

NLP

Assignment # 1



September 29, 2024

20F-0340

9A

# Challenges Faced

* The first and foremost challenge encountered was dealing with the unique characteristics of the Urdu language. Urdu, being a right-to-left script, presents various challenges such as the absence of spaces between words in some cases and compound words. Tokenization, stemming, and lemmatization were particularly difficult because traditional NLP libraries like NLTK, spaCy, and others are designed with languages like English.
* The morphological richness of Urdu (due to its inflection and the use of suffixes) made it challenging to develop an effective stemming and lemmatization process.
* Unlike English, there is a lack of high-quality, readily available resources for Urdu NLP tasks. Pre-trained language models for Urdu sentiment analysis are not as mature or accessible as they are for English. This limited the potential of using advanced techniques such as transfer learning or pre-trained embeddings without investing significant time into creating or fine-tuning models.
* Another significant challenge was handling the noisy nature of social media data. Urdu posts often mix Roman Urdu, emojis, slang, abbreviations, and other non-standard textual elements, making it difficult to maintain a clean dataset. This noise required heavy pre-processing, including removing irrelevant symbols, handling misspellings, and normalizing text.

# Optimization

* To enhance the accuracy of text processing, leveraging advanced tokenizers and developing a more sophisticated lemmatization pipeline is crucial. As an improvement, I would explore utilizing models like LughaatNLP or other Urdu-specific libraries in combination with custom rules to handle the complexities of Urdu morphology more efficiently.
* While traditional models like SVM and Naive Bayes were useful, deep learning techniques such as LSTMs (Long Short-Term Memory) or GRUs (Gated Recurrent Units) can better capture the sequence and context in Urdu text. Recurrent neural networks (RNNs) would allow the model to better understand the dependencies between words, especially in complex sentences.
* To address the challenge of limited data, data augmentation techniques such as back-translation (translating Urdu to another language and back) or synthetic data generation could be employed. This would help increase the diversity of the training set and improve the model's generalizability.