SENTIMENT ANALYSIS OF TURKIYE-RELATED NEWS USING MULTILINGUAL DEEP LEARNING MODELS

Ahmet Can Karataş & Erhan Alasar
Department of Computer Engineering
Izmir Institute of Technology
Izmir, Turkiye

ABSTRACT

This project implements a multilingual sentiment analysis system for Türkiye-related news articles. It incorporates data collection, translation, and sentiment classification using deep learning models. Over 2,000 news articles were collected through NewsAPI and RSS feeds, translated into Turkish using MarianMT and DeepL API, and analyzed with multilingual models such as BERT, MiniLM, and Bart Large MNLI. Analysis revealed that zero-shot learning provided consistent sentiment predictions, while few-shot learning showed a limited improvement due to conflicting predictions caused by multimodal language complexities and model-specific biases. This project demonstrates the potential and limitations of multilingual NLP techniques in addressing cross-lingual sentiment analysis challenges.

1 INTRODUCTION

1.1 BACKGROUND AND MOTIVATION

Sentiment analysis is a critical tool for understanding public opinion, especially in geopolitically significant contexts. Türkiye frequently appears in international news, requiring analysis that bridges linguistic and contextual gaps. This project aims to address this by integrating multilingual NLP models with machine translation to classify sentiments effectively.

1.2 OBJECTIVE

The objective of this project is to create a sentiment analysis pipeline for Türkiye-related news articles by leveraging multilingual deep learning models and exploring zero-shot and few-shot learning methods.

2 IMPLEMENTATION

2.1 DATA COLLECTION

- Sources: Articles were collected via NewsAPI and RSS feeds from outlets like BBC, Reuters, and Al Jazeera.
- Dataset Size: Over 2,000 articles, filtered for keywords such as ”Turkey,” ”Ankara,” and ”Istanbul.”
- Preprocessing: Text was cleaned to remove HTML tags, special characters, and irrelevant information.

2.2 TRANSLATION

- Models Used:
  - MarianMT Model: Used for English-to-Turkish translation.

  - DeepL API: Enhanced the quality of translations for complex phrases.
- Outcome: The dataset was successfully translated into Turkish while retaining contextual accuracy.

2.3 SENTIMENT CLASSIFICATION

- Models Used:
  - BERT Base Multilingual Uncased.
  - MiniLM Multilingual.
  - Bart Large MNLI.
- Approaches:
  - Zero-Shot Learning: Models classified sentiment without task-specific training, yielding consistent predictions.
  - Few-Shot Learning: Curated examples were introduced, but conflicting predictions limited its effectiveness.

3 EXPERIMENTAL WORK

3.1 SETUP

- Environment: Python, Hugging Face Transformers, and Google Colab.
- Dataset: Preprocessed and translated dataset of 3,000+ articles.
- Evaluation Metrics: Consistency and accuracy of predictions across models.

3.2 RESULTS

- Zero-Shot Learning:
  - The BERT model predicted 1,373 negative and 415 positive sentiments.
  - MiniLM identified 1,381 positive and 407 negative sentiments.
  - Bart Large MNLI leaned toward 1,491 negative and 297 positive sentiments.
- Few-Shot Learning:
  - Predictions varied significantly: e.g., Bart Large MNLI predicted 1,786 positive sentiments but only 3 negative ones.
  - High conflict rates (1,699 predictions) indicated limited agreement across models.

3.3 ERROR ANALYSIS

- Translation Impact: Misinterpretation of idiomatic expressions led to classification errors.
- Model Conflicts: Few-shot learning amplified inconsistencies among models due to varying sensitivities to input prompts.
- Data Quality: Multimodal language complexities affected the reliability of both zero-shot and few-shot predictions.

4 CONCLUSION AND DISCUSSION

4.1 KEY FINDINGS

- Zero-shot learning provided robust baseline predictions, showcasing the adaptability of multilingual NLP models.
- Few-shot learning failed to enhance accuracy significantly, likely due to conflicting predictions and training biases.

4.2 IMPLICATIONS

This project highlights the utility of multilingual NLP tools in cross-lingual sentiment analysis but underscores the need for more refined methodologies to handle multimodal linguistic challenges.

4.3 FUTURE WORK

- Fine-tune models for Türkiye-specific contexts to address linguistic nuances.
- Expand the dataset with additional sources and diverse topics.
- Develop ensemble methods to resolve conflicting predictions across models.

5 REFERENCES

- NewsAPI: https://newsapi.org/
- BBC RSS Feeds: https://www.bbc.com/news/10628494
- Reuters RSS Feeds: https://www.reutersagency.com/en/reutersbest/rss-feeds/
- Al Jazeera RSS Feeds: https://www.aljazeera.com/xml/rss/all.xml
- MarianMT: https://huggingface.co/Helsinki-NLP/opus-mt-en-tr
- DeepL API: https://www.deepl.com/pro-api
- BERT Base Multilingual: https://huggingface.co/bert-base-multilingual-uncased
- MiniLM Multilingual: https://huggingface.co/nreimers/MiniLMv2-L6-H384-distilled-from-XLMR-Large
- Bart Large MNLI: https://huggingface.co/facebook/bart-large-mnli