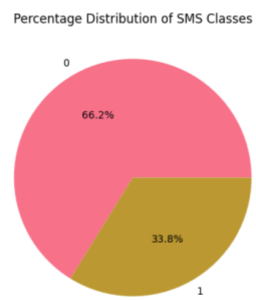
Part A- SMS Spam Classification

Introduction

I have developed a SMS spam classification system to accurately identify spam messages across different languages. I have first analyzed the data, then used TF-IDF and Logistic Regression to predict spam messages.

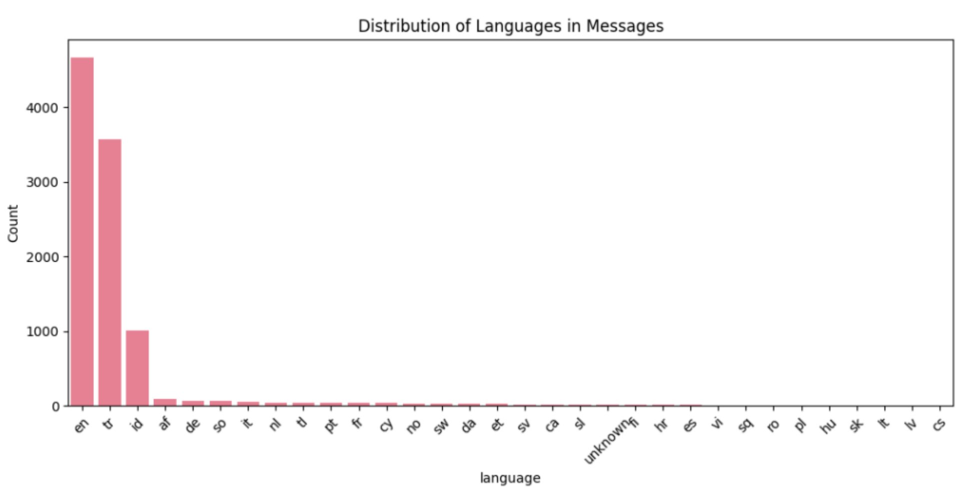
Data Analysis

Dataset size: 10,000 SMS messages
Class distribution: 66.% non-spam vs 33% spam messages



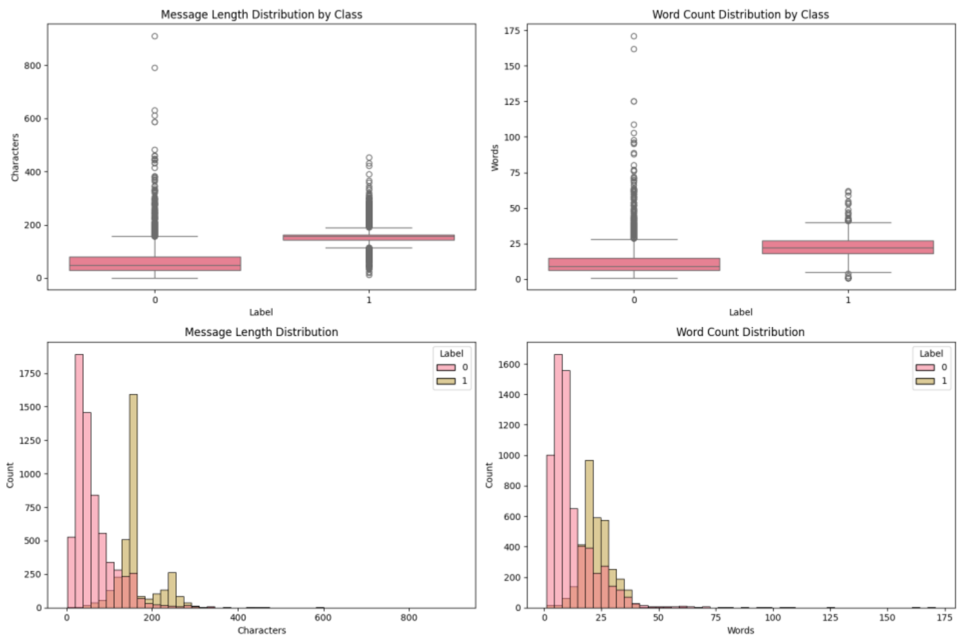
Further analysis shows;

Main languages are English, Turkish and Indonesian.



Spam messages are longer with average 164 characters vs 64 for non-spam

Spam messages contain more numbers and URLs

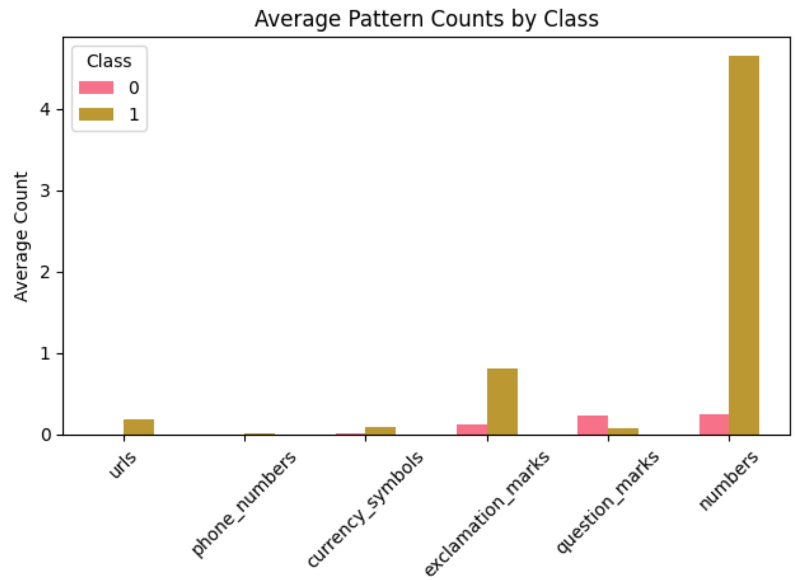


Language distribution shows different spam rates across different languages

Spam Percentage by Language:

Label	0	1
language		
af	94.25	5.75
ca	95.45	4.55
cs	100.00	0.00
cy	100.00	0.00
da	100.00	0.00
de	44.29	55.71
en	80.14	19.86
es	90.00	10.00
et	82.61	17.39
fi	100.00	0.00

The comparison of spam vs non-spam based on messages containing different types of text can be seen here:



Modeling

To model the data, I first removed stop-words and duplicates, split the data, created features. Then I created a pipeline using TfidfVectorizer and LogisticRegression(with L2 penalty).

I chose Logistic Regression because:

1. Works well with sparse text features
2. Provides interpretable feature importance
3. Fast training and prediction times

Results

Achieved 98% accuracy on test set with F1-score: 0.97 for spam detection.

Test Set Classification Report:

	precision	recall	f1-score	support
0	0.98	0.99	0.98	1276
1	0.97	0.96	0.97	658
accuracy			0.98	1934
macro avg	0.98	0.97	0.98	1934
weighted avg	0.98	0.98	0.98	1934

Feature Importance

After analyzing results, we see that the most important features are:

1. Presence of phone numbers
2. URLs
3. Keywords like free, ozel, gratis,
4. Message length