TF-IDF Case Study Report

# 1. SMS Spam Detection

Libraries Installed:

- datasets

- scikit-learn

Dataset Loading:

```python

from datasets import load\_dataset

dataset = load\_dataset("sms\_spam")

texts = dataset['train']['sms']

labels = dataset['train']['label']

```

Process:

- Dataset: SMS Spam

- Vectorization: TfidfVectorizer with stop\_words='english', max\_df=0.95, min\_df=2

- Classifier: Logistic Regression

- Evaluation: Accuracy, Classification Report, Confusion Matrix, Top TF-IDF Terms

Results:

- Accuracy: ~0.9785

- Precision: 1.0

- Recall: 0.84

- F1-Score: 0.91

# 2. IMDb Subset Sentiment Analysis

Dataset Loading:

```python

from datasets import load\_dataset

dataset = load\_dataset("imdb")

texts\_pos = dataset['train'].filter(lambda x: x['label'] == 1).select(range(150))['text']

texts\_neg = dataset['train'].filter(lambda x: x['label'] == 0).select(range(150))['text']

texts = texts\_pos + texts\_neg

labels = [1]\*150 + [0]\*150

```

Process:

- Dataset: IMDb (150 positive, 150 negative)

- Vectorization: TfidfVectorizer with stop\_words='english', max\_df=0.95, min\_df=2

- Classifier: Logistic Regression

- Evaluation: Accuracy, Classification Report, Confusion Matrix

Results:

- Accuracy: ~0.9333

# 3. Yelp Polarity Subset Sentiment Analysis

Dataset Loading and Cleaning:

```python

from datasets import load\_dataset

import re

dataset = load\_dataset("yelp\_polarity", split="train[:500]")

texts = dataset['text']

labels = dataset['label']

def clean\_text(text):

text = re.sub(r'[^\w\s]', '', text.lower())

text = re.sub(r'\d+', '', text)

return text

texts\_cleaned = [clean\_text(t) for t in texts]

```

Process:

- Dataset: Yelp Polarity (500 samples)

- Preprocessing: Lowercasing, punctuation and digit removal

- Vectorization: TfidfVectorizer with stop\_words='english', max\_df=0.9, min\_df=2, ngram\_range=(1,2)

- Classifier: Logistic Regression

- Evaluation: Accuracy, Classification Report, Cosine Similarity, Confusion Matrix

Results:

- Accuracy: ~0.72

# Model Comparison Visualizations

- Pie Chart: Accuracy comparison among models

- Bar Graph: Accuracy comparison among models

Summary Table:

| Model | Accuracy |

|------------------------------------|----------|

| Logistic Regression (SMS Spam) | 0.9785 |

| Logistic Regression (IMDb Subset) | 0.9333 |

| Logistic Regression (Yelp Subset) | 0.72 |