Models
Model based on Japanese -> English
Model based on French -> English
Model based on German -> English
Test on
French-> English Dataset
German-> English Dataset
Japanese-> English Dataset
Test sets

*Figure B.8: Graph of Performance of Translation Augmentation with DeepL Translator on Sentiment Analysis*

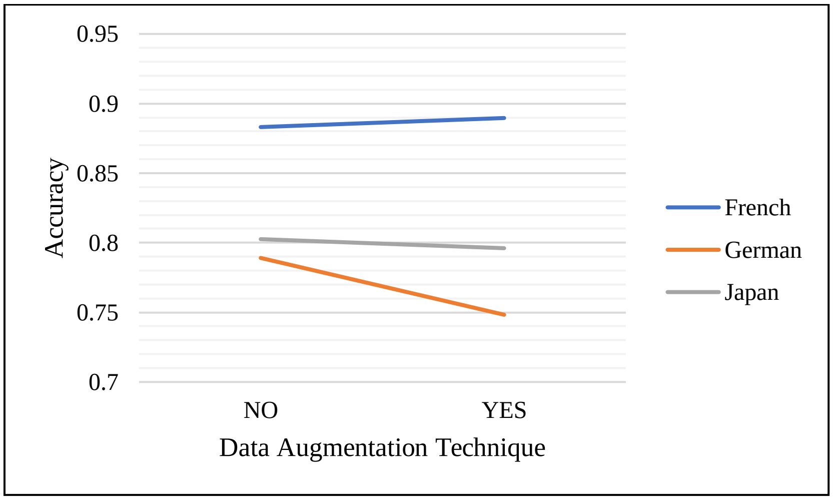


*Figure B.9: Graph of Performance of Translation Augmentation with Google Translator on Sentiment Analysis*

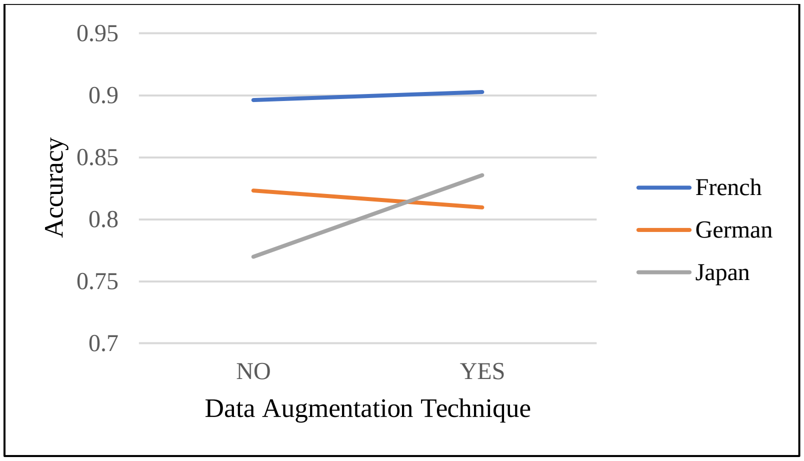


*Figure B.10: Graph of Performance of Generalization of Translation Augmentation on French Training Dataset*

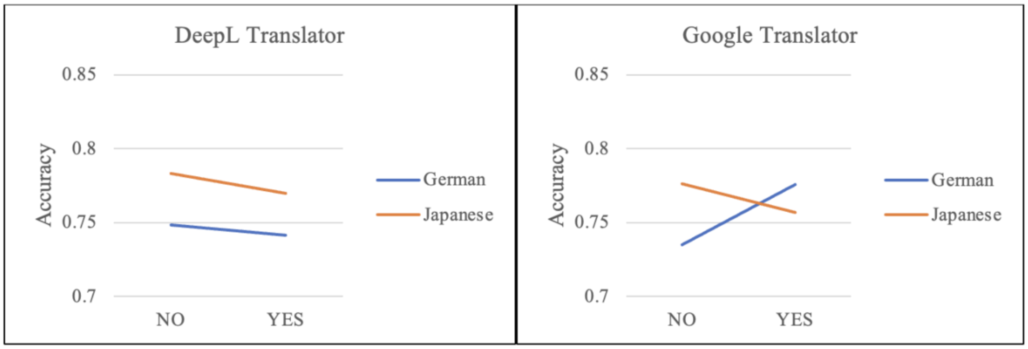


*Figure B.11: Graph of Performance of Generalization of Translation Augmentation on German Training Dataset*

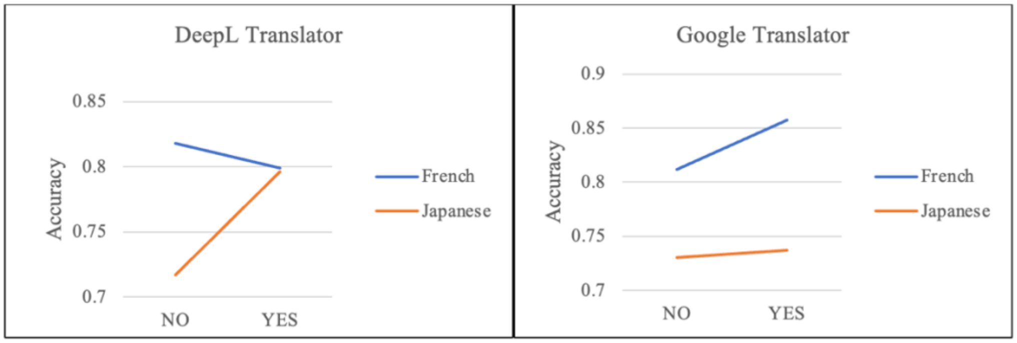
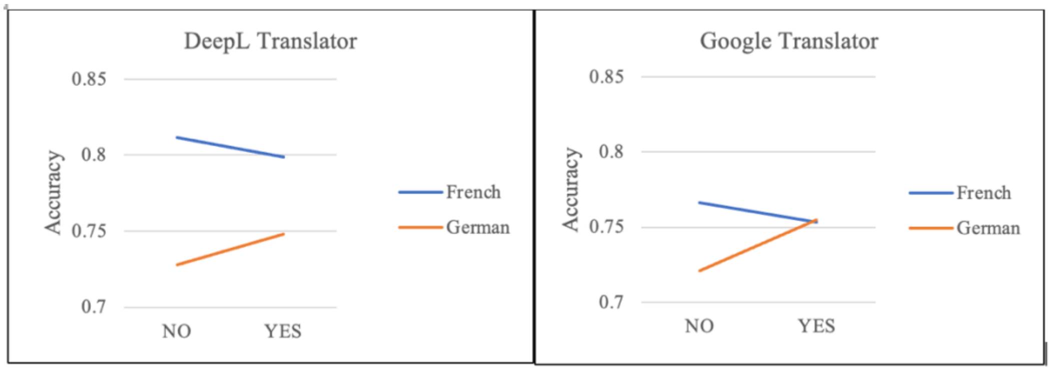


*Figure B.11: Graph of Performance of Generalization of Translation Augmentation on Japanese Training Dataset*



# Appendix C: Literature Review Summary Table

*Table C.1: Data Augmentation*

| Title | Author | Objective | Method | Result |
|---|---|---|---|---|
| Gradient-Based Learning Applied to Document Recognition [16] | L. Bottou, Y. Lecun, Y. BENGIO and P. HAFFNER | To investigate several techniques applied to handwritten character recognition and to compare the techniques on a standard handwritten digit recognition task | Multilayer neural network | The techniques Improved recognition performance |
| Improving Deep Learning with Generic Data Augmentation [1] | L. Talyor and G. Nitschke | To explore various appropriate data augmentation techniques for their dataset and to improve Convolutional Neural Network task performance | Baseline, Flipping, Rotating, Cropping, Color Jippering, Edge Enhancement, Fancy PCA | Generally, most of the methods were good, especially, using cropping method improved 13.82% |
| EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks [29] | J. Wei and K. Zou | To improve performance on text classification tasks by EDA data augmentation technique | EDA technique | EDA technique demonstrated particularly strong results for smaller datasets |
| Data augmentation using back-translation for context-aware neural machine translation [30] | A. Sugiyama and N. Yoshinaga | To obtain large-scale pseudo parallel corpora by back-translating target-side monolingual corpora, and then investigate its impact on the translation performance of context-aware NMT models. | NMT model | The data augmentation impacted the context aware NMT models in terms of BLEU score and specialized test sets on ja→en1 and fr→en. |
| Using back-and-forth translation to create artificial augmented textual data for sentiment analysis models [11] | T. Body, X. Tao, Y. Li, L. Li and N. Zhong | To propose new data augmentation technique and to investigate it on different sample size dataset | Back-and-forth data augmentation technique. | The new data augmentation technique reduced the error rate in binary sentiment classification |
| Comparison of Different Image Data Augmentation Approaches [19] | L. Nanni, M. Paci, S. Brahnam and A. Lumini | To investigate the performance the different sets of data augmentation methods, with two novel approaches proposed here: one based on the discrete wavelet transform and the other on the constant-Q Gabor transform | Several data augmentation technique with Convolutional neural networks (CNNs) | Using diverse data augmentation techniques is a appropriate approach to build an ensemble of classifiers for image classification |

| | | | | |
|---|---|---|---|---|
| A two-stage balancing strategy based on data augmentation for imbalanced text sentiment classification [2] | Z. Peng, H. Li, C. Wang, J. Shi and J. Zhou | To propose new data augmentation technique, balancing strategies for the text classification | Synonym replacement augmentation technique with oversampling stages and noise modification stages | The balancing strategies are able to help balanced the dataset efficiently |
| Toward Text Data Augmentation for Sentiment Analysis [27] | H. Q. Abonizio, E. C. Paraiso and S. Barbon | To investigate the advantages and drawbacks of text augmentation techniques with recent classification algorithms | Easy data augmentation (EDA), back-translation, BART, and pretrained data augmenter | DA improved the performance of the algorithms with both kinds of datasets |
| M-DA: A Multifeature Text Data-Augmentation Model for Improving Accuracy of Chinese Sentiment Analysis [12] | L. Wang, X. Xu, C. Lui and Z. Chen | To propose multifeatured text data augmentation model with a multiple-input single-output network structure for Chinese | Multifeatured text data augmentation model and BilLSTM | The model showed superior to the traditional deep learning |
| Distress Detection in Subway Tunnel Images via Data Augmentation Based on Selective Image Cropping and Patching [18] | K. Maeda, S. Takada, T. Haruyama, R. Togo, T. Ogawa and M. Haseyama | To improve the performance of deep learning-based distress detection to support the maintenance of subway tunnels with a new data augmentation technique | Data Augmentation Based on Selective Image Cropping and Patching | The detection performance can be improved by the data augmentation. |
| Tailored text augmentation for sentiment analysis [28] | Z. Feng, H. Zhou, Z. Zhu and K. Mao | To propose a tailor text augmentation algorithm to improve the synonym replacement augmentation technique. | Data augmentation technique with probabilistic synonym replacement and irrelevant words zero masking. | The DA improved the model's generalization capability |
| Data augmentation in natural language processing: a novel text generation approach for long and short text classifiers [13] | M. Bayer, M.-A. Kaufhold, . B. Buchhold, M. Keller, J. Dallmeyer and C. Reuter | To evaluate a text generation to increase the performance of classifier for long and short text. | Text generation methods | Data augmentation increased the performance by increasing accuracy |

| | | | | |
|---|---|---|---|---|
| Data augmentation techniques in natural language processing [15] | L. F. A. O. Pellicer, T. M. Ferreira and A. H. R. Costa | To compare different data augmentation techniques performances and compare with baseline. | Several data augmentation techniques | The use of DA algorithms is more critical the greater the scarcity of data. |

*Table C.2: Sentiment Analysis*

| **Title** | **Author** | **Objective** | **Method** | **Result** |
|---|---|---|---|---|
| Insights from hashtag #supplychain and Twitter Analytics: Considering Twitter and Twitter data for supply chain practice and research [9] | B. (. Chae | To analyze supply chain tweets, highlighting the current use of Twitter in supply chain contexts, and to develop insights into the potential role of Twitter for supply chain practice and research | Descriptive analytics (DA), content analytics (CA) integrating text mining and sentiment analysis, and network analytics (NA) relying on network visualization and metrics | |
| Sentiment Analysis of Amazon Products Using Ensemble Machine Learning Algorithm [20] | J. Sadhasivam and R. B. Kalivaradhan | To analyze the product review by two machine learning models | Naïve Bayes, SVM, Ensemble algorithms | Ensemble algorithms gave better accuracy compared to existing algorithms |
| Aspect-based sentiment analysis of movie reviews on discussion boards [21] | T. T. Thet, J.-C. Na and C. S. Khoo | To propose method performs fine-grained analysis to determine both the sentiment orientation and sentiment strength of the reviewer towards various aspects of a movie. | A linguistic approach | |
| Sentiment Analysis of Customer Feedback in Online Food Ordering Services [23] | B. Nguyen, V.-H. Nguyen and T. Ho | To analyze the customers' views and sentiments by businesses to assess consumer behavior or a point of view on certain products or services. | Decision Tree, Naïve Bayes, Logistics, SVM | Results can help enterprise managers and service providers get insight into customers' satisfaction with their products or services |
| Deep Recurrent neural network vs. support vector machine for aspect-based sentiment analysis of Arabic hotels' reviews [24] | M. Al-Smadi, O. Oawasmeh, M. Al-Ayyoub, Y. Jararweh and B. Gupta | To evaluate the sentiment analysis of the Arabic hotel reviews | State-of-the-art approaches based on supervised machine learning. Two approaches of deep recurrent neural network (RNN) and support vector machine | SVM approach outperforms the other deep RNN approach in the research investigated tasks (T1: aspect category identification, T2: aspect opinion target expression (OTE) |

| | | | | |
|---|---|---|---|---|
| | | | (SVM), trained along with lexical, word, syntactic, morphological, and semantic features | extraction, and T3: aspect sentiment polarity identification) |
| Sentiment analysis of hotel online reviews using the BERT model and ERNIE model-Data from China [25] | Y. Wen, Y. Liang and X. Zhu | To investigate the emotion analysis of hotel online reviews by using the neural network model BERT | BERT and ERNIE models | Both models can lead to good classification results, but the latter performs better. ERNIE exhibits stronger classification and stability than BERT |
| Efficient Long Short-Term Memory-Based Sentiment Analysis of E-Commerce Reviews [26] | N. K. Gondhi, Chaahat, E. Sharma, A. H. Alharbi, R. Verma and M. A. Shah | To analyze the customers' review | Long Short-Term Memory (LSTM) has been combined with word2vec representation | Improved the overall performance |
| Sentiment analysis in hospitality and tourism: a thematic and methodological review [35] | F. Mehraliyav, I. C. C. Chan and A. P. Kirilenko | To conduct a systematic review and critically analyze the sentiment analysis literature in hospitality and tourism from methodological (data sets and analyzes) and thematic (topics, theories, key constructs, and their relationships) perspectives | Several machine learning models | |