# **Sentimental Analysis**

# **Assignment Report**

# 

# **Submitted to:**

# Ms. Qurat-ul-ain

# **Submitted by:**

# Sohaib Ikram 2022-CS-621

# Department of Computer Science, New Campus

# **University of Engineering and Technology, Lahore**

# Introduction

Sentiment analysis, or opinion mining, is the computational study of people's opinions, sentiments, emotions, and attitudes expressed in text. With the growing digital content in regional languages, sentiment analysis in Urdu has become crucial for understanding user-generated content on social media, video platforms, and forums. This project focuses on building a machine learning model to classify Urdu sentences from drama subtitles into positive or negative sentiments.

# Methodology

We adopted a supervised learning approach using the Naive Bayes classifier, which is well-suited for text classification problems. The methodology includes:

**Text Preprocessing:** Removing Urdu punctuation, whitespace normalization, tokenization, and optional removal of Urdu stopwords.

**Vectorization:** Sentences are transformed into numerical features using TF-IDF (Term Frequency–Inverse Document Frequency).

**Model Training:** A MultinomialNB classifier was trained using a pipeline that combines the preprocessing and vectorization steps.

**Evaluation:** We split the dataset into training and testing subsets and evaluated the model using accuracy, precision, recall, and F1-score.

# Dataset / Development

The dataset consists of **39,473** labeled Urdu sentences extracted from Pakistani drama subtitles. Key features include:

**Drama\_Name**, **Episode Number**, and **URL** of the source.

**Urdu\_Sentence\_text:** The sentence to be analyzed.

**English\_Sentence\_text:** Its translation (used only for reference).

**Sentiment(P/N):** The ground truth sentiment label — "P" for positive and "N" for negative.

The dataset was balanced reasonably well, and preprocessing steps were applied before training.

# Results

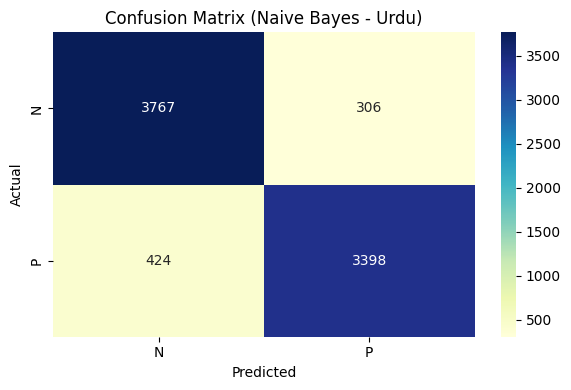
The Naive Bayes classifier achieved the following performance metrics on the test set:

**Accuracy:** 90.75%

**Precision:** High precision for both classes (**90%** for Negative Class and **92%** for Positive Class).

**Recall:** High precision for both classes (**92%** for Negative Class and **89%** for Positive Class).

**F1-score:** High precision for both classes (**91%** for Negative Class and **90%** for Positive Class).



# Error Analysis

|  |  |  |
| --- | --- | --- |
| Original Sentence | Actual Sentiment | Predicted Sentiment |
| Correct Assessment | | |
| میں نے ہمیشہ خوشی سے ایک ساتھ رہنے کی کوشش کی ہے | P | P |
| بس اپنی بہنوں کو بھول جاؤ | N | N |
| کیا ہوا آپ کیوں پریشان نظر آتے ہیں | N | N |
| وہ ٹھیک ہیں اسے ہوش میں آ گیا ہے | P | P |
| ہاں میڈم میں سمجھتا ہوں | P | P |
| Incorrect Assessment | | |
| فکر نہ کرو وہ ابھی بہت ناراض ہے | N | P |
| آپ اپنی ساس کے فرمانبردار ہیں | P | N |
| لیکن کسی کے بچوں کی خوشی اور قسمت نہیں | N | P |
| میں انتظار نہیں کرسکتا حمزہ | P | N |
| پیسہ ایسی بری چیز ہے | N | P |