# Development of an Automatic Sentiment Analysis Tool for Urdu Text on Social Media Platforms

**Abdullah Basit | 20I-0623**

**7D – NLP Assignment 1**

---

1- Main.py file is attached alongside this report

2- Text Preprocessing Results, Feature Extraction Results, N-Gram Analysis, and Sentiment Classification Model is displayed below in the complete code output

```
C:\Users\alpha\PycharmProjects\NLPA1\.venv\Scripts\python.exe C:\Users\alpha\PycharmProjects\NLPA1\main.py
[nltk_data] Downloading package punkt to
[nltk_data]     C:\Users\alpha\AppData\Roaming\nltk_data...
[nltk_data]   Package punkt is already up-to-date!
Dataset loaded successfully. Here's a preview:
                                      urdu_text  ...  Unnamed: 7
NaN         ...  ...ہو لینے دے میری شادی فسادن ٹھ 😂😂_xDD23__xD83E_  0
NaN         ...  ...چل مہمانوں میں کھانا سرو کر چڑیل چاچی نوں دسدی  1
NaN         ...  ...کامران خان آپکی دن بھریہ زمہ داری لگائی گئی اپ  2
NaN         ...  😍 نہیں بائین                                      3
NaN         ...  ...مراد علی شاہ کے بھیس میں ڈی جی آئی ایس آئی `` 4

[5 rows x 8 columns]

Stopwords loaded successfully. Number of stopwords: 515

Text cleaning completed.
Stopword removal completed.
Short posts filtered.
Stemming completed.
Lemmatization completed.

Preprocessing completed. Here's a preview of the preprocessed data:
                                      urdu_text  ...                                      cleaned_text
🤣 لین میر شاد فسادن ٹھیک کوج نہ چاہی xDD23xD83E    ...  ...ہو لینے دے میری شادی فسادن ٹھ 😂😂_xDD23__xD83E_  0
چل مہمان م کھان سرو کر چڑیل چاچ ن دسد آں م         ...  ...چل مہمانوں میں کھانا سرو کر چڑیل چاچی نوں دسدی  1
...کامر خ آپک دن بھری زم دار لگائ اپوزیشن کر کش ا  ...  ...کامران خان آپکی دن بھریہ زمہ داری لگائی گئی اپ  2
مراد عل شا بھیس م ڈ ج ایس حامد میر                ...  ...مراد علی شاہ کے بھیس میں ڈی جی آئی ایس آئی `` 3
قابل اعتبار قاتل اعتبار                            ...  🔥💔 قابل اعتبار ہی اکثر قاتل اعتبار ہوتے ہیں        4

[5 rows x 9 columns]
```

```
Tokenization completed.

Sample tokenized text:
                                        cleaned_text                                                tokens
[لین, میر, شاد, فسادن, ٹھیک, کوچ, نہ, چاہی]        🤣 لین میر شاد فسادن ٹھیک کوچ نہ چاہی xDD23xD83E    0
[چل, مہمان, م, کھان, سرو, کر, چڑیل, چاچ, ن, دس...  چل مہمان م کھان سرو کر چڑیل چاچ ن دسد آں م          1
[کامر, خ, آپک, دن, بھری, زم, دار, لگائ, اپوزیش...  کامر خ آپک دن بھری زم دار لگائ اپوزیشن کر کش ۱۰۰۰ 2
[مراد, عل, شا, بھیس, م, ڈ, چ, ایس, حامد, میر]      مراد عل شا بھیس م ڈ چ ایس حامد میر               3
[قابل, اعتبار, قاتل, اعتبار]                        قابل اعتبار قاتل اعتبار                          4

TF-IDF computation completed.

Top 10 words with the highest TF-IDF scores:
0.0495 :م
0.0419 :ن
0.0407 :س
0.0388 :کر
0.0383 :کو
0.0382 :ک
0.0364 :نہ
0.0330 :بھ
0.0235 :اس
0.0218 :آپ

Word2Vec model training completed.

.not found in the vocabulary 'اچھا' Word

Top 10 unigrams:
7630 :م
6026 :ن
5894 :س
5666 :کو
5329 :ک
5264 :کر
4438 :نہ
```

```
5264 :کر
4438 :نہ
3921 :بھ
2659 :اس
2131 :و

Top 10 bigrams:
495 :عمر خ
442 :نواز شریف
394 :م ن
315 :آپ کو
294 :سندھ پولیس
240 :بھ نہ
225 :ن لیگ
224 :آرم چیف
223 :آپ ن
209 :م بھ

Top 10 trigrams:
120 :صل الل علی
92 :جزاک الل خیر
86 :پ ڈ ایم
75 :ووٹ کو عزت
69 :عمر خ ن
66 :وال کو فالو
66 :فالو فالو بیک
64 :ن جواب دی
62 :لسٹ م شامل
61 :فالورز ک اضاف
```

```
Classification Report:
              precision    recall  f1-score   support

           0       0.75      0.75      0.75      1860
           1       0.76      0.76      0.76      1948

    accuracy                           0.75      3808
   macro avg       0.75      0.75      0.75      3808
weighted avg       0.75      0.75      0.75      3808

Confusion Matrix:
[[1390  470]
 [ 463 1485]]
```

## 3- Reflection:

The development of an NLP pipeline for Urdu sentiment analysis presented several challenges that required careful consideration and innovative solutions. One of the primary obstacles was managing stopwords without losing critical sentiment-bearing words. Standard Urdu stopword lists included essential sentiment words like "خوش" (happy) and "غم" (sadness), which are pivotal for accurate sentiment analysis. To address this, a custom list was created to exclude these sentiment words from the stopword list, ensuring that the emotional context of the text was preserved during preprocessing.

Another significant challenge involved data conversion and text cleaning. The Urdu text data faced issues with encoding and inconsistent formatting, leading to errors in the cleaning functions. Some text entries were not read correctly due to special characters and encoding mismatches. This was resolved by explicitly converting all entries in the `urdu_text` column to string format before applying cleaning operations. This step standardized the data format, allowing for effective cleaning and normalization processes.

Library compatibility issues also arose during the project. Specifically, an error was encountered with scikit-learn's `TfidfVectorizer` when using the `get_feature_names_out()` method, which was not recognized in the installed version of scikit-learn. After investigating, it was discovered that this method was available in newer versions of the library. Updating scikit-learn and adjusting the code to use `get_feature_names()` instead allowed the code to execute correctly and proceed with feature extraction.

After preprocessing—which included cleaning, stopword removal, stemming, lemmatization, and tokenization—a Logistic Regression model was trained using TF-IDF features. The dataset was split into training and testing sets, and the model achieved an accuracy of approximately 75%. The precision and recall for both classes (sarcastic and non-sarcastic) hovered around 75-76%. However, the confusion matrix revealed a substantial number of misclassifications, indicating room for improvement in the model's performance.

To enhance the pipeline, several future optimizations were proposed. Expanding the lemmatization dictionary to include more verb forms, nouns, and common inflections in Urdu could improve text normalization and feature representation. Implementing advanced stemming and tokenization techniques using specialized NLP libraries like UrduHack or Hazm could address the complexities of the Urdu script more effectively. Addressing class imbalance through techniques like SMOTE or adjusting class weights could improve the model's ability to learn from both classes equally.

Further, exploring different feature extraction methods—such as combining TF-IDF with word embeddings from Word2Vec or utilizing advanced models like BERT tailored for Urdu—could capture semantic relationships more effectively. Experimenting with other classification algorithms like Support Vector Machines, Random Forests, or Gradient Boosting Machines,

along with hyperparameter tuning using Grid Search or Random Search, could optimize model performance. Incorporating k-fold cross-validation would provide a more robust estimate of the model's generalizability.

Lastly, integrating contextual and domain-specific knowledge, such as linguistic rules specific to Urdu sarcasm or irony, could enhance the model's ability to detect subtle sentiment expressions.

In conclusion, despite the challenges related to language-specific processing, data handling, and technical compatibility, a functional baseline model was established. The moderate accuracy achieved indicates a solid foundation upon which to build. By focusing on the proposed optimizations, the NLP pipeline can be further refined to achieve higher accuracy and more reliable sentiment classification in Urdu, contributing valuable advancements to sentiment analysis in low-resource languages.