# TF-IDF Case Study

## By

Hari Om, Vishwajeet Kumar, Shivam Singh

## Introduction

TF-IDF, which stands for Term Frequency-Inverse Document Frequency, is a statistical measure used in NLP to evaluate word importance in documents. It is widely used in information retrieval and text mining.

## Why TF-IDF?

- Highlights important terms specific to a document.
- Reduces the weight of common and less informative words.
- Enhances classification and similarity-based tasks.

## Mathematical Background

Term Frequency (TF):

TF(t, d) = (No. of times term t appears in d) / (Total terms in d)

Inverse Document Frequency (IDF):

IDF(t) = log(N / (1 + df(t)))

TF-IDF:

TF-IDF(t, d) = TF(t, d) * IDF(t)

## SMS Spam Dataset (Highest Accuracy)

# TF-IDF Case Study

Spam messages contain specific vocabulary such as 'free', 'win', etc. TF-IDF scores these highly in spam. Logistic Regression works well due to clear linear separation.

Accuracy: 97.85%

## IMDb Subset Dataset (Moderate Accuracy)

Movie reviews are nuanced and varied. TF-IDF helps but doesn't capture context. Logistic Regression still finds moderately effective boundaries.

Accuracy: 93.33%

## Yelp Polarity Subset Dataset (Worst Accuracy)

Reviews are noisy and context-heavy. Vocabulary overlap between positive and negative reviews makes classification harder. TF-IDF and Logistic Regression struggle.

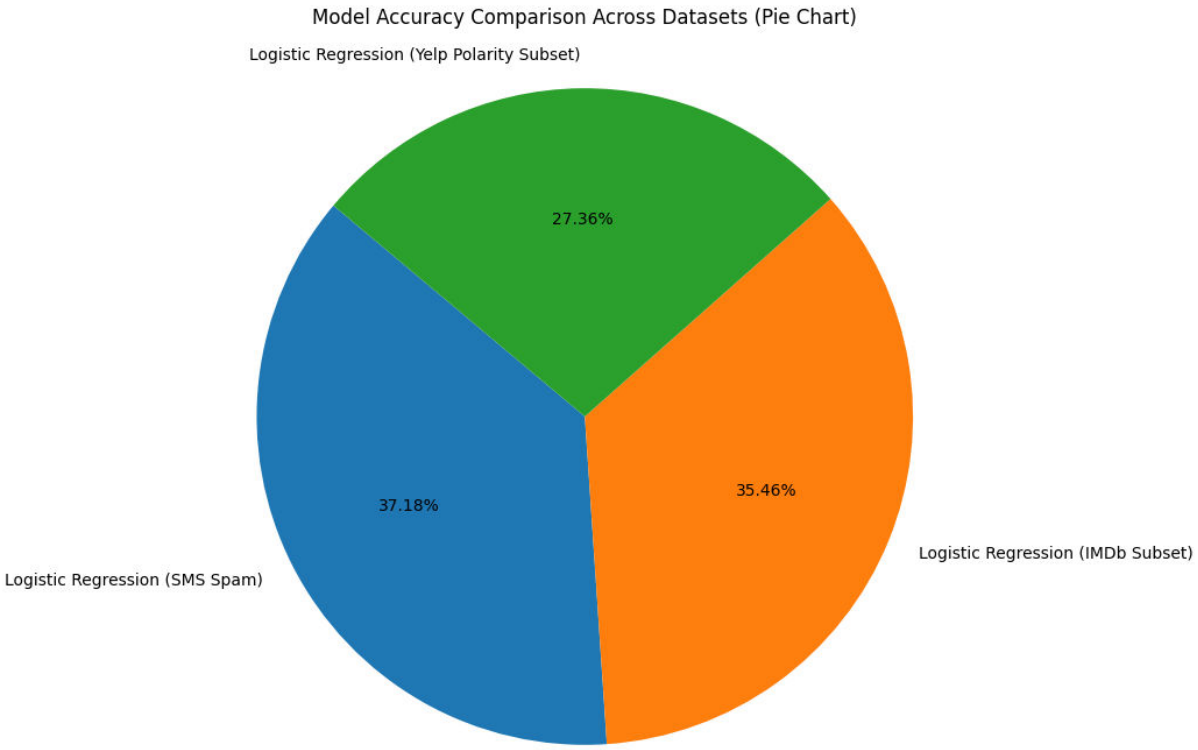
Accuracy: 72.00%

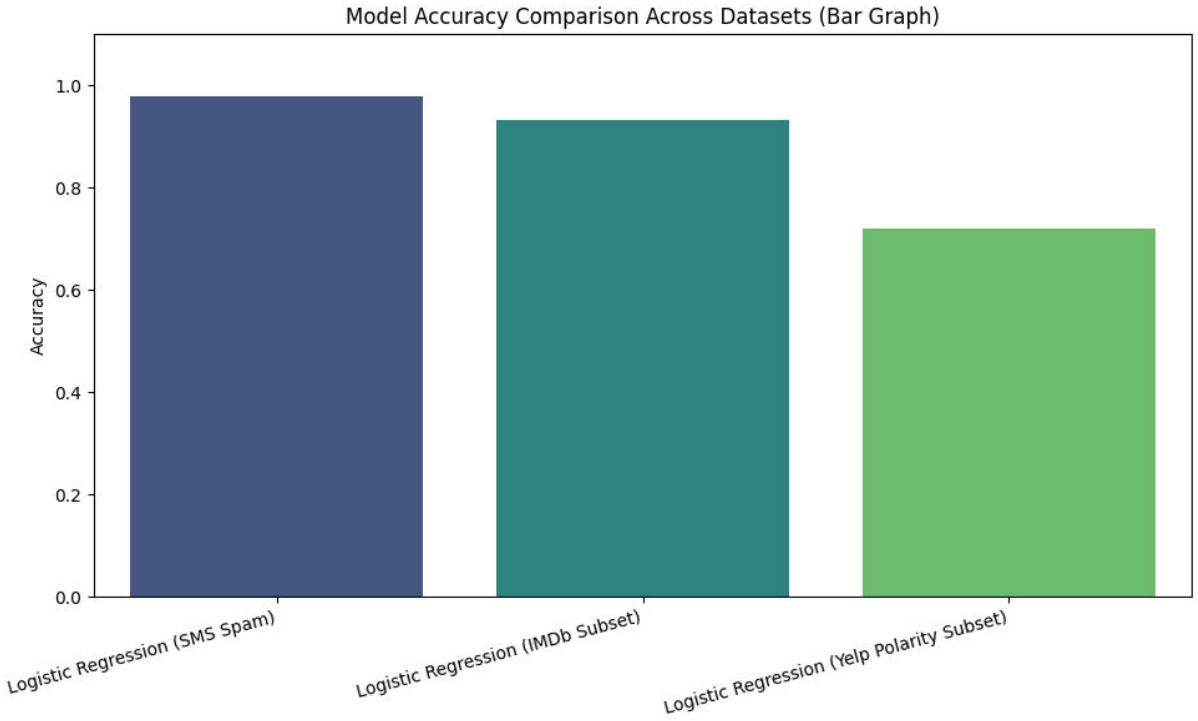
## Conclusion

TF-IDF is effective where vocabulary is clearly distinguishable across classes. For nuanced or noisy datasets, more complex models are preferred.

## Screenshot

# TF-IDF Case Study



# TF-IDF Case Study