# 21F-9138 Muavia Abdul Moiz NLP-9A

## Problems faced

### Stop word removal and Punctuation

A significant obstacle was the absence of robust libraries for removing stop words in Urdu. Consequently, we had to compile a comprehensive list of Urdu stop words.

Urdu punctuation marks differ from their English counterparts, featuring symbols like the reverse question mark (؟) and unique comma variations. Existing libraries did not adequately address these differences, necessitating the creation of a specialized punctuation removal tool.

### Lemmatization problem

Lemmatization is a critical step in reducing words to their base forms, but the lack of built-in Urdu lemmatization libraries posed a major hurdle. We overcame this by creating a custom **Urdu lemmatization dictionary**, which allowed us to reduce many words to their root forms. However, the limited size and scope of our dictionary meant that not all word variations were covered. As a result, certain inflected forms may have been missed during pre-processing, leading to incomplete normalization and potential loss of sentiment-bearing information

### Stemming

Stemming in Urdu was similarly challenging due to the absence of ready-made solutions. We developed our own approach by identifying common prefixes and suffixes in Urdu words, which allowed us to reduce some words to their stems. However, the complexity and diversity of Urdu morphology made this task intricate, and some words may not have been stemmed correctly. A more advanced stemming approach, possibly using rule-based or machine learning techniques, could improve the accuracy of text normalization.

### Word 2 Vec Model Performance

Despite pre-processing efforts, residual stop words impacted the performance of our Word2Vec **model**. The presence of these stop words caused contextual variations in the word vectors, leading to inaccurate similarity measures. Additionally, similar words were mapped inconsistently across different documents, diminishing the model's ability to capture true semantic relationships. This issue highlighted the need for more effective stop word removal and better context handling to improve the quality of the word embeddings

## Areas for Improvement

Although the model performed reasonably well with an accuracy of 82.5%, there are several areas where improvements could significantly enhance its performance:

### 1. Handling Implicit Sentiment

**- Current Challenge:** The model struggled with posts that contained sarcasm or subtle emotional cues. For example, sarcastic phrases like "واہ، بہت اچھا!" were misclassified as positive, even though the sentiment was negative.

- **Improvement Strategy:** Incorporating more context-aware models, such as BERT or UrduBERT, could help detect sarcasm by understanding the tone and context of the entire sentence. Training the model on a dataset with labelled sarcastic posts could also improve its ability to recognize this form of implicit sentiment.

## 2. Colloquial Language and Phonetic Spellings

- **Current Challenge:** Social media users frequently employ colloquial language, slang, and phonetic spellings, which the model struggled to interpret accurately. For instance, "شکریہ" could be spelled as "شکریا," and the model would fail to identify both forms as the same word.

- **Improvement Strategy**: The use of phonetic spell-checkers or training on more diverse datasets containing colloquial language and slang could help the model recognize alternative spellings. Additionally, building a custom Urdu slang dictionary could improve classification accuracy for colloquial expressions.

## 3. Emoji Sentiment Integration

- **Current Challenge:** The model currently ignores emojis, which are widely used on social media to express emotions. For example, a post containing the emoji "😥" was classified as neutral, even though the emoji implied a negative sentiment.

- **Improvement Strategy:** Integrating an emoji-to-sentiment dictionary into the model could allow it to interpret emojis as sentiment-bearing tokens. By translating emojis into positive, negative, or neutral sentiments, the model would better capture the emotional tone of posts.

## 4. Handling Hashtags

- **Current Challenge:** During preprocessing, hashtags were removed to reduce noise, but in some cases, they carried important context. Hashtags like "#برا" (bad) often provided sentiment that the model missed.

- **Improvement Strategy:** Instead of removing hashtags entirely, extracting meaningful hashtags and interpreting them as sentiment clues could enhance the model’s ability to capture the full sentiment of posts.