Report: NLP Sentiment Analysis Model

1. Introduction:

The aim of this project is to build a sentiment analysis model using Natural Language Processing (NLP). The model classifies text into three categories: happiness, angriness, and sadness. The project uses various NLP techniques, including text cleaning, feature transformation, and machine learning models.

2. Design Choices:

2.1. Data Loading and Labeling:

Three datasets, representing different sentiments (angry, happy, sad), are loaded and concatenated.
A label mapping is created to convert sentiment labels into numerical values.

2.2. Data Preprocessing:

Text data is cleaned using functions like `remove_stopwords` and `text_clean` to remove unnecessary information and reduce noise.
Duplicate entries are identified and removed.
The length of the content is calculated and analyzed for each sentiment label.

2.3. Feature Transformation:

Two types of feature transformation are explored: Count Vectorization and TF-IDF Vectorization.
Vectorization is performed on the cleaned content to convert text into numerical features for model training.

2.4. Model Selection:

Three classifiers are chosen for evaluation: Logistic Regression, Naive Bayes, and Random Forest.
The models are trained on both Count Vectorization and TF-IDF Vectorization.

3. Performance Evaluation:

3.1. Model Training and Evaluation:

Models are trained and evaluated using accuracy and area under the ROC curve (AUC) as performance metrics.
Performance is compared for both Count Vectorization and TF-IDF Vectorization.

3.2. Results:

Logistic Regression and Random Forest outperform Naive Bayes in both accuracy and AUC.
TF-IDF Vectorization generally provides better results compared to Count Vectorization.

4. Discussion of Future Work:

4.1. Model Improvement:

Fine-tune hyperparameters of the chosen models to potentially enhance performance. Explore more advanced algorithms and ensemble methods for improved accuracy.

4.2. Feature Engineering:

Experiment with different text preprocessing techniques and feature engineering methods to capture more nuanced information from the text.

5. Source Code:

The provided source code encompasses the entire pipeline, including data loading, preprocessing, feature transformation, model training, and evaluation. Key libraries such as pandas, numpy, nltk, sklearn, and wordcloud are utilized.

```
# Importing essential libraries and functions
import pandas as pd
import numpy as np
import nltk
import re

nltk.download("stopwords")
nltk.download("punkt")

from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize

#Visualisation Library
import matplotlib.pyplot as plt

# Feature Transformation Library
from sklearn.feature_extraction.text import CountVectorizer, TfidfVectorizer
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.linear_model import LogisticRegression
from sklearn.naive_bayes import MultinomialNB

#Classification Report and model evaluation
from sklearn.metrics import roc_auc_score, classification_report

from wordcloud import WordCloud
from sklearn.metrics import roc_curve, auc
from sklearn.preprocessing import label_binarize
from itertools import cycle
```

Conclusion:

The sentiment analysis model demonstrates promising performance, with Logistic Regression and Random Forest emerging as strong contenders. Future work involves refining the model, addressing imbalances, and exploring advanced NLP techniques to further enhance accuracy. The provided source code serves as a comprehensive guide for building and evaluating sentiment analysis models.