Jazba-e-Urdu: Transformer-Based Sentiment Analysis for Urdu Tweets

NLP Semester Project

Hanzala Khalid
2021-SE-05

Ahtshtam-ul-Haq
2021-SE-25

Abdullah Shahid
2021-SE-32

Abstract

This project implements an end-to-end sentiment analysis system for Urdu tweets using a pretrained multilingual transformer. The pipeline developed in Google Colab covers data ingestion from Excel files, thorough text cleaning and sentiment mapping, dataset balancing, tokenization, and fine-tuning of the "xlm-roberta-base" model for classification into negative, neutral, and positive categories. Evaluation on a balanced test set shows high performance with approximately 98% accuracy and F1-score. Furthermore, the model is deployed on Hugging Face Spaces using Gradio, providing an interactive web interface for real-time sentiment prediction. This work lays a robust foundation for sentiment analysis in low-resource languages while illustrating best practices for model deployment.

1. Introduction

Sentiment analysis is crucial for understanding the ideas and emotions conveyed in user-generated content, such social media or customer evaluations. Although a lot of research has been done on resource-rich languages like English, low-resource languages like Urdu are still not widely explored. Because of its rich grammar and right-to-left script, Urdu provides distinct challenges requiring need the application of particular modeling and preprocessing methods.

We develop an Urdu sentiment analysis system in this project. After data extraction from an Excel dataset of Urdu tweets, we map raw sentiment labels into three predefined categories negative, neutral, and positive using cleaning and normalizing methods.

In this project, we develop a sentiment analysis system for Urdu. We start with data extraction from an Excel dataset of Urdu tweets, apply cleaning and normalization techniques and map raw sentiment labels into three standardized categories: negative, neutral, and positive. The processed data is then passed through tokenization techniques and converted into a format suitable for deep learning. We have used multilingual transformer (XLM-Roberta) to perform sentiment classification and achieve high evaluation metrics. Finally, we deploy the model on Hugging Face Spaces, providing a user-friendly Gradio web interface for real time inference.

2. Related Work

Lexicon based techniques and statistical models remained the main techniques of early sentiment analysis methodologies (Pang & Lee, 2008). When machine learning emerged, machine learning models' methods like Naïve Bayes and Support Vector Machines became popular, but they required a lot of features to develop. A new era of deep learning was brought in by the automated feature collection and processing of sequential data by CNNs (Zhang et al., 2015) and traditional neural networks (Kim, 2014). Transformer-based models, like as BERT (Devlin et al., 2019), have recently produced state of the art outcomes for NLP tasks.
These capabilities are extended to languages with fewer resources, such as Urdu, using multilingual models like XLM-Roberta (Conneau et al., 2020). Meanwhile, frameworks like Hugging Face's Transformers and Datasets have made it easier to create, train, and implement NLP models. By focusing on a full pipeline from data pretreatment to interactive deployment specifically for Urdu sentiment analysis, our work builds on existing developments and makes even more contributions.

3. Proposed Methodology

3.1. Data Acquisition and Preprocessing

● Dataset Loading

By Mounting Google Drive in a Colab notebook and loading an Excel file with two sheets of Urdu tweets is the first step in the project. To produce a single dataset, the sheets are concatenated.

● Data Cleaning

Text cleaning functions remove extra spaces and unwanted characters to ensure uniformity. This step is crucial given the challenges in processing Urdu's script and punctuation.

● Sentiment Mapping

Using keyword-based mapping, raw sentiment categories are simplified to three classes: negative, neutral, and positive. Consistent model training needs this standardization

● Dataset Balancing

To avoid class imbalance, a balanced subset is sampled equally from each sentiment category. The balanced data is visualized via bar plots and then converted into a Hugging Face Dataset.

● Dataset Splitting

For the purpose of impartial evaluation during training, the dataset is divided into training (80%), validation (10%), and test (10%) sets.

3.2. Tokenization and Feature Extraction

● Tokenizer Setup

The "xlm-roberta-base" tokenizer is loaded to handle the Urdu text. A custom tokenization function applies truncation and padding (max_length=128) to standardize input sizes across the dataset.

- **Embedding Representation**

The pretrained transformer, which naturally captures syntactic and semantic information in Urdu, is used to turn tokens into dense embeddings.

3.3. Model Architecture and Training

- **Model Configuration**

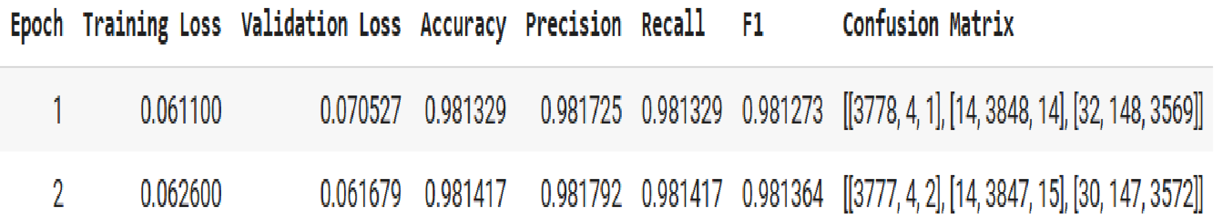
The pretrained XLM-Roberta model is fine-tuned for sequence classification with three output labels. The model's architecture is adapted by configuring the final classification layer to predict negative, neutral, or positive sentiment.

- **Training Parameters**

Training is conducted using the Hugging Face Trainer API with a learning rate of 2e-5, a batch size of 16, and two training epochs. Mixed precision (bf16) is enabled to optimize resource usage. Regular evaluations are performed using metrics such as accuracy, precision, recall, F1-score, and confusion matrices.

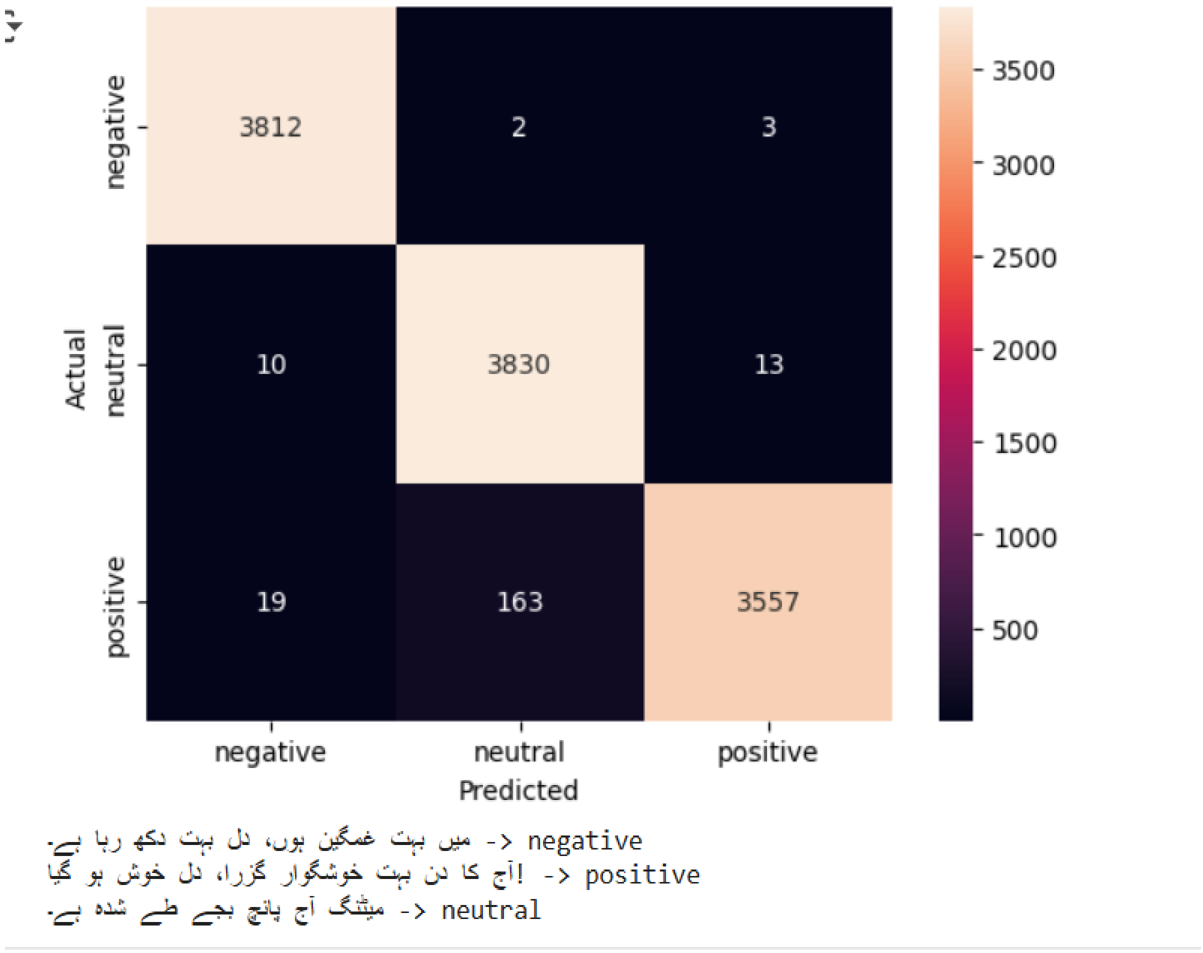
- **Evaluation**

High performance is obtained from post-training evaluation on the test set (about 98% accuracy and F1-score). Insights into model performance across sentiment categories are provided by detailed classification reports and confusion matrix representations.



Epoch	Training Loss	Validation Loss	Accuracy	Precision	Recall	F1	Confusion Matrix
1	0.061100	0.070527	0.981329	0.981725	0.981329	0.981273	[[3778, 4, 1], [14, 3848, 14], [32, 148, 3569]]
2	0.062600	0.061679	0.981417	0.981792	0.981417	0.981364	[[3777, 4, 2], [14, 3847, 15], [30, 147, 3572]]

Figure 1 - Evaluation Metrics



```
میں بہت غمگین ہوں، دل بہت دکھ رہا ہے۔ -> negative
آج کا دن بہت خوشگوار گزرا، دل خوش ہو گیا! -> positive
میٹنگ آج پانچ بجے طے شدہ ہے۔ -> neutral
```

Figure 2 - Confusion Matrix

3.4. Model Deployment on Hugging Face Spaces

- **Repository and Model Upload**

After training, the model and tokenizer are saved and uploaded to a Hugging Face repository. This facilitates version control and sharing with the community.

- **Deployment with Gradio**

A Gradio app (app.py) is developed to create an interactive web interface. The app loads the model from the repository and defines a prediction function that cleans input text, tokenizes it, and returns the predicted sentiment label. Example inputs are provided to guide users, and the interface is designed to be user-friendly.

- **Web Interface Interaction**

The Gradio interface includes textboxes for input and output, a descriptive title, and examples demonstrating usage. This deployment enables real-time sentiment prediction and broad accessibility via the Hugging Face Spaces platform.

4. Limitations

While the project shows promising results, while several limitations remain:

- **Dataset Limitations**

The dataset, though balanced for the experiment, may not fully capture the diverse expressions of sentiment in Urdu. Broader datasets could help improve model robustness.

- **Preprocessing Challenges**

However useful, custom text cleaning and tokenization for Urdu might not take into account all language complexities, which could compromise accuracy in edge situations.

- **Model Complexity**

The current approach uses a fine-tuned transformer with a straightforward classifier. More sophisticated architectures or ensemble methods might further enhance performance, particularly in noisy real-world environments.

- **Computational Resources**

Training in a Colab environment limits extensive hyperparameter tuning and experimentation with larger datasets or more complex models.

- **Deployment Considerations**

The deployed model on Hugging Face Spaces uses a publicly available interface that, while user-friendly, might require additional security and scalability measures for production-level usage.

5. Conclusions

This project successfully demonstrates a complete pipeline for Urdu sentiment analysis, from data preprocessing and fine-tuning of a multilingual transformer model to interactive deployment on Hugging Face Spaces. The high evaluation metrics underscore the effectiveness of the approach, while the deployment via Gradio makes the model accessible for real-time inference. Future work can focus on expanding the dataset, enhancing preprocessing routines, exploring advanced architectures, and optimizing the deployment framework for broader applications in low-resource language processing.

6. References

1. Ghafoor, A., Imran, A. S., Daudpota, S. M., Kastrati, Z., Shaikh, S., & Batra, R. (n.d.). *SentiUrdu-1M: A large-scale tweet dataset for Urdu text sentiment analysis using weakly supervised learning*. Retrieved from SentiUrdu-1M.
2. Pang, B., & Lee, L. (2008). *Opinion Mining and Sentiment Analysis*. Foundations and Trends in Information Retrieval. Retrieved from Opinion Mining and Sentiment Analysis.
3. Sukhbaatar, A., Szlam, A., Weston, J., & Fergus, R. (2015). *End-to-End Memory Networks*. NYU. Retrieved from End-to-End Memory Networks.
4. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. NAACL-HLT. Retrieved from BERT Pre-training of Deep Bidirectional Transformers.
5. Conneau, A., et al. (2020). *Unsupervised Cross-lingual Representation Learning at Scale*. ACL. Retrieved from Unsupervised Cross-lingual Representation.
6. Vaswani, A., et al. (2017). *Attention Is All You Need*. NIPS. Retrieved from Attention is All you Need.
7. Kim, Y. (2014). *Convolutional Neural Networks for Sentence Classification*. EMNLP. Retrieved from CNN for Sentence Classification.
8. Zhang, X., Zhao, J., & LeCun, Y. (2015). *Character-level Convolutional Networks for Text Classification*. NIPS. Retrieved from Character-level Convolutional Networks.
9. Wolf, T., et al. (2020). *Transformers: State-of-the-art Natural Language Processing*. EMNLP. Retrieved from Transformers.
10. Pedregosa, F., et al. (2011). *Scikit-learn: Machine Learning in Python*. Journal of Machine Learning Research. Retrieved from Journal of Machine Learning Research.