MIDS Assignment 2

1. Loading the required libraries:

```
import pandas as pd
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize

data = pd.read_csv("Tweets.csv")
```

2. Data Pre-processing:

```
def preprocess_text(text):
  text = text.lower()
  tokens = word_tokenize(text)
  stop_words = set(stopwords.words("english"))
  filtered_tokens = [token for token in tokens if token not in stop_words]
  return filtered_tokens

data["text"] = data["text"].apply(preprocess_text)
```

3. Feature Engineering:

```
from sklearn.feature_extraction.text import CountVectorizer
vectorizer = CountVectorizer(max_features=2000)
features = vectorizer.fit_transform(data["text"])
```

4. Train-Test Split:

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(features, data["airline_sentiment"], test_size=0.2)
```

5. KNN Model:

```
from sklearn.neighbors import KNeighborsClassifier

knn = KNeighborsClassifier(n_neighbors=5)
knn.fit(X_train, y_train)
```

6. **Model Evaluation:**

- **Prediction on Test Set:**

```
y_pred = knn.predict(X_test)
```

- **Evaluation Metrics:**

```
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score

print("Accuracy:", accuracy_score(y_test, y_pred))
print("Precision:", precision_score(y_test, y_pred, average="weighted"))
print("Recall:", recall_score(y_test, y_pred, average="weighted"))
print("F1-Score:", f1_score(y_test, y_pred, average="weighted"))
```