Translating NLI: Investigating the Impact of Cross-lingual Data on Classification Performance in Fever NLI

Guruprasanna Rajukannan Suresh
grajukan@uwaterloo.ca

Sindhuja Chaduvula
schaduvu@uwaterloo.ca

Abstract

Natural Language Inference (NLI) is a crucial task in natural language processing, involving the determination of the logical relationship between two given statements - a premise and a hypothesis. In this context, the FEVER dataset, initially designed for Fact Extraction and Verification tasks, serves as a valuable resource for NLI-based research. The dataset consists of pairs of premises and hypotheses in English, which are then translated into French for multilingual evaluation. These pairs are annotated with labels indicating whether the statements exhibit an entailment, contradiction, or neutral relationship. This project explores the application of NLI models in a multilingual setting by evaluating the performance of translated English premise and hypothesis pairs. The project uses state-of-the-art machine translation techniques to convert English pairs into French. The translated French pairs undergo an NLI classification task to predict logical relationships between the statements.The NLI classifications of the translated pairs are compared with actual French premise and hypothesis pairs to assess the accuracy and reliability of the machine translation process. Through this comprehensive evaluation, valuable insights are gained into the cross-lingual transferability of NLI models, as they encounter diverse linguistic contexts. Ensuring the quality of the translation process is crucial in preserving the integrity of the NLI task outcomes. The outcomes of this study significantly contribute to the advancement of multilingual NLI research, guiding the development of robust NLI models capable of effectively handling multiple languages. Moreover, the project's insights have practical implications in various applications, such as cross-lingual information retrieval and automated fact-checking, enabling more sophisticated understanding and processing of multilingual data in real-world scenarios.

1 Introduction

Advancements in Natural Language Processing (NLP) have addressed the inherent ambiguity of natural language and achieved significant progress in areas such as cross-lingual information extraction and fact verification. Multitask learning, a strategy combining translation and classification tasks on a single dataset, is now being widely adopted. This approach enhances model performance by leveraging insights from one task to another.

In the context of the Fact Extraction and Verification (FEVER)(Thorne et al., 2018)Natural Language Inference task, we integrated translation and classification. This task involves converting English premises and hypotheses into French and then predicting one of three provided stances. To achieve this, we initially implemented a Convolution Neural Network (CNN) for classification(Kim, 2014), which surpassed the baseline performance of the Multinomial Naïve Bayes (MNB) model. For translation, we employed a Sequence2Sequence model due to its proficiency in managing context and variable-length sentences, as well as its attention mechanism.

Finally, the translated French sentences were classified using various transformer models, including BERT (Bidirectional Encoder Representations from Transformers)(Devlin et al., 2019), RoBERTa (Robustly Optimized BERT)(Liu et al., 2019), ELECTRA (Efficiently Learning an Encoder that Classifies Token Replacements Accurately)(Clark et al., 2020), XLNet(Yang et al., 2020). These pre-trained models have the capability to encapsulate intricate language features, significantly enhancing their performance on sentence classification tasks. Among them, RoBERTa emerged as the superior model.We published our code in Github.[1]

[1] https://github.com/GuruprasannaRS/

2 Related Works

2.1 BERT

"BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding"(Devlin et al., 2019)is a groundbreaking paper by Jacob Devlin et al., published in 2018.BERT, or Bidirectional Encoder Representations from Transformers, influences natural language processing model with its unique approach to contextual word representations. BERT uses bidirectional transformers and self-attention for context-aware embeddings, further enhanced by Masked Language Model and Next Sentence Prediction tasks. Its fine-tuning ability leads to unprecedented performance in various NLP tasks, marking a significant shift in the field.

2.2 CNN

The paper "Convolutional Neural Networks for Sentence Classification"(Kim, 2014)by Yoon Kim, published in 2014, introduces a novel approach to sentence classification using Convolutional Neural Networks (CNNs). This work innovatively applies Convolutional Neural Networks to Natural Language Processing, utilizing one-dimensional convolutions on word embeddings to surpass traditional methods in sentiment analysis and topic categorization.

2.3 Bag of Tricks for Efficient Text Classification

The paper "Bag of Tricks for Efficient Text Classification" by Armand Joulin and his co-authors, published in 2017(Joulin et al., 2016), introduces a set of practical techniques to improve the efficiency and effectiveness of text classification models. The paper presents a "Bag of Tricks" - including pretrained embeddings, varied activation functions, adaptive learning rates, and data augmentation - to significantly enhance efficiency and performance in text classification models.

2.4 Attention is all you need

"Attention Is All You Need" by Vaswani et al. (2017)(Vaswani et al., 2023)introduces the transformer architecture, a revolutionary neural network model based on self-attention mechanisms. The Transformer model leverages attention mechanisms for efficient long-sequence processing and global dependency capturing, outperforming traditional models in text classification and inspiring further attention-based NLP research.

2.5 Cross-Lingual Language Pretraining

"Cross-lingual Language Model Pretraining" by Guillaume Lample and his co-authors, published in 2019(Lample and Conneau, 2019), introduces a method for pretraining language models that can effectively transfer knowledge across multiple languages. The authors introduce a cross-lingual pretraining approach using shared vocabulary and translation objectives, which learns multilingual representations and excels in cross-lingual tasks, representing a significant step towards efficient multilingual language models, even for low-resource languages.

3 Dataset

Our dataset(Laurer et al., 2022) is derived from English datasets such as MultiNLI, Fever-NLI, ANLI, LingNLI, and WANLI, transformed into 26 languages, including English, using the latest open-source machine translation models. Each language subset includes a randomly selected sample of 25,000 hypothesis-premise pairs from MultiNLI, Fever-NLI, ANLI, and WANLI, and 5,000 texts from LingNLI due to its smaller size. From this multi-language collection, we focus on the French FEVER NLI subset, containing 25,000 pairs with classification labels. Each pair includes the original English premise and hypothesis, their machine-translated versions in French, and a classification label indicating entailment (0), neutral (1), or contradiction (2). The dataset utilized for the tasks of translation and classification is divided into two subsets: training and testing. The testing subset represents 20% of the total dataset, with the remaining 80% allocated for training.

4 Approach

In addressing the dual challenge of language translation and label classification, our objective was to apply a broad spectrum of deep learning and Natural Language Processing (NLP) strategies.

4.1 Data Preprocessing

In the data preprocessing steps, we employed the Natural Language Toolkit (NLTK)(Bird et al., 2009) and leveraged pre-trained tokenizers for data preprocessing of both English and French premises and hypotheses. NLTK's extensive capabilities and

`translational_nli`

support for pre-trained models provided a powerful combination to process text data effectively.

For the English text, NLTK's pre-trained tokenizer allowed us to efficiently perform tokenization, breaking down the sentences into individual words. This pre-trained model, trained on a large corpus of English text, contributed to accurate and contextually appropriate segmentation. For the French text, we preserved special characters and contextual meaning by employing a French-specific pre-trained tokenizer, considering the unique linguistic nuances of the language, thereby maintaining sentence integrity and preventing any information loss.

Integrating pre-trained tokenizers into our data preprocessing pipeline significantly reduced the burden of building tokenizers from scratch and provided a robust foundation for subsequent translation and classification tasks.

4.2 Sequence2Sequence Model

Seq2Seq models, perfect for tasks like language translation with different input-output sequence lengths, utilize an encoder and decoder, usually recurrent neural networks. The encoder turns a source sentence into a context vector, while the decoder produces the target sentence token-by-token using the context vector and preceding tokens. Contemporary Seq2Seq models(Sutskever et al., 2014) often feature an attention mechanism, establishing a soft alignment between source and target sentences via attention scores.

Pretraining a Seq2Seq model(Bahdanau et al., 2016) on a large, diverse corpus captures language complexities, including vocabulary, syntax, semantics, and pragmatics. Following this, the pretrained model is fine-tuned on a task-specific dataset, adjusting the weights to optimize the model for a specific task, often leading to superior performance compared to models trained from scratch.

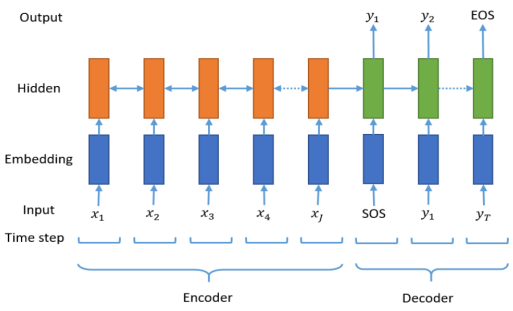

Figure 1: seq2seq model architecture

4.3 Baseline Models (MNB, CNN)

Multinomial Naive Bayes is a probabilistic classifier commonly used for text classification tasks. It is based on Bayes' theorem and assumes that the features are conditionally independent given the class label. In the context of text classification, MNB is particularly effective for tasks involving word frequencies or counts. It works well with sparse data and requires relatively fewer training examples.

Convolutional Neural Networks are deep learning models that have shown remarkable success in various computer vision tasks. In recent years, they have also been applied to natural language processing tasks, including text classification. In the context of text classification(Kim, 2014), CNNs use convolutional layers to extract local patterns or features from the input text. They can capture important n-gram features and are capable of learning hierarchical representations of the text.

4.4 BERT (Bidirectional Encoder Representations from Transformers)

BERT (Bidirectional Encoder Representations from Transformers) is a game-changing model in Natural Language Processing (NLP), introduced by Google researchers. The model leverages the Transformer architecture, which processes all words in a sequence in parallel, providing an enhanced contextual understanding. BERT's process consists of two steps: pretraining and fine-tuning. During pretraining, BERT learns from a massive corpus of text (like Wikipedia and various books) to understand word semantics and relationships. It then fine-tunes these prelearned representations on a specific task with a smaller labeled dataset.

BERT's bidirectional processing provides a more nuanced understanding of text, aiding in classification tasks by comprehensively capturing context from both directions. Its multiple sizes and language capabilities allow it to be fine-tuned for various classification tasks across languages(Devlin et al., 2019). Leveraging BERT's sophisticated text representations, state-of-the-art results have been achieved in text classification tasks including sentiment analysis, topic classification, and intent detection.

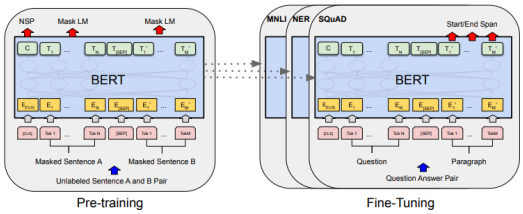

Figure 2: BERT model architecture

4.5 RoBERTa (A Robustly Optimized BERT Pretraining approach)

RoBERTa, standing for Robustly Optimized BERT Pretraining Approach, is an optimized variant of BERT, introduced by Facebook AI. It adapts and refines BERT's pre-training process with more extensive hyperparameters and a change in the masking strategy of the model. RoBERTa's training leverages larger batch sizes and extends the training time, allowing the model to learn richer, deeper representations. It also makes use of more training data, providing a broader linguistic context. These enhancements allow RoBERTa to surpass BERT's performance on various NLP benchmarks, showing substantial improvements. For classification tasks, the refined training process equips RoBERTa(Liu et al., 2019) with a more robust understanding of textual context and nuances, leading to more accurate and efficient classification of text. This could include tasks like sentiment analysis, topic classification, or document categorization, thus making RoBERTa a valuable tool in a wide range of NLP applications.

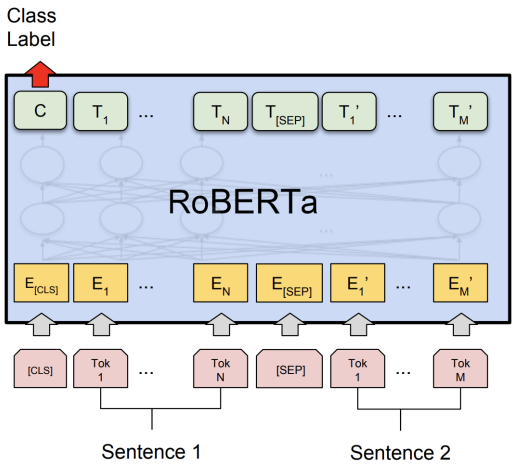

Figure 3: RoBERTa model architecture

4.6 ELECTRA (Efficiently Learning an Encoder that Classifies Token Replacements Accurately)

ELECTRA (Efficiently Learning an Encoder that Classifies Token Replacements Accurately)(Clark et al., 2020) is a transformer-based language model that brings a new pre-training objective to the table - "replaced token detection". Developed by researchers at Google Research/Stanford University, ELECTRA diverges from BERT's masked language modeling by introducing a discriminative task where the model predicts if a token in the sentence is original or replaced by another token from a small generator network. This results in the model training on all tokens instead of a subset, thereby leading to more efficient use of compute resources during pre-training. Additionally, ELECTRA's pre-training aligns better with the fine-tuning phase, yielding improved performance. When applied to classification tasks, the in-depth understanding of linguistic context and relations fostered by ELECTRA's training method results in high-performance text classification.

4.7 XLNet (Generalized Autoregressive Pretraining for Language Understanding)

XLNet(Yang et al., 2020) is an advanced transformer-based model developed by researchers at Google Brain and Carnegie Mellon University, which employs a generalized autoregressive pretraining method, combining the advantages of both autoregressive and autoencoding models. Unlike BERT, which is an autoencoder and predicts masked tokens independently, XLNet uses a permutation-based training objective to keep the autoregressive property, predicting each token considering the context of all other tokens in the sequence, regardless of their order. This innovative method allows XLNet to model bidirectional contexts, similar to BERT, but without the issues of the "masking" technique used in BERT. When it comes to classification tasks, the deep and broad understanding of contextual relationships gleaned from XLNet's unique training process results in highly accurate classification of text.

5 Experiments

5.1 Translation

Our Seq2Seq model(Sutskever et al., 2014) is designed for English to French translation, using the pre-trained "Helsinki-NLP/opus-mt-en-

fr"(Tiedemann and Thottingal, 2020) transformer model from Hugging Face. Input sentences are tokenized, padded, and converted into tensors using AutoTokenizer. The model uses the Adam optimizer with weight decay for regularization and is fine-tuned over 3 epochs. Employing this model, we have trained on our dataset and used the test data to generate predictions for French sentences.

5.2 Classification

For classification, our model is optimized using the AdamW optimizer with a learning rate of 2e-5. The model and data are moved to the GPU if one is available, otherwise, they use the CPU. The loss and accuracy for training and validation stages for 5 epochs is calculated. During training, after zeroing the gradients, the model computes a forward pass the model makes predictions based on the inputs and these predictions are compared to the actual labels to calculate the loss. Backpropagation is then performed by calling loss.backward(), and the optimizer updates the model parameters. The model is then set to evaluation mode for validation. Accuracy is computed by comparing the model's highest-probability predictions to the actual labels. Both the training and validation losses and accuracies are tracked.

5.3 Influence of translation on Classification

The translated French sentences, obtained from a seq2seq model, serve as inputs for various classification models, including Multinomial Naive Bayes (MNB), Convolutional Neural Network (CNN), and several transformer models like BERT, RoBERTa, ELECTRa, and XLNet. These models, especially transformer models pre-trained on extensive corpora, can capture complex language features beneficial for sentence classification. The classifications produced by these models are compared to the actual classifications of the French sentences in the test dataset. This process enables the measurement of model performance using metrics such as accuracy, precision, recall, or F1 score. By evaluating multiple models, you can discern which model(s) demonstrate superior performance for this specific classification task.

5.4 Evaluation

BLEU score(Papineni et al., 2002), a standard measure in machine translation, assesses precision by considering the proportion of n-grams in the translated text that also appear in the reference text. It factors in a penalty for brevity and evaluates up to 4-gram precision, with varying weights for each n-gram level to balance the importance of individual word matching versus longer phrase matching. In parallel, sentiment analysis, conducted with TextBlob, gauges the emotional tone of text, assigning sentiment scores from -1.0 (negative) to 1.0 (positive). Semantic similarity, calculated using the Universal Sentence Encoder and cosine similarity, quantifies how closely two texts align in meaning, with scores ranging from 0 (dissimilar) to 1 (identical). Combined, these methods provide a comprehensive overview of a text's emotional tone and semantic content.

Classification model performance is assessed using a range of metrics including precision, recall, F1 score, test accuracy, and kappa (Powers, 2020). Precision is key when minimizing false positives is critical, while recall becomes important when capturing all actual positives is the goal. Test accuracy provides an overall measure of model effectiveness, while the F1 score offers a balance between precision and recall. The kappa score is a robust metric that accounts for chance agreement. Misclassification accuracy, computed per label and overall, measures how frequently a model incorrectly classifies instances, which can help to identify challenging classes and guide improvements. These metrics, when combined, offer a comprehensive view of a model's ability to correctly classify data, aiding model selection and optimization. This process also identifies the strengths and weaknesses of individual models, guiding future model refinement and deployment.

6 Results

6.1 Comparison of Translation Results

The BLEU score(Papineni et al., 2002) is a key machine translation metric, that quantifies the similarity between machine-generated and reference translations through overall and n-gram overlaps.Figure 4 depicts the BLEU scores for both premises and hypothesis. The BLEU scores for both premises and hypothesis are highest at the 1-gram level, indicating good word-level matching, with the hypothesis performing slightly better (0.73 vs. 0.70). As the n-gram length increases, scores decrease, suggesting less success in matching longer phrases and maintaining sentence structure, but the hypothesis still outperforms the premises. The overall scores reinforce this, with

Sentences	Sentiment Analysis
Original Premises	0.11384
Translated Premises	0.04092
Original Hypothesis	0.02542
Translated Hypothesis	0.00714

Table 1: Comparison of Sentiment Analysis of original and translated sentences

the hypothesis showing a marginally better performance (0.56 vs. 0.50). However, both show room for improvement, especially in matching longer phrases (as indicated by lower scores for higher n-grams).

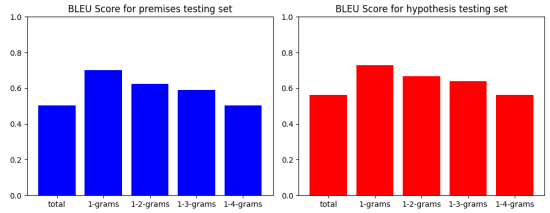

Figure 4: BLEU score

Sentiment analysis is a field of study that computationally determines the emotional tone behind words to gain an understanding of the attitudes, opinions, and emotions expressed within an online mention. The sentiment scores range from -1 (very negative) to +1 (very positive). The sentiment analysis results from table 1 reveal interesting insights. The original hypothesis has an average sentiment of 0.025, decreasing slightly to 0.007 in the translated hypothesis and the original premises have a higher average sentiment of 0.114, decreasing to 0.041 in the translated premises, indicating a minor loss in positive sentiment during translation. The premises in both original and translated forms exhibit a higher positive sentiment than the hypothesis.

Semantic similarity is a measure of the degree to which two pieces of text carry the same meaning. From the table 2 The semantic similarity scores, ranging from 0 (completely dissimilar) to 1 (identical), show the hypothesis and premises have averages of 0.758 and 0.774 respectively, suggesting good preservation of meaning during translation. On average, both hypothesis and premises translations are closely semantically aligned with their originals, with premises translations exhibiting slightly higher similarity.

Sentences	Semantic Similarity
Premises	0.77360
Hypothesis	0.75781

Table 2: Comparison of Semantic Similarity of original and translated sentences

6.2 Comparison of Classification Results

As depicted in Figure 5, transformer-based models (Electra, XLNet, BERT, and RoBERTa) outshine traditional models (MNB and CNN) in terms of both train and validation accuracy. These transformer models exhibit high train accuracy ranging from 83.5% to 92.0%, signifying their efficiency in comprehending complex language patterns. Conversely, the MNB model presents a lower train accuracy of 44.9%, and the CNN, albeit learning more complex representations, registers a train accuracy of 59.8%, lower than the transformer models. In terms of validation accuracy, a measure of model generalization, Electra and XLNet report similar values around 75.6%-75.8%. BERT and RoBERTa stand out by attaining the highest validation accuracy of 78.2%, underpinned by their ability to capture intricate linguistic relationships and contexts using bidirectional attention mechanisms.

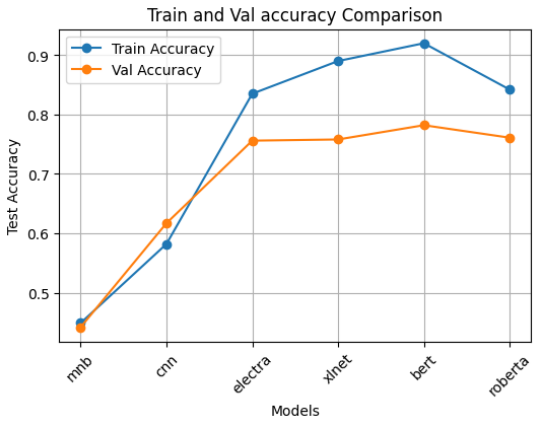

Figure 5: Train and Validation Accuracy Comparison for Classification Models

The Figure 6 illustrates the evaluation of the 7 classification models on the "translated" test accuracy. The results closely align with their "original" test accuracy, validating the reliability of the evaluation process. Multinomial Naive Bayes (MNB) exhibits limited effectiveness, achieving a "translated" test accuracy of 44.9%, which closely matches the "original" test accuracy of 44.6%. Convolutional Neural Network (CNN) performs

slightly better with a "translated" test accuracy of 59.8%, consistent with the "original" test accuracy. Among the transformer-based models, Electra and XLNet show moderate performance, achieving "translated" test accuracies of 72.4% and 73.8%, respectively, closely aligned with their "original" test accuracies depicted in Figure 6.

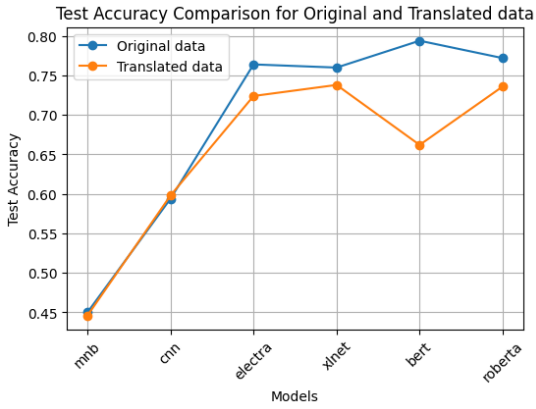

Figure 6: Test Accuracy Comparison for Original and Translated data

In the classification tasks, transformer-based models, including BERT, RoBERTa, Electra, and XLNet, exhibited robust performance, surpassing traditional models like MNB and CNN in terms of accuracy and adaptability to language patterns. Remarkably, RoBERTa stood out with an "original" and "translated" test accuracy of 79.4%, demonstrating exceptional consistency. BERT also performed well with a "translated" test accuracy mirroring the "original" at 66.2%. However, the MNB model had a lower performance with around 41% precision, recall, and F1-scores, and kappa scores of 0.13 for both "original" and "translated" data. The CNN model demonstrated limited precision (22.9%), indicating a higher false positive rate, and moderate recall (33.3%), resulting in a lower F1-score of 27.2% for the original data, as shown in Figure 7. Furthermore, CNN's kappa score of 0.0 suggested a poor agreement between the model's predictions and the actual data, likely due to its lower precision and recall.

Transformer models, Electra and XLNet, showcase consistent performance with F1-scores around 63.9% to 65.7% for both original and translated data. BERT and RoBERTa models indicate an interesting trend with a slight improvement in performance on translated data; BERT's F1-score rises from 60.3% (original) to 66.3% (translated) Figure 7 and RoBERTa's F1-score remains high,

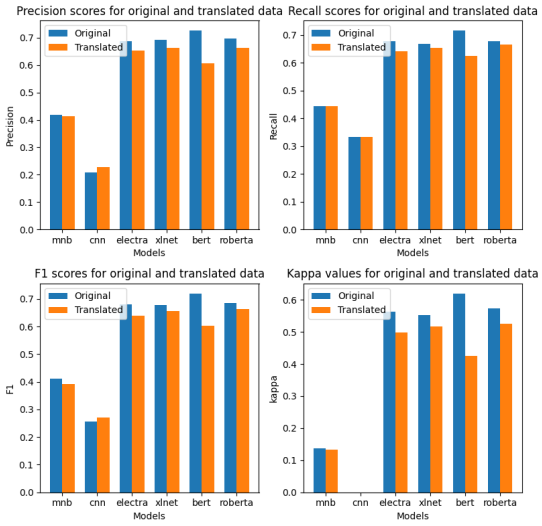

Figure 7: Comparison of Precision, Recall, F1 Score and Kappa values for original and translated data

though it dips slightly from 72.0% (original) to 68.5% (translated).

Misclassification analysis across models (MNB, CNN, BERT, ROBERTA, ELECTRA, XLNet) shows varied performances. CNN excels in entailment classification (2.35 misclassification rate) but struggles with neutral instances (16.55). The MNB model performs best for neutral (8.10) and contradiction (14.50) labels. Despite difficulties with entailment, MNB maintains the lowest overall misclassification (14.50), whereas CNN, despite strength in entailment, exhibits the highest overall misclassification rate (22.35) as shown in table 3. Among transformer models, ELECTRA displays a slightly better average misclassification rate (17.00), reflecting nuanced performance characteristics based on model type and classification category.

7 Conclusion and Future work

In your study, English sentences were translated into French using a seq2seq model, achieving a BLEU score of 0.6, which indicates a decent level of translation quality. The sentiment analysis showed neutrality, suggesting that the emotional tone of the sentences was maintained in translation. The translation process maintained an average semantic similarity, yet, classification of translated sentences saw XLNet and RoBERTa outperforming BERT - the best performer on original sentences. However, a slight reduction in precision, recall, and F1 scores for translated versus original sentences suggests a possible loss of nuances during transla-

Models	Entailment(%)	Neutral(%)	Contradiction(%)	Overall(%)
MNB	32.35	8.10	14.50	18.31
CNN	2.35	16.55	22.35	13.75
BERT	22.95	13.75	17.10	17.93
RoBERTa	22.35	14.25	17.25	17.95
ELECTRA	23.30	14.70	17.00	18.33
XLNet	22.20	14.35	17.80	18.11

Table 3: Comparison of Misclassification Accuracy

tion impacting classification.

To improve the precision, recall, and F1 scores for translated data, a multifaceted approach is beneficial. Enhancements could include refining translation models, expanding and diversifying training data, post-translation adjustments, and fine-tuning the classification models on the translated data. Knowledge distillation(Hinton et al., 2015) is a process where a smaller, simpler model (the student) is trained to replicate the performance of a larger, more complex model (the teacher). This technique can help in handling the nuances of translated sentences more effectively by learning from a teacher model that performs well on the task.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2016. Neural machine translation by jointly learning to align and translate.

Steven Bird, Ewan Klein, and Edward Loper. 2009. Natural language processing in python. In *Proceedings of the 7th Python in Science Conference (SciPy 2008)*, pages 61–66.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

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

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification.

Yoon Kim. 2014. Convolutional neural networks for sentence classification.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining.

Moritz Laurer, Wouter van Atteveldt, Andreu Salleras Casas, and Kasper Welbers. 2022. Less Annotating, More Classifying – Addressing the Data Scarcity Issue of Supervised Machine Learning with Deep Transfer Learning and BERT - NLI. *Preprint*. Publisher: Open Science Framework.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318.

David M. W. Powers. 2020. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2020. Xlnet: Generalized autoregressive pretraining for language understanding.