Topic : 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 Natural Language Processing (NLP) to evaluate the importance of a word in a document relative to a collection of documents (corpus). It is commonly 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 in NLP

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 Formula:

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

Observation on Study-Cases

Here , in this case study of the above topic on various datasets available in various libraries accessible from python. In the below data on the operation of the TF_IDF model on the datasets are different as :-

❖ SMS Spam Dataset (Highest Accuracy):

Nature of Data: Spam messages often use very specific, repetitive, and often unusual vocabulary (e.g., "free," " win," "cash," "prize," "claim," "urgent," misspellings, excessive punctuation, all caps). Ham messages tend to be more conversational and less focused on these distinct keywords.
How TF-IDF Helps: TF-IDF excels at identifying words that are important in a specific document but not very common across the entire dataset. Spam-specific words tend to fit this description – they appear frequently in spam but rarely in ham. This gives them high TF-IDF scores in spam messages, making them strong indicators for the "spam" class.
Why Logistic Regression Works: Since the distinguishing features (spam-specific words) have high TF-IDF scores in one class and low scores in the other, a linear model like Logistic Regression can easily find a clear boundary (a hyperplane in the feature space) to separate spam from ham. The features are highly discriminative and linearly separable to a large extent.
Conclusion: The vocabulary differences between spam and ham are often very stark and consistent, making it a relatively easy classification problem for TF-IDF and Logistic Regression.

❖ IMDb Subset Dataset (Moderate Accuracy):

Nature of Data: Movie reviews use a much wider range of vocabulary. While there are positive and negative words (e.g., "amazing," "brilliant," vs. "terrible," "boring"), the language is more nuanced. Positive reviews might contain some negative words in a comparative context, and vice versa. Sarcasm, complex sentence structures, and figurative language are more common.

How TF-IDF Helps: TF-IDF can still identify important words for positive and negative reviews. Words like " great," "loved," "excellent" might get high scores in positive reviews, and "awful," "waste," "disappointed" in negative ones. However, the signal isn't as clean as in spam detection. Words can have different meanings based on context, which TF-IDF alone doesn't capture.

Why Logistic Regression Works (Moderately): Logistic Regression can still learn from the weights of individual words identified by TF-IDF. Words strongly associated with positive sentiment will have positive coefficients, and negative words will have negative coefficients. It can find a reasonable boundary, but the separation isn't as clear-cut as with spam because sentiment is expressed more subtly and contextually.

Conclusion: Sentiment analysis is more complex than spam detection due to the complexity and variability of human language expressing opinions. TF-IDF captures word importance but misses context, and Logistic Regression is limited by its linearity. This leads to a moderate accuracy score.

❖ Yelp Polarity Subset Dataset (Worst Accuracy):

Nature of Data: Yelp reviews are often shorter, more informal, can contain misspellings, slang, and are highly subjective.They cover a vast range of topics (restaurants, services, etc.), leading to very diverse vocabulary that isn't necessarily sentiment-specific. A review might mention factual details (e.g., "the service was slow," "the portion size was small") that are objectively negative but don't use strongly negative sentiment words. Conversely, a review might be factually positive ("great location") but lack strong positive adjectives.

How TF-IDF Helps (Less Effectively): While TF-IDF will still highlight words frequent in specific reviews and rare overall, the most important terms might be about the specific business or service ("burger," "waitress," "cleanliness") rather than sentiment-laden words. The overlap in vocabulary between positive and negative reviews can be higher compared to spam/ham or even positive/negative movie reviews. N-grams (like ngram_range=(1,2)) were used, which helps capture simple phrases, but the noise and variability in Yelp reviews are high.

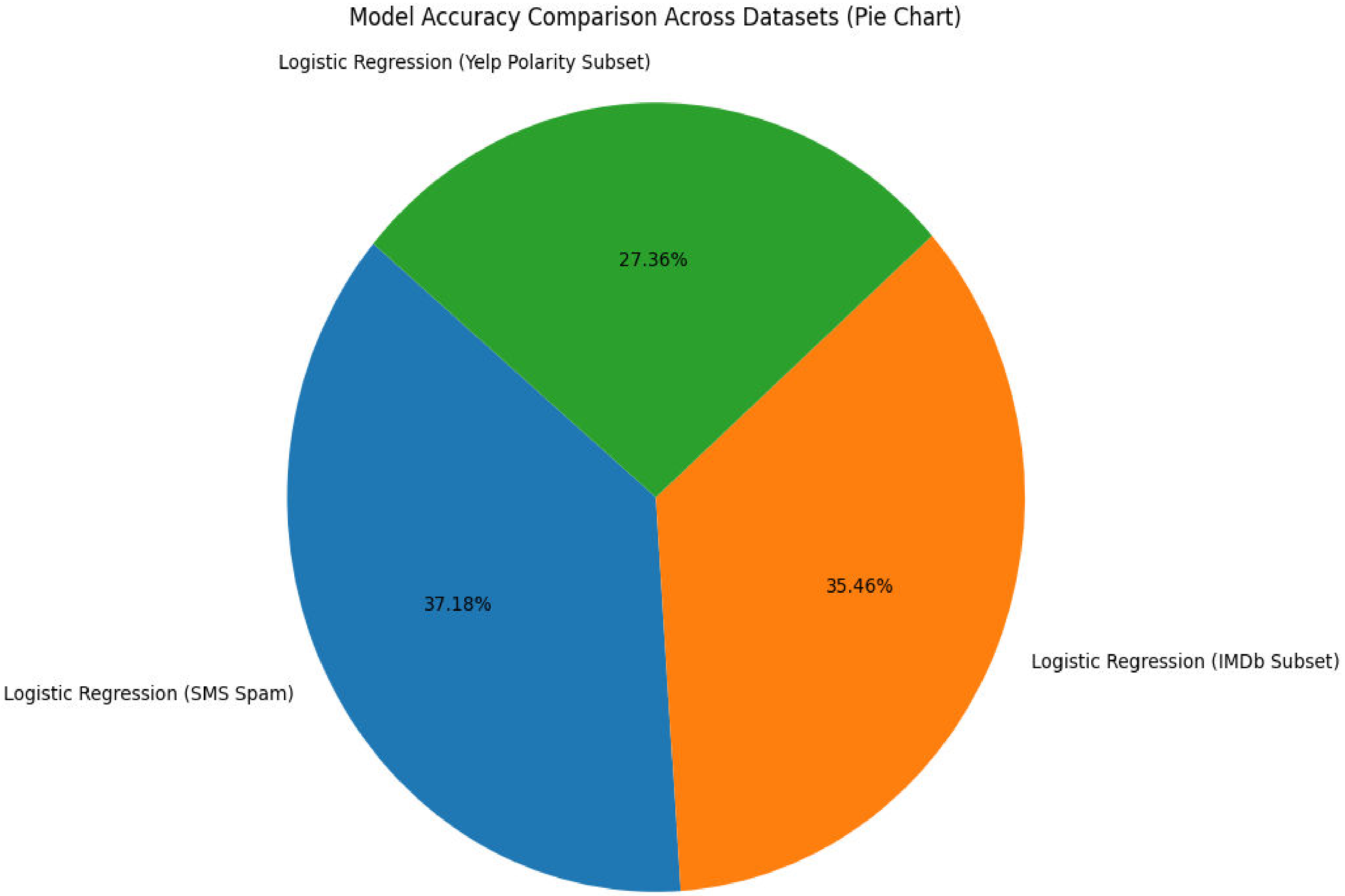
Why Logistic Regression Struggles: With a less clear distinction in important TF-IDF terms between positive and negative reviews, the linear boundary that Logistic Regression tries to learn becomes less effective. The noise and diverse topics make it harder for the model to rely solely on word frequencies and bigram frequencies to determine overall sentiment.

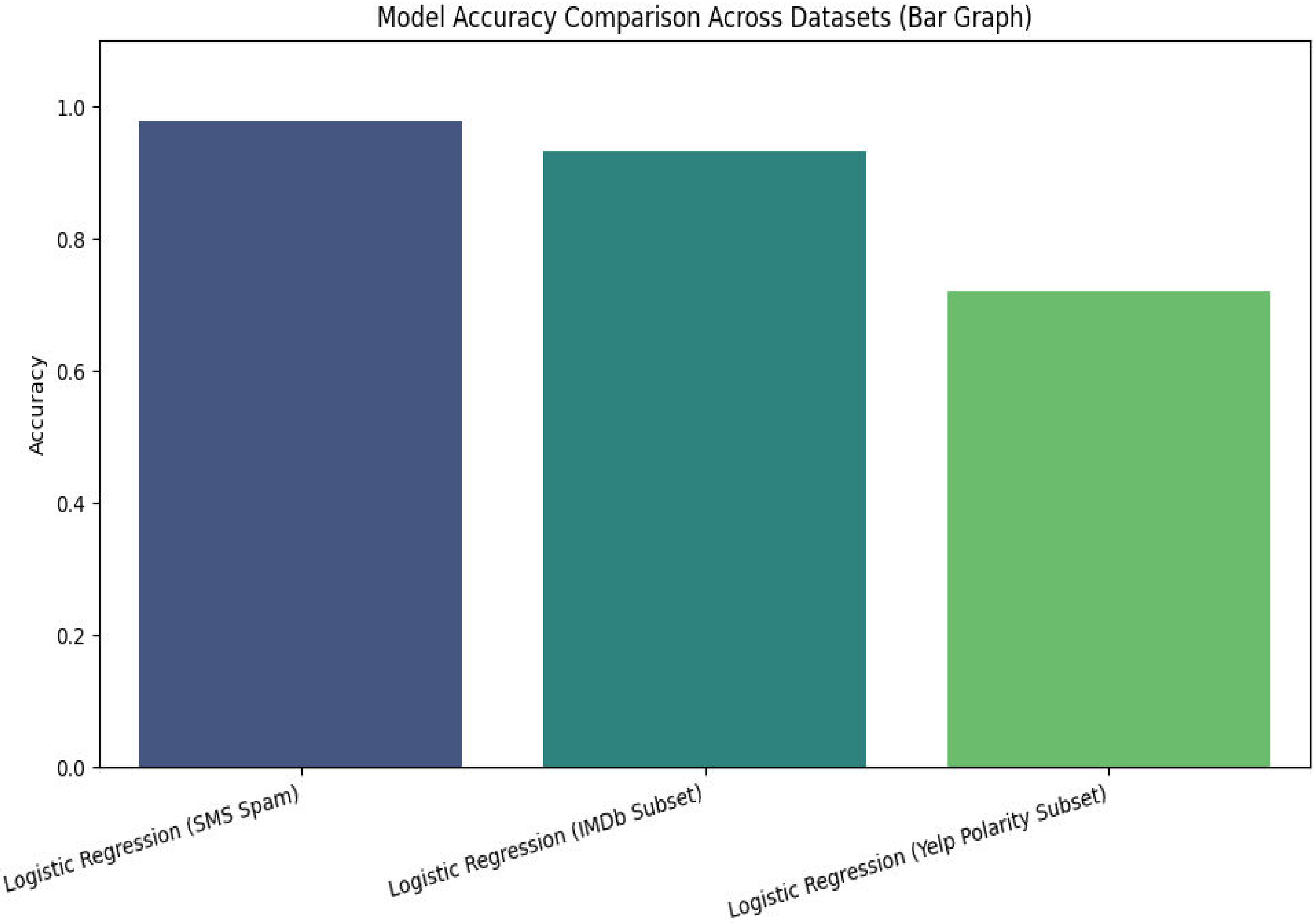
Conclusion: Yelp reviews are generally considered noisier and harder to classify sentimentally than movie reviews or spam/ham. The diverse topics and informal language dilute the effectiveness of TF-IDF, and a simple linear model like Logistic Regression struggles to capture the nuances required for accurate sentiment prediction on this type of data. More sophisticated techniques (like using embeddings, transformers, or more complex models) are often needed for better performance on such datasets.

Metrics used for the pie chart:

```
                         model  accuracy

1         Logistic Regression (SMS Spam)                  0.9785
2         Logistic Regression (IMDb Subset)               0.9333
3         Logistic Regression (Yelp Polarity Subset)      0.7200
```


Model Accuracy Comparison Across Datasets (Pie Chart)
Logistic Regression (Yelp Polarity Subset)
27.36%
37.18%
35.46%
Logistic Regression (SMS Spam)
Logistic Regression (IMDb Subset)


Model Accuracy Comparison Across Datasets (Bar Graph)
Accuracy
1.0
0.8
0.6
0.4
0.2
0.0
Logistic Regression (SMS Spam)
Logistic Regression (IMDb Subset)
Logistic Regression (Yelp Polarity Subset)