Natural Language Processing Project

Dataset details:

The dataset chosen for this particular assignment is an Amazon multilingual reviews dataset. This dataset was download from Kaggle and contains a total of 30,000 reviews. It covers several languages other than English like Deutsch, Spanish, Japanese, French and Chinese.
The set contains 8 columns, each conveying a specific information regarding the order. There is the review ID, product ID, reviewer ID, stars, review title, review body, language and product category. The stars for the products range from 1 to 5 (1 being the least and 5 being the most) and the product category is also very varied as it contains items from domains like home, kitchen, industrial supplies, pet products, wireless, drugstore, automotive, digital ebook purchase etc.
All these features make this dataset versatile and a good fit for this assignment.

Models Used and Rationale:

The model used for monolingual task is “distilbert-base-uncased”. It is a lightweight English only BERT variant for efficient fine tuning. It is suitable for English sentiment tasks, and is expected to perform best on English subset.
To handle the multilingual task, “distilbert-base-multilingual-cased” model was employed. It has been trained on 104 languages and it was chosen to evaluate generalization across non-English data. It is expected to generalize better but it might trade off English-specific accuracy. Comparing the monolingual and multilingual models helps us to assess cross-lingual transfer and performance trade-offs in multilingual NLP.

Training setup and Hyperparameters:

Both the models were fine-tuned using Hugging Face “Trainer” API on the same training to test split which was of 60-40. The other details are as follows:

Learning Rate = 2e-5
Optimizer used = AdamW (as it is default in Trainer)
Epochs ran = 1
Batch size = 8
Weight decay = 0.01
Evaluation Strategy = To per epoch
Tokenization = distilbert-base-uncased (for monolingual)
distilbert-base-multilingual-cased (for multilingual)

Performance Comparison and Analysis:

Here are the details regarding the overall results obtained after running the code:

For the monolingual model,

Training loss = 1.441100
Validation loss = 1.358682
Accuracy = 0.363417
F1 = 0.368255

Evaluation runtime = 172.3781 seconds
Evaluation samples per second = 69.614
Evaluation steps per second = 8.702

Whereas for the multilingual model,

Training loss = 1.313200
Validation loss = 1.214602
Accuracy = 0.465917
F1 = 0.461538
Evaluation runtime = 172.3698 seconds
Evaluation samples per second = 69.618
Evaluation steps per second = 8.702

Metrics per Language:

First, for the monolingual model:

Deutsch:

Validation loss = 1.333480715751648
Accuracy = 0.39568345323741005
F1 = 0.3866444412688506
Evaluation runtime = 27.8738
Evaluation samples per second = 69.815
Evaluation steps per second = 8.754

English:

Validation loss = 1.073514699935913
Accuracy = 0.5183913683178029
F1 = 0.5088215782657002
Evaluation runtime = 28.9292
Evaluation samples per second = 70.482
Evaluation steps per second = 8.815

Spanish:

Validation loss = 1.2600048780441284
Accuracy = 0.42525900345337936
F1 = 0.41951629299824317
Evaluation runtime = 28.6495
Evaluation samples per second = 70.752
Evaluation steps per second = 8.866

French:

Validation loss = 1.2771832942962646
Accuracy = 0.43921960980490243
F1 = 0.4343123853376648
Evaluation runtime = 28.3699
Evaluation samples per second = 70.462
Evaluation steps per second = 8.812

Japanese:

Validation loss = 1.6072771549224854
Accuracy = 0.2049221496735309
F1 = 0.09617664760374081
Evaluation runtime = 27.81
Evaluation samples per second = 71.593
Evaluation steps per second = 8.954

Chinese:

Validation loss = 1.6081726551055908
Accuracy = 0.1931931931931932
F1 = 0.08138635337498518
Evaluation runtime = 27.7871
Evaluation samples per second = 71.904
Evaluation steps per second = 8.997

Now, for the multilingual model:

Deutsch:

Validation loss = 1.1864467859268188
Accuracy = 0.4784172661870504
F1 = 0.467475685514327
Evaluation runtime = 27.1512
Evaluation samples per second = 71.673
Evaluation steps per second = 8.987

English:

Validation loss = 1.1609491109848022
Accuracy = 0.48406081412457086
F1 = 0.47915005836283353
Evaluation runtime = 28.5028
Evaluation samples per second = 71.537
Evaluation steps per second = 8.946

Spanish:

Validation loss = 1.1800543069839478
Accuracy = 0.47755303404045385
F1 = 0.47941078795840336
Evaluation runtime = 28.3477
Evaluation samples per second = 71.505
Evaluation steps per second = 8.96

French:

Validation loss = 1.1932499408721924
Accuracy = 0.471735867933967
F1 = 0.46396795741264457

Evaluation runtime = 27.7865
Evaluation samples per second = 71.941
Evaluation steps per second = 8.997

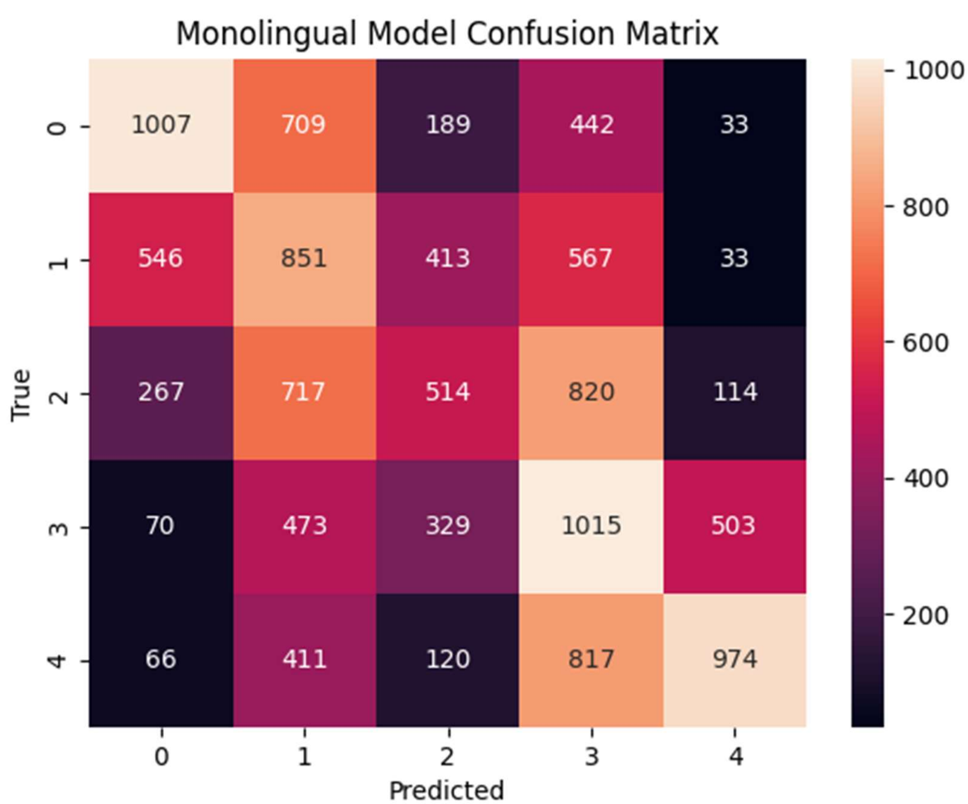
Japanese:

Validation loss = 1.2846189737319946
Accuracy = 0.42591662481165243
F1 = 0.4228613870107223
Evaluation runtime = 27.7086
Evaluation samples per second = 71.855
Evaluation steps per second = 8.986

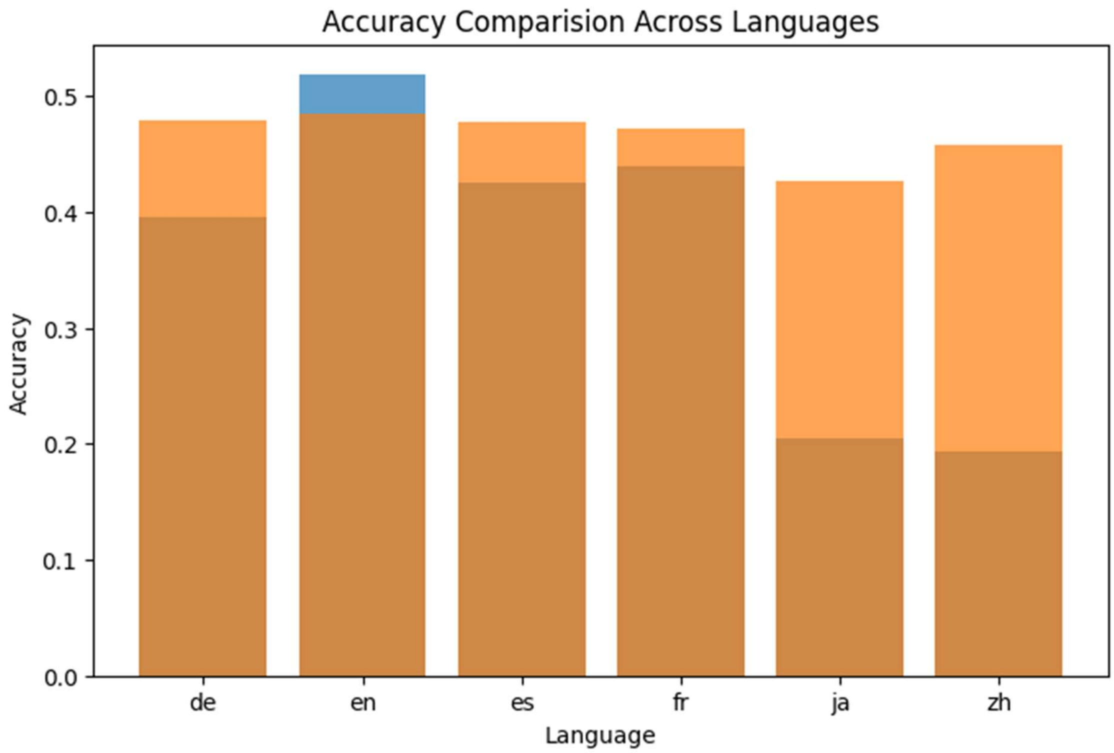
Chinese:

Validation loss = 1.2834198474884033
Accuracy = 0.4574574574574575
F1 = 0.45366420854993905
Evaluation runtime = 28.1189
Evaluation samples per second = 71.055
Evaluation steps per second = 8.891

Monolingual Model Confusion Matrix diagram:



Bar graph based on accuracy across different languages:



Note: de = Deutsch, en = English, es = Spanish, fr = French, ja = Japanese, zh = Chinese

Key insights on multilingual generalization:

The multilingual model handles unseen languages, even with minimal fine-tuning data per language. Though it has a slightly lower English accuracy but it has a better average multilingual performance (especially for Deutsch, Japanese and Chinese), hence this can be considered as a fair trade-off. It can be said that the multilingual model requires more parameters and longer training time but for mixed language user reviews, multilingual models offer superior real-world applicability.

Conclusion:

Monolingual distilbert excels primarily at English sentiment detection while Multilingual distilbert is a lot more efficient at generalizing across languages with minimal loss on English. Combining language specific and multilingual approaches can balance both performance and generalization however for constraints involving time, it is better to use a multilingual model as it is a lot more versatile and does not impact English understanding by a large margin.