**A PROJECT REPORT**

**On**

**NLP Transformer-based Models**

**(Multilingual Sentiment Analysis)**

**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-Based Models for Multilingual Sentiment Analysis"* 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**

|  |  |
| --- | --- |
| **DESCRIPTION** | **PAGE NUMBER** |
| DECLARATION | **4** |
| ACKNOWLEDGEMENTS | **5** |
| LIST OF FIGURES | **6** |
| ABBREVIATIONS | **6** |
| ABSTRACT | **7** |
| CHAPTER-1: Introduction | **8** |
| CHAPTER-2: Literature Review | **8** |
| CHAPTER-3: Methodology | **9** |
| CHAPTER-4: Result | **10** |
| CHAPTER-5: Challenge | **14** |
| CHAPTER-6: Future Work and Scope | **15** |
| CHAPTER-7: Conclusion | **16** |
| REFERENCES | **17** |

**DECLARATION**

I declare that this project report titled **“Transformer-Based Models for Multilingual Sentiment Analysis”** 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**

|  |  |  |
| --- | --- | --- |
| **S. No** | **Figure name** | **Page Number** |
| **1** | Architecture of Transformer Model |  |
| **2** | Sentiment Analysis Classification Levels |  |
| **3** | Workflow of Multilingual Sentiment Analysis Pipeline |  |
| **4.** | Tokenization Process using XLM-RoBERTa Tokenizer |  |
| **5.** | Training and Validation Accuracy Curve for mBERT |  |
| **6.** | Accuracy Comparison across Languages |  |
| **7.** | Zero-shot Sentiment Transfer from English to French |  |

**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-Based Models for Multilingual Sentiment Analysis " 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. Data Collection and Preprocessing

Datasets were sourced from publicly available multilingual sentiment corpora, such as the Multilingual Amazon Reviews and Twitter sentiment datasets. Each dataset was labeled with sentiment classes (positive, negative, neutral). Basic preprocessing was applied:

* Text normalization (lowercasing, punctuation removal)
* Language-specific filtering
* Tokenization using the SentencePiece (for XLM-R) and WordPiece (for mBERT) tokenizers
* Ensured balance across sentiment classes for fair evaluation where possible

2. Model Selection

Two state-of-the-art multilingual transformer models were selected:

* mBERT (Multilingual BERT): Pretrained on Wikipedia dumps of 104 languages.
* XLM-R (XLM-RoBERTa): Trained on 2.5TB of CommonCrawl data in 100 languages with improved pretraining objectives.

Both models use Transformer encoder architecture, enabling contextual embeddings that generalize across languages without requiring translation.

3. Fine-Tuning Procedure

Each model was fine-tuned on the English sentiment dataset using a classification head (a linear layer with softmax) added to the final hidden state of the [CLS] token. Training configuration:

* Optimizer: AdamW
* Learning rate: 2e-5
* Epochs: 3
* Batch size: 16
* Loss function: Cross-entropy loss

After fine-tuning on English, the models were evaluated zero-shot on the non-English datasets to test cross-lingual transferability.

4. Evaluation Metrics

Model performance was evaluated using:

* Accuracy: Overall correct predictions across all sentiment classes
* F1-score: To handle class imbalance, particularly in neutral/negative categories
* Confusion matrices: To analyze misclassification patterns across languages

5. Languages and Testing Setup

The models were tested on:

* High-resource languages: English (EN), Spanish (ES), French (FR)
* Medium-resource: German (DE)
* Low-resource with complex script: Japanese (JA)

This setup helps analyze the impact of resource availability, linguistic distance, and script diversity on model performance.

**Chapter 4: Results**

4.1 Dataset Analysis and Sentiment Distribution

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:

A close-up of a computer screen

Description automatically generated

4.2 Validation Dataset Evaluation

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:

A close-up of a text

Description automatically generated

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:

A screenshot of a data

Description automatically generated

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:

A screenshot of a computer

Description automatically generated

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.

A pie chart with numbers and percentages

Description automatically generated

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

A graph of a bar chart

Description automatically generated with medium confidence

**Chapter 5: Challenges**

Transformer-based models such as Multilingual BERT (mBERT) and XLM-RoBERTa (XLM-R) have advanced the field of multilingual sentiment analysis. However, despite their powerful architecture and training on massive multilingual corpora, several key challenges remain when applying these models to sentiment tasks across diverse languages.

1. Semantic and Cultural Variability

Sentiment expression is highly language- and culture-dependent. Idiomatic phrases, sarcasm, politeness strategies, and even emotional intensity vary across languages. Transformer models trained on large text corpora may struggle to capture these nuanced expressions, especially in low-resource or less culturally represented languages. For example, a phrase perceived as neutral in one language might carry negative sentiment in another due to cultural context, which models often fail to interpret correctly. The lack of explicit cultural modeling within transformer architectures remains a bottleneck.

2. Tokenization and Script Diversity

Transformers rely on subword tokenization, such as WordPiece or SentencePiece, which affects how inputs are segmented and encoded. Languages with complex scripts — like Chinese, Japanese, or Arabic — often do not use spaces between words, making tokenization more error-prone. Japanese, for instance, uses multiple scripts (Kanji, Hiragana, Katakana), and improper segmentation can distort word meanings. Moreover, tokenizers tend to favour frequent subword units from high-resource Latin-script languages, leading to less efficient representations for scripts with fewer training samples.

3. Data Imbalance and Resource Scarcity

Multilingual pretraining does not ensure equal performance across languages, primarily due to data imbalance. Languages like English dominate pretraining corpora, while many low-resource languages are underrepresented. This imbalance leads to skewed model capabilities, where high-resource languages benefit from richer contextual embeddings and fine-tuned parameters, while underrepresented languages suffer from lower accuracy and robustness. This challenge is further magnified during fine-tuning, where task-specific annotated data is often unevenly distributed across languages.

4. Computational Complexity

Transformer models are resource-intensive, requiring significant computational power for training and inference. This presents a practical barrier for deploying multilingual sentiment systems in low-resource settings or on-device applications. Larger models like XLM-R, although more accurate, demand more memory and processing time, which may not be feasible in real-time or edge environments.

**Chapter 6: Future Work and Scope**

As transformer-based models like mBERT and XLM-R have significantly improved multilingual sentiment analysis, ongoing research is exploring ways to address current limitations and expand their applicability. The future scope for this domain spans improvements in model architecture, resource development, cross-cultural understanding, and deployment strategies.

1. Incorporating Cultural and Contextual Awareness

Current models lack explicit understanding of cultural context, which plays a crucial role in sentiment expression. Future research could focus on integrating external knowledge sources, such as sentiment lexicons, ontologies, or culturally specific datasets, to help models interpret idioms, sarcasm, and region-specific emotional cues. Embedding cultural context into pretraining or fine-tuning stages may improve sentiment prediction across diverse populations.

2. Enhancing Low-Resource Language Support

To achieve true multilingual equity, future work must address the limitations faced by low-resource languages. This includes developing more annotated datasets, leveraging unsupervised and semi-supervised learning techniques, and employing data augmentation methods (e.g., back-translation or synthetic data generation). Cross-lingual knowledge transfer methods, such as adapter-based tuning or multitask learning, could help bridge the performance gap without retraining large models from scratch.

3. Improved Tokenization Techniques

Current subword tokenizers often underperform on non-Latin scripts and morphologically rich languages. Future models might benefit from language-specific tokenization strategies or adaptive tokenization mechanisms that learn to segment text more effectively based on script or morphology. This could reduce fragmentation in encoding and yield better semantic representations across diverse languages.

4. Multimodal and Context-Aware Sentiment Analysis

A promising direction involves extending sentiment analysis beyond text to include multimodal signals, such as speech, facial expressions, or social media images. Combining textual data with metadata or user information can help resolve ambiguities and provide more accurate sentiment predictions. Transformer architectures like Vision-Language Transformers (e.g., CLIP) could be adapted for multilingual, multimodal sentiment tasks.

5. Explainability and Fairness

Finally, future research must prioritize explainable NLP and fairness across languages and demographics. Developing tools to interpret model predictions, detect biases, and ensure equitable performance will be crucial for deploying these systems responsibly, especially in high-stakes domains like healthcare, education, or social media monitoring.

**Chapter 7: Conclusion**

The emergence of transformer-based NLP models, particularly Multilingual BERT (mBERT) and XLM-RoBERTa (XLM-R), has significantly transformed the field of multilingual sentiment analysis. These models have demonstrated a remarkable ability to understand and classify sentiment across diverse languages using shared contextual representations learned from large-scale multilingual corpora. By enabling zero-shot and few-shot learning, they reduce the dependency on labeled data in low-resource languages and make sentiment analysis more accessible globally.

This study examined the performance of mBERT and XLM-R across five typologically diverse languages — English, Spanish, German, French, and Japanese — highlighting their strengths and limitations. Results showed that XLM-R consistently outperformed mBERT, particularly in high-resource languages. However, challenges remained evident in handling syntactically and semantically divergent languages like Japanese, emphasizing the ongoing limitations of even the most advanced multilingual models.

Key challenges identified included the difficulty of interpreting cultural nuances, data imbalance, and tokenization issues in non-Latin scripts. While transformer models are trained on vast datasets, they still struggle to capture localized sentiment expressions, idioms, and sarcasm, particularly when cultural context is absent. Moreover, underrepresentation of many languages during pretraining leads to performance disparities that affect fairness and generalizability.

Despite these challenges, the potential of transformer models in multilingual sentiment analysis is undeniable. Their capacity to learn deep semantic relationships, adapt to new languages with minimal fine-tuning, and generalize across language families offers a scalable and powerful foundation for future research and applications. From business intelligence and customer feedback to social media monitoring and public health communication, these models can drive sentiment understanding in truly global contexts.

Looking ahead, continued improvements in model architecture, training strategies, and cross-lingual transfer learning will be essential to achieving greater inclusivity and accuracy. Incorporating cultural context, improving language-specific tokenization, and ensuring fairness and transparency in predictions will help address current limitations. Furthermore, optimizing these models for real-time deployment will extend their usefulness in practical, real-world settings.

**References**

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. NAACL.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., ... & Stoyanov, V. (2020). *Unsupervised Cross-lingual Representation Learning at Scale*. ACL.

Pires, T., Schlinger, E., & Garrette, D. (2019). *How Multilingual is Multilingual BERT?*. ACL.

Hu, J., Ruder, S., Siddhant, A., Neubig, G., Firat, O., & Johnson, M. (2020). *XTREME: A Massively Multilingual Multi-task Benchmark for Evaluating Cross-lingual Generalization*. ICML.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. arXiv preprint arXiv:1907.11692.

Artetxe, M., & Schwenk, H. (2020). *Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond*. TACL.