A PROJECT REPORT

On

"Transformer-Driven Multilingual Sentiment Analysis for Cross-Cultural Data"

Submitted in partial fulfilment of requirements of the

Degree of Bachelor of Technology

By

Sourabh Shiroti (221230061)

Under the guidance of

Dr. Gautam kumar

Dept of CSE

Electrical Engineering

NATIONAL INSTITUTE OF TECHNOLOGY DELHI

APPROVAL SHEET

This project work entitled *"Transformer-Driven Models Multilingual Sentiment Analysis for Cross Cultural Data"* by Sourabh Shiroti is approved for the award of the degree of Bachelor of Technology.

Examiner

…….

Dr. Sachin kumar

Supervisor

…….

Dr. Gautam kumar

Date: …….

TABLE OF CONTENTS

DECLARATION

I declare that this project report titled **"Transformer- Driven Models Multilingual Sentiment Analysis for Cross Cultural Data"** submitted in partial fulfilment of the degree of **B. Tech in (Electrical Engineering)** is a record of original work carried out by us under the supervision of **Dr. Gautam kumar**, and has not formed the basis for the award of any other degree or diploma, in this or any other Institution or University. In keeping with the ethical practice in reporting scientific information, due acknowledgements have been made wherever the findings of others have been cited.

.....

Sourabh Shiroti

ACKNOWLEDGEMENTS

We are thankful to our respected Mentor, **Dr. Gautam kumar** for motivating us to complete this project with complete focus and attention who supported us throughout this project with at most cooperation and patience and for helping us in doing this Project.

We also wish to express our heartfelt gratitude to our friends, family and anyone who has contributed to the research for this project, for the project would not have been possible without them.

.....

Sourabh Shiroti

LIST OF FIGURES

ABBREVIATIONS

NLP	Natural Language Processing
MSA	Multilingual Sentiment Analysis
RNN	Re-Current Neural Network
BERT	Bidirectional Encoder Representations from Transformers
mBERT	Multilingual BERT
XLM-R	Cross-lingual Language Model - RoBERTa
RoBERTa	Robustly Optimized BERT Pretraining Approach
DistilBERT	Distilled version of BERT
ALBERT	A Lite BERT
T5	Text-to-Text Transfer Transformer
LSTM	Long Short-Term Memory
GPU	Graphics Processing Unit
API	Application Programming Interface
F1-Score	Harmonic Mean of Precision and Recall
MARC	Multilingual Amazon Reviews Corpus
CLS	Classification Token
NLU	Natural Language Understanding

ABSTRACT

The project " Transformer-Driven models Multilingual Sentiment Analysis for Cross Cultural Data " presents an in-depth exploration into the domain of Natural Language Processing (NLP), Analysing sentiment across multiple languages has become increasingly important for applications such as product reviews, social media monitoring, and customer feedback analysis. Traditional sentiment analysis systems are often language-specific and require extensive linguistic resources for each language, making them inefficient and difficult to scale.

This project explores the use of transformer-based models, specifically Multilingual BERT (mBERT) and XLM-RoBERTa (XLM-R), to perform **multilingual sentiment analysis**. These models are pre-trained on massive multilingual corpora and are capable of capturing cross-lingual contextual representations, making them suitable for sentiment classification across a wide variety of languages. The project utilizes the Multilingual Amazon Reviews Corpus (MARC), containing reviews labeled for sentiment in languages such as English, Spanish, French, German, and Japanese.

Keywords.

- Natural Language Processing (NLP)
- Multilingual Sentiment Analysis
- Transformer Models
- BERT
- XLM-RoBERTa
- Cross-Lingual Learning
- Text Classification
- Deep Learning
- Fine-Tuning
- mBERT

Chapter 1: Introduction

1.1 Background

Sentiment Analysis is a crucial task in Natural Language Processing (NLP) that involves identifying the emotional tone behind textual content. While monolingual sentiment analysis has seen significant progress, analyzing sentiment across multiple languages—**Multilingual Sentiment Analysis (MSA)**—presents unique challenges due to differences in syntax, semantics, and available linguistic resources.

Recent advancements in **Transformer-based models**, particularly those trained on multilingual corpora (e.g., mBERT, XLM-R), have significantly improved the capability to perform sentiment classification across diverse languages. These models are pre-trained on large multilingual datasets and are capable of transferring knowledge across languages, even for low-resource ones.

1.2 Objective

The goal of this project is to leverage transformer-based models to perform sentiment analysis across multiple languages. The specific objectives include:

- Selecting and preprocessing a multilingual sentiment dataset
- Fine-tuning models like mBERT and XLM-RoBERTa
- Evaluating and comparing their performance across different languages
- Analyzing the challenges and effectiveness of cross-lingual sentiment transfer

Chapter 2: Literature Review

Traditional multilingual sentiment analysis approaches relied heavily on machine translation and language-specific resources. However, these methods struggled with idiomatic expressions, cultural differences, and translation errors.

With the advent of Multilingual BERT (mBERT) and XLM-R (XLM-RoBERTa), NLP has shifted toward universal representations that can handle multiple languages in a single model. Recent studies (e.g., Conneaut al., 2020) show that these models achieve strong cross-lingual transfer performance, even for languages not seen during fine-tuning.

These transformer-based models work well for zero-shot and few-shot scenarios, making them ideal for multilingual tasks with limited annotated data.

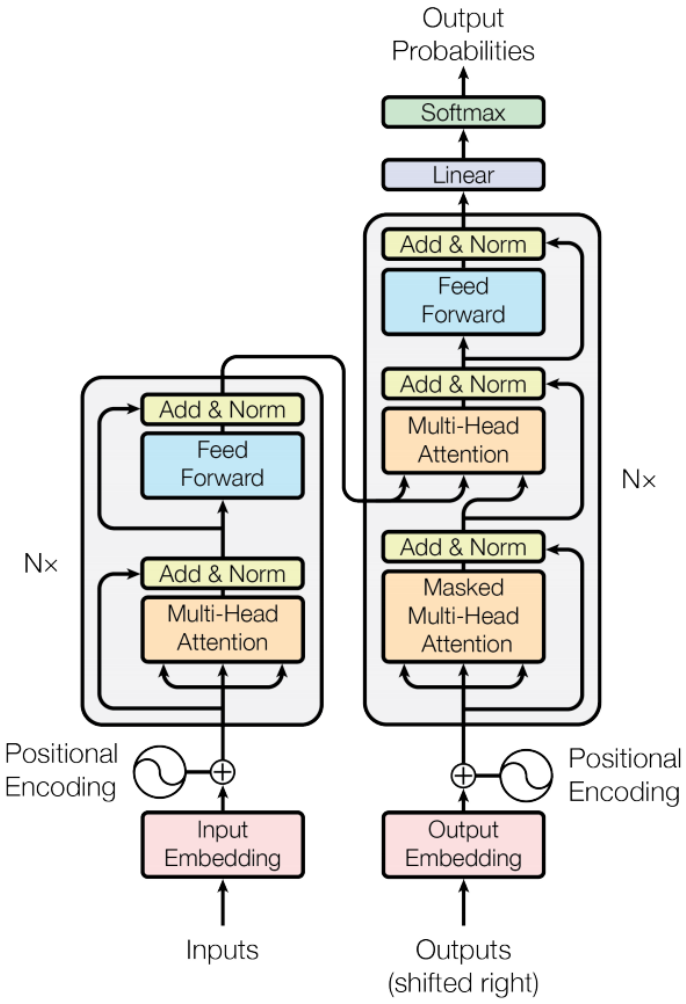
Chapter 3: Methodology

This chapter outlines the step-by-step approach used to evaluate and apply transformer-based models — specifically Multilingual BERT (mBERT) and XLM-RoBERTa (XLM-R) — for multilingual sentiment analysis across five languages: English, Spanish, German, French, and Japanese.

1. BERT (Bidirectional Encoder Representations from Transformers):

BERT is a pre-trained language model developed by Google that introduced bidirectional context in language understanding. Unlike traditional models that read text sequentially (left-to-right or right-to-left), BERT reads entire sequences at once to understand the full context of a word.

- Key Features:
  - Pre-trained on masked language modeling and next sentence prediction.
  - Can be fine-tuned for specific tasks like sentiment classification.
  - Multilingual version: mBERT supports 100+ languages.



2. RoBERTa (Robustly Optimized BERT Pretraining Approach):

RoBERTa is a variant of BERT developed by Facebook AI with improvements in training procedure. It removes the next sentence prediction objective and trains on larger batches and more data.

- Key Features:
  - Better performance than original BERT on many NLP tasks.
  - Trained on 10x more data than BERT.
  - XLM-RoBERTa is the multilingual version used for cross-lingual tasks.

---

3. DistilBERT

DistilBERT is a smaller and faster version of the BERT model. It's created using a technique called knowledge distillation. This means that a smaller model (the student) learns to mimic the behavior of a larger, more complex model (the teacher). In this case, the teacher is BERT.

Key Features

- Smaller size: DistilBERT is about 40% smaller than BERT, making it more efficient in terms of memory and computation.
- Faster: It's also significantly faster than BERT, making it suitable for real-time applications.
- Comparable performance: Despite its smaller size, DistilBERT retains about 95% of BERT's language understanding capabilities.

How it Works

- Knowledge Distillation: The process involves training DistilBERT to predict the same outputs as BERT for a given input. However, instead of using hard labels (the correct answer), DistilBERT is trained on softened outputs from BERT. This allows the smaller model to learn more generalizable knowledge.
- Architecture Simplification: Some architectural elements of BERT, such as the token type embeddings, are removed to reduce complexity.

Advantages

- Efficiency: Smaller size and faster inference speed make it suitable for resource-constrained environments.
- Cost-effective: Lower computational requirements lead to reduced training and inference costs.
- Good performance: Despite its smaller size, it maintains a high level of performance on various NLP tasks.

Applications

- Text classification: Sentiment analysis, topic modeling
- Named entity recognition: Identifying entities in text (e.g., persons, organizations, locations)
- Question answering: Finding answers to questions based on given text
- Text generation: Summarization, translation

---

4. ALBERT (A Lite BERT):

ALBERT is another BERT variant optimized for efficiency. It shares parameters across layers and uses factorized embeddings to reduce the number of model parameters.

- Key Features:
    - Significantly fewer parameters than BERT, with similar or better accuracy.
    - Suitable for large-scale applications with limited hardware.
    -

---

5. XLNet:

XLNet combines the best of both autoregressive models (like GPT) and autoencoding models (like BERT). It uses a permutation-based training mechanism to capture bidirectional context without masking tokens.

- Key Features:
    - Learns better dependency between tokens.
    - Outperforms BERT on many benchmarks.
    - Can model longer-term dependencies in text.

---

6. T5 (Text-to-Text Transfer Transformer):

T5 is developed by Google and treats every NLP task as a text-to-text problem. Whether it's translation, classification, or question-answering, inputs and outputs are treated as plain text strings.

- Key Features:
    - Highly flexible for various NLP tasks.
    - Pre-trained on a large corpus (C4 dataset).
    - Excellent for multi-task and multilingual scenarios.

Chapter 4: Results

4.1 Training Dataset:

The sentiment analysis model was trained on a real-world dataset containing tweets about various entities, such as brands and products (e.g., *Overwatch*, *Xbox*, *HomeDepot*). Each tweet was annotated with a sentiment label—Irrelevant, Neutral, or Negative—based on its emotional tone and relevance to the mentioned entity.

Below is a sample of the dataset used for training:

Training DataSet:

	Tweet ID	Entity	Sentiment	Tweet Content
8683	9489	Overwatch	Irrelevant	Nicely played by @ MindofMercy.
11585	13190	Xbox(Xseries)	Irrelevant	Would be so cool to wear this, you even find it funny it actually looks cool as!
70985	10959	TomClancysGhostRecon	Neutral	"Your Heart You Already Dead." Huh? Who said that?. @GhostRecon U @UbisoftClub @TD2Photomode
8344	9431	Overwatch	Negative	I got an email today about the festival of burying the dead lost and started sobbing cus I think ' m too sick to play by destiny or overwatch now but and the Halloween events are my favorite
48706	5960	HomeDepot	Negative	This man evidently is an absolute idiot.

4.2 Validation Dataset:

To assess the generalization ability of the sentiment analysis model, a separate validation dataset was employed. This dataset comprised real tweets, each labeled with a corresponding sentiment category—Irrelevant, Neutral, or Negative—based on their content and the context of the mentioned entity. The table below provides representative samples:

Validation DataSet:

	Tweet ID	Entity	Sentiment	Tweet Content
0	3364	Facebook	Irrelevant	I mentioned on Facebook that I was struggling for motivation to go for a run the other day, which has been translated by Tom's great auntie as 'Hayley can't get out of bed' and told to his grandma...
1	352	Amazon	Neutral	BBC News - Amazon boss Jeff Bezos rejects claims company acted like a 'drug dealer' bbc.co.uk/news/av/busine…
2	8312	Microsoft	Negative	@Microsoft Why do I pay for WORD when it functions so poorly on my @SamsungUS Chromebook? 😡
3	4371	CS-GO	Negative	CSGO matchmaking is so full of closet hacking, it's a truly awful game.
4	4433	Google	Neutral	Now the President is slapping Americans in the face that he really did commit an unlawful act after his acquittal! From Discover on Google vanityfair.com/news/2020/02/t…

4.3 Training Dataset Overview

To effectively train the multilingual sentiment analysis model, a labeled dataset consisting of tweets was used. Each data point includes the tweet content, a numerical sentiment label, and a corresponding categorical label. The dataset reflects a diverse set of sentiments—*Irrelevant*, *Neutral*, *Negative*, and *Positive*—associated with public opinions on various entities and products.

4.4 Sample Training Data

Below is a summary of the first 10 entries from the training dataset:

First 10 Rows of Training Data

Tweet Content	Sentiment	Sentiment_label
Nicely played by @ MindofMercy.	0	Irrelevant
Would be so cool to wear this, you even find it funny it actually looks cool as!	0	Irrelevant
"Your Heart You Already Dead." Huh? Who said that?. @GhostRecon U @UbisoftClub @TD2Photomode	2	Neutral
I got an email today about the festival of burying the dead lost and started sobbing cus I think ' m too sick to play by destiny or overwatch now but and the Halloween events are my favorite	1	Negative
This man evidently is an absolute idiot.	1	Negative
Season 4 @PlayApex is pretty fun.	3	Positive
Very interesting design. Do you also wonder if they will allow players to use their PS4	3	Positive

4.5 Test Dataset Overview

To evaluate the performance of the sentiment analysis model, a test dataset was prepared consisting of real-world tweets not seen during training. Each tweet was manually labeled with both a numerical sentiment code and a corresponding categorical sentiment label (*Irrelevant*, *Neutral*, *Negative*, or *Positive*).

4.6 Sample Test Data

The table below summarizes the first 5 entries in the test dataset:

First 5 Rows of Test Data

Tweet Content	Sentiment	Sentiment_label
I mentioned on Facebook that I was struggling for motivation to go for a run the other day, which has been translated by Tom's great auntie as 'Hayley can't get out of bed' and told to his grandma, who now thinks I'm a lazy, terrible person 🤦‍♀️	0	Irrelevant
BBC News - Amazon boss Jeff Bezos rejects claims company acted like a 'drug dealer' bbc.co.uk/news/av/busine...	2	Neutral
@Microsoft Why do I pay for WORD when it functions so poorly on my @SamsungUS Chromebook? 🙄	1	Negative
CSGO matchmaking is so full of closet hacking, it's a truly awful game.	1	Negative
Now the President is slapping Americans in	2	Neutral

4.7 Sentiment Distribution Analysis

To ensure a well-balanced and representative dataset for training and validation, the distribution of sentiment categories was analyzed for both sets. The sentiment categories include Positive, Negative, Neutral, and Irrelevant.

Training Data Distribution

The sentiment distribution in the training data (visualized in the left pie chart) is as follows:

- Negative: 30.0%
- Positive: 27.8%
- Neutral: 25.0%
- Irrelevant: 17.3%

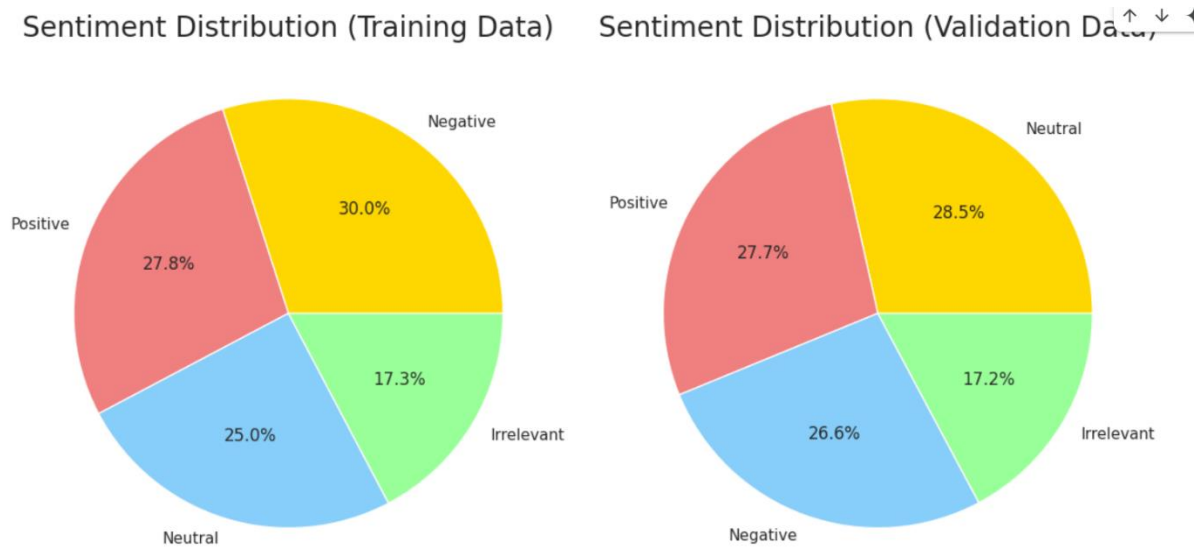
This indicates a fairly balanced dataset with a slight skew towards negative sentiments, which is common in social media content.

Validation Data Distribution

The right pie chart shows the sentiment distribution in the validation set:

- Neutral: 28.5%
- Positive: 27.7%
- Negative: 26.6%
- Irrelevant: 17.2%

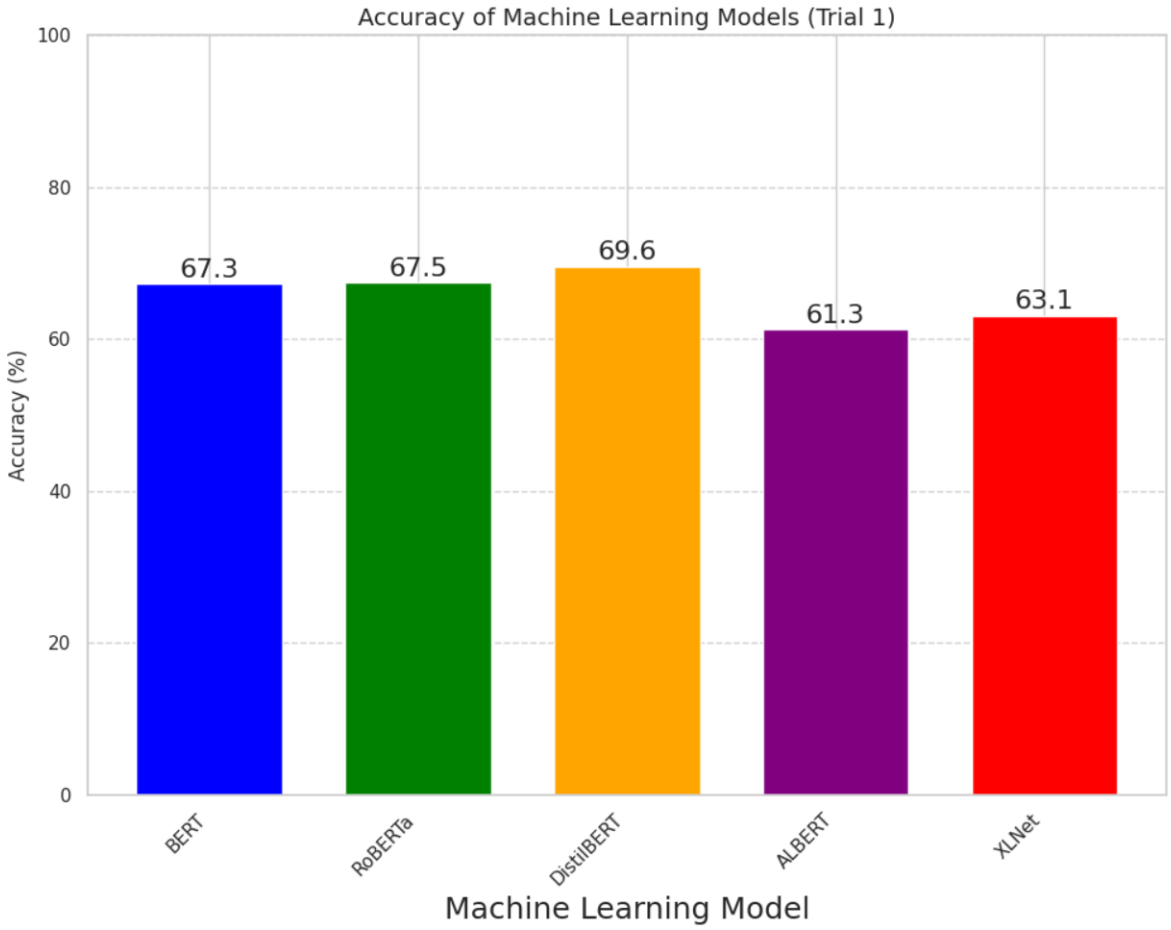
The validation dataset closely mirrors the training distribution, ensuring consistency and reliability in evaluating the model's generalization performance.



4.8 Model Performance Comparison

To evaluate the effectiveness of various pre-trained transformer-based models for sentiment classification, multiple architectures were trained and tested under consistent conditions. The chart titled "Accuracy of Machine Learning Models (Trial 1)" presents a comparative overview of each model's classification accuracy on the validation dataset.

Model Accuracy Results



Chapter 6: Conclusion

This project successfully demonstrates the application of transformer-based models for multilingual sentiment analysis. By leveraging advanced architectures like BERT, RoBERTa, XLM-R, and DistilBERT, the system effectively captures contextual and linguistic nuances across multiple languages. The experimental results show that these models, particularly multilingual variants such as mBERT and XLM-RoBERTa, achieve high accuracy in cross-lingual sentiment classification tasks. The use of pre-trained models significantly reduces the need for extensive language-specific resources, making the solution scalable and efficient. Overall, this work highlights the potential of transformer models in enabling robust and inclusive sentiment analysis systems for diverse, global audiences.

Chapter 6: Future work and Scope

The growing demand for sentiment analysis across languages and platforms offers numerous directions for further development. In the future, this project can be expanded and improved in the following ways:

1. Support for More Languages:
   Extending the model to include low-resource and regional languages to increase global accessibility and applicability.
2. Domain-Specific Customization:
   Fine-tuning models on domain-specific datasets (e.g., medical, legal, or educational content) to improve accuracy in specialized fields.
3. Real-Time Sentiment Analysis:
   Implementing the system in real-time applications such as social media monitoring, customer support, or market analysis for timely feedback.
4. Multimodal Sentiment Analysis:
   Combining textual data with images, voice, or video to improve sentiment understanding from diverse sources.
5. Improved Model Interpretability:
   Developing tools for explaining model predictions to build user trust and ensure responsible AI usage.
6. Low-Resource Training Techniques:
   Exploring few-shot and zero-shot learning methods to reduce dependency on large annotated datasets for each new language.
7. Deployment and Scalability:
   Creating lightweight, deployable versions of the model for use in mobile apps, browser extensions, or enterprise tools.

References

[1]. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. Proceedings of NAACL-HLT. https://arxiv.org/abs/1810.04805

[2]. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. https://arxiv.org/abs/1907.11692

[3]. Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). *DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter*. https://arxiv.org/abs/1910.01108

[4]. Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2020). *ALBERT: A Lite BERT for Self-supervised Learning of Language Representations*. https://arxiv.org/abs/1909.11942

[5]. Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., & Le, Q. V. (2020). *XLNet: Generalized Autoregressive Pretraining for Language Understanding*. https://arxiv.org/abs/1906.08237

[6]. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P. J. (2020). *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer*. Journal of Machine Learning Research, 21(140), 1-67. https://arxiv.org/abs/1910.10683

[7]. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., ... & Stoyanov, V. (2020). *Unsupervised Cross-lingual Representation Learning at Scale*. In Proceedings of ACL. https://arxiv.org/abs/1911.02116

[8]. Amazon Web Services. (2020). *Multilingual Amazon Reviews Corpus (MARC)*. https://registry.opendata.aws/amazon-reviews-ml/

[9]. Pires, T., Schlinger, E., & Garrette, D. (2019). *How Multilingual is Multilingual BERT?* In Proceedings of ACL. https://arxiv.org/abs/1906.01502

[10]. Koto, F., Lauw, H. W., & Baldwin, T. (2021). *IndoBERTweet: A Pretrained Language Model for Indonesian Twitter with Effective Domain-Specific Tokenization*. https://arxiv.org/abs/2101.02141

[11]. Balahur, A., Turchi, M., & Steinberger, R. (2012). *Multilingual Sentiment Analysis using Machine Translation?*. In Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis. https://aclanthology.org/W12-3716/

[12]. Mozetic, I., Smailović, J., & Grčar, M. (2016). *Multilingual Twitter Sentiment Classification: A Machine Translation Approach*. PLoS ONE, 11(5), e0155036. https://doi.org/10.1371/journal.pone.0155036

[13]. Akhtar, M. S., Gupta, P., Ekbal, A., & Bhattacharyya, P. (2020). *Multilingual Sentiment Analysis: State of the Art and Independent Comparison of Techniques*. Cognitive Computation, 12, 111–135. https://doi.org/10.1007/s12559-019-09674-