- Null values in the dataset were replaced with an empty string.
- The 'Category' column (containing 'spam' or 'ham') was label encoded, with 'spam' as 0 and 'ham' as 1.
- The dataset was split into training and test sets, with 80% of the data used for training and 20% for testing.
- The TfidfVectorizer from scikit-learn was used to convert the text messages into numerical feature vectors.
- Words with a frequency of more than 1 were considered, and common English stop words were removed.
- A Logistic Regression model was used for the binary classification task (spam or not spam).
- The model was trained on the feature vectors from the training set.
- The trained model achieved an accuracy of 96.70% on the training data.
- The model achieved an accuracy of 96.59% on the test data.
- The similar accuracy scores on both training and test data suggest that the model is not overfitting or underfitting.
- Two example SMS messages were provided: one expressing gratitude (predicted as 'Ham SMS') and another offering a gift (predicted as 'Spam SMS').
- The prediction process involves converting the input SMS message into a feature vector using the same TfidfVectorizer, and then passing it through the trained Logistic Regression model.