SMS SPAM FILTER

MACHINE LEARNING & TEXT ANALYTICS

SUBMITTED BY:

D22010 – ASWIN KUMAR I S

D22024 – LAXMI PANCHAL

D22027 – MAHESHKUMAR N

GUIDED BY:

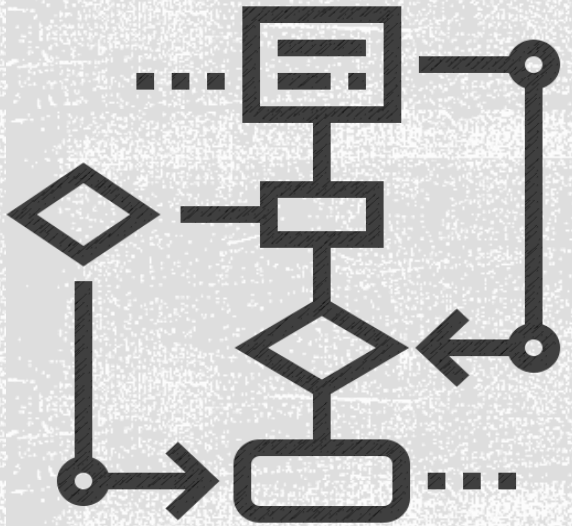
PROF. GOURAB NATH

Praxis Business School

CELEBRATE YOUR WORTH

DATA PRE-PROCESSING FLOW DIAGRAM

1. Raw Text Messages
2. Removal of Extra White Spaces
3. Removal of Stop Words
4. Lemmatization
5. Cleaned Text Corpus



DATA EXPLORATION USING WORD CLOUD

HAM MESSAGE

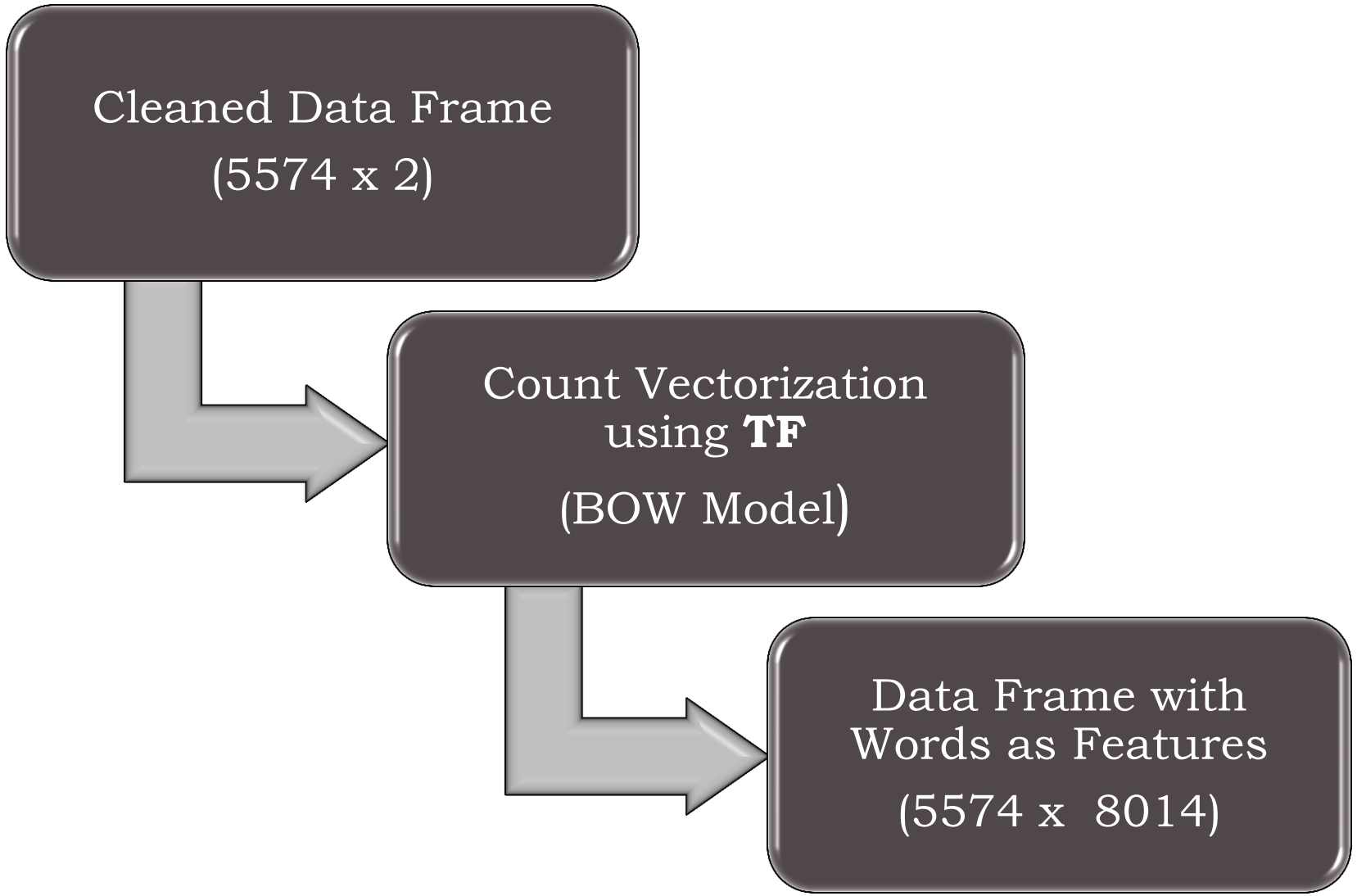
OBSERVATIONS:

The words like go, call, know, come, u, m, lt, ok, tell, time, want, need, make, send, love, good, take, one, think, say, today, ok, now are frequent words appeared in ham messages

SPAM MESSAGE

OBSERVATIONS:

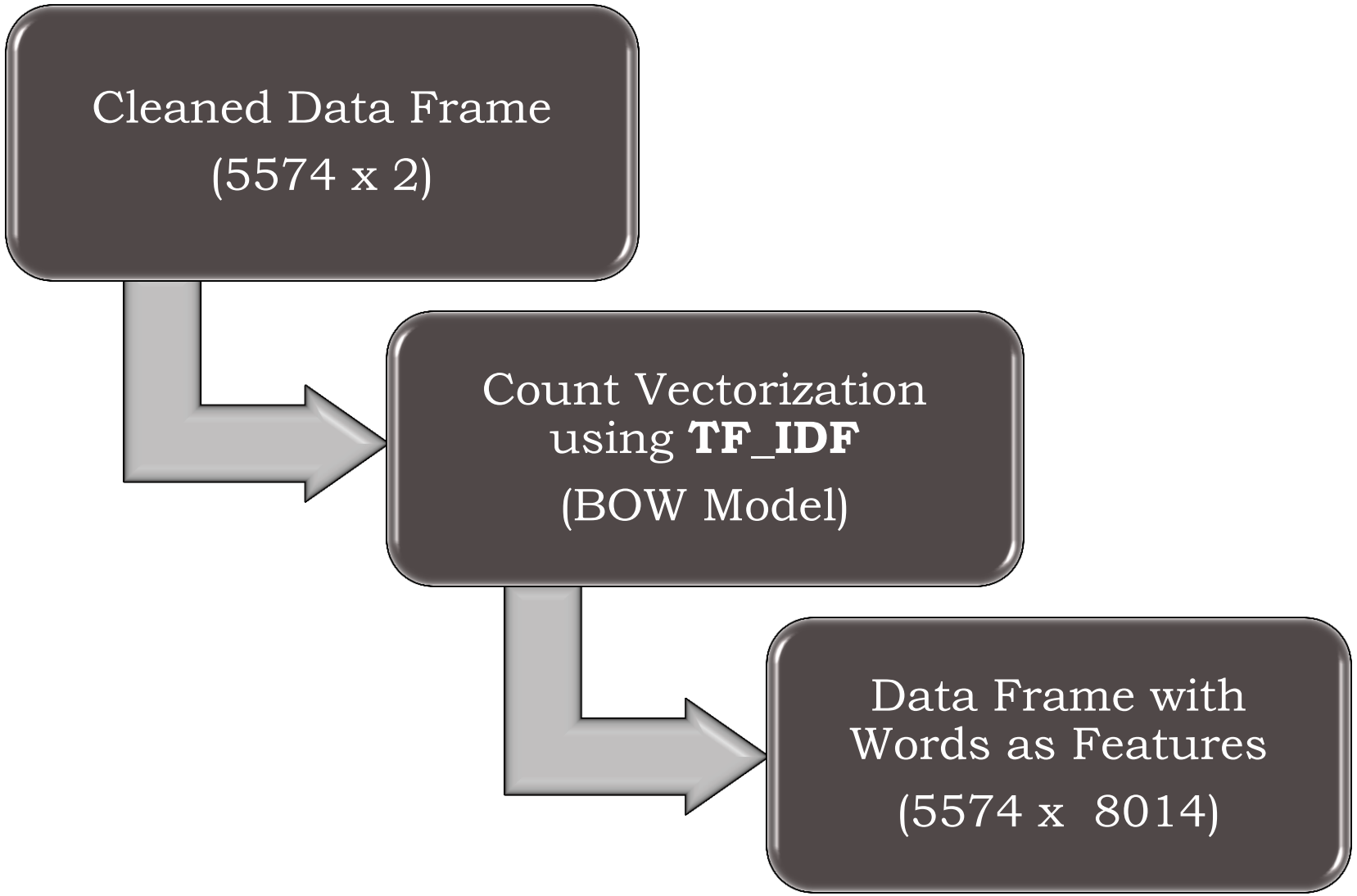
The words like free, call, txt, stop, reply, contact, award, please, claim, prize, service, customer, now, mobile, win, chance, voucher are frequent words appeared in spam messages



TERM FREQUENCY

VECTORIZATION PIPELINE

By computing the term-frequency of each word in a given document, we can get a good understanding of what the document is about and its relevance to other documents



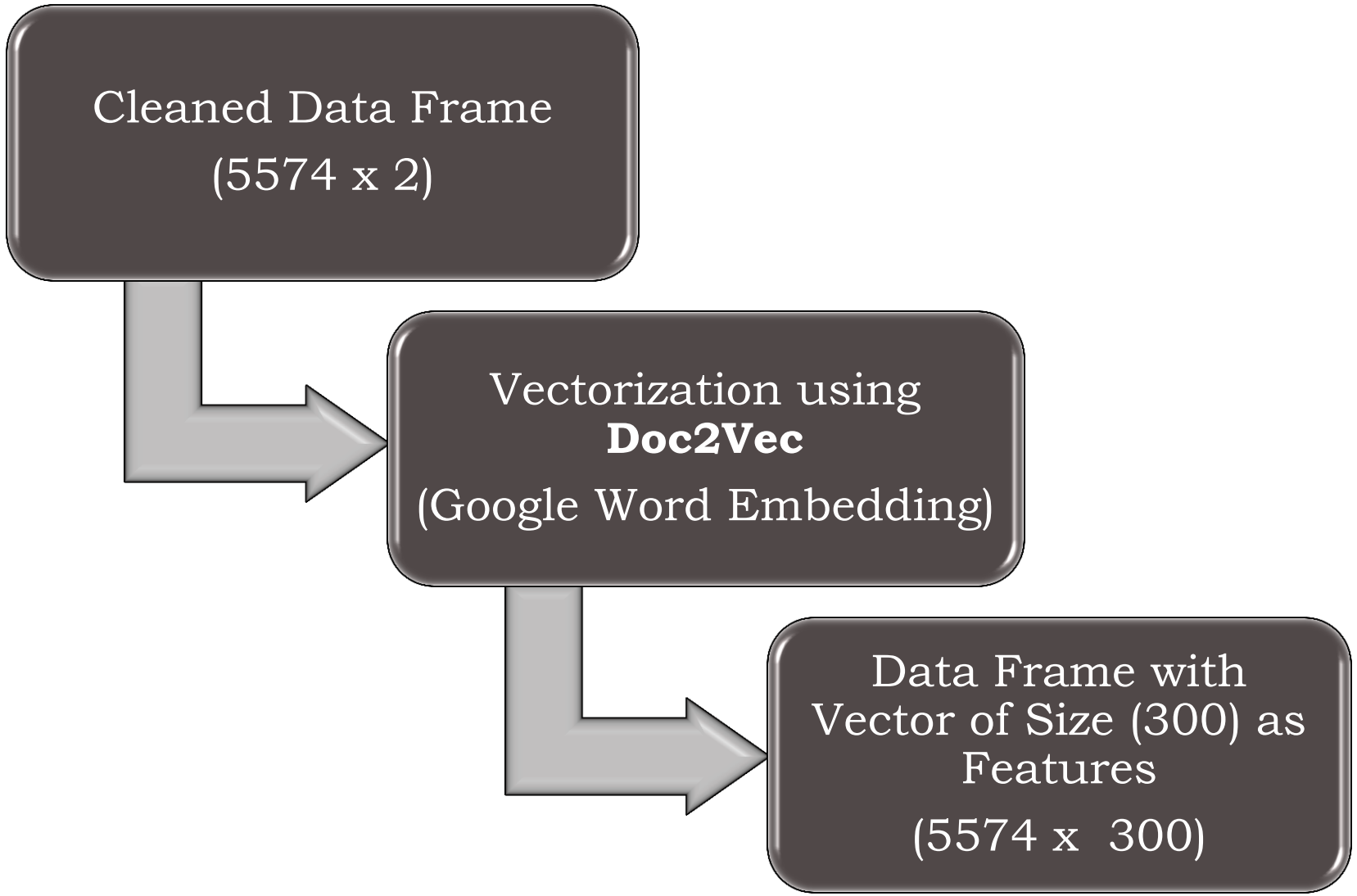
INVERSE DOCUMENT TERM FREQUENCY

VECTORIZATION PIPELINE

The idea behind IDF is that if a word appears in many documents, it's less likely to be a good indicator of the content of any one document.

On the other hand, if a word appears in only a few documents, it's more likely to be a good indicator of the content of those documents.

By combining term-frequency with inverse document frequency, we can get a better understanding of what words are important for each document, and we can use this information to classify the documents.

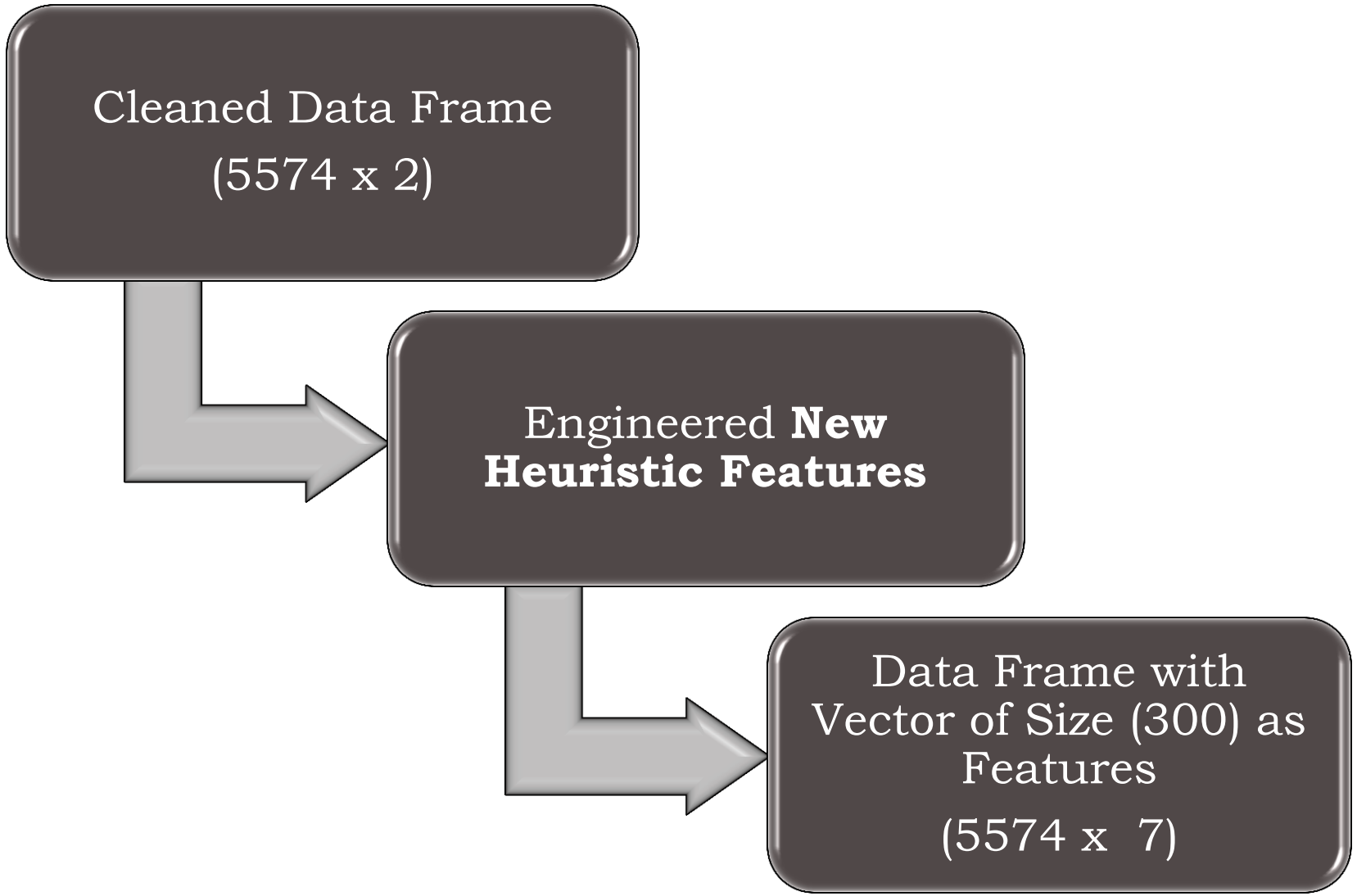


DOC2VEC

VECTORIZATION PIPELINE

This technique is based on the idea that words that are similar in meaning should have similar vectors, so words that are close together in the vector space are more likely to be related in meaning.

By using word2vec vectorization, we can convert our SMS messages into vectors that can be used as inputs to our classification models.



HEURISTIC

HEURISTIC FEATURES CREATED:

1) Presence of **Phone Number**(Binary Feature)

2) Presence of **$ sign** (Binary Feature)

3) Presence of **Capital letter Word** (Binary Feature)

4) Proportion of **Spelling Mistakes** (Continuous numeric)

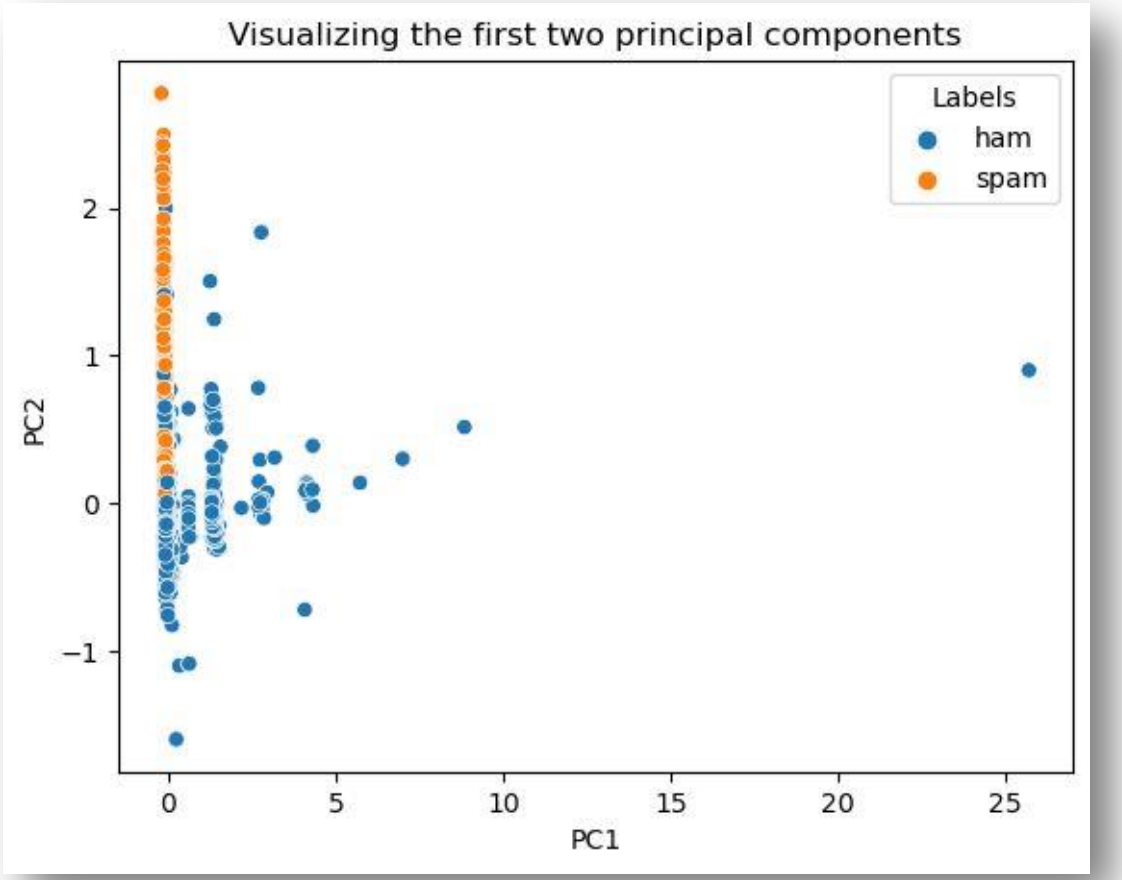
5) Proportion of **Punctuations** (Continuous numeric)

6) **Subjectivity Score** ()

7) **Sentiment Score** ()

VISUALIZATION

First Two Principal Components

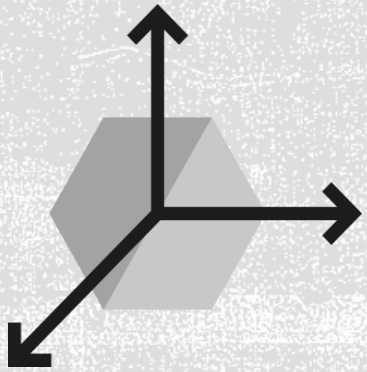


- Used PCA (Principal Component Analysis) on TF Vectors.
- First Two Principal Components for Plotting the 2D Graph

DIMENSION REDUCTION

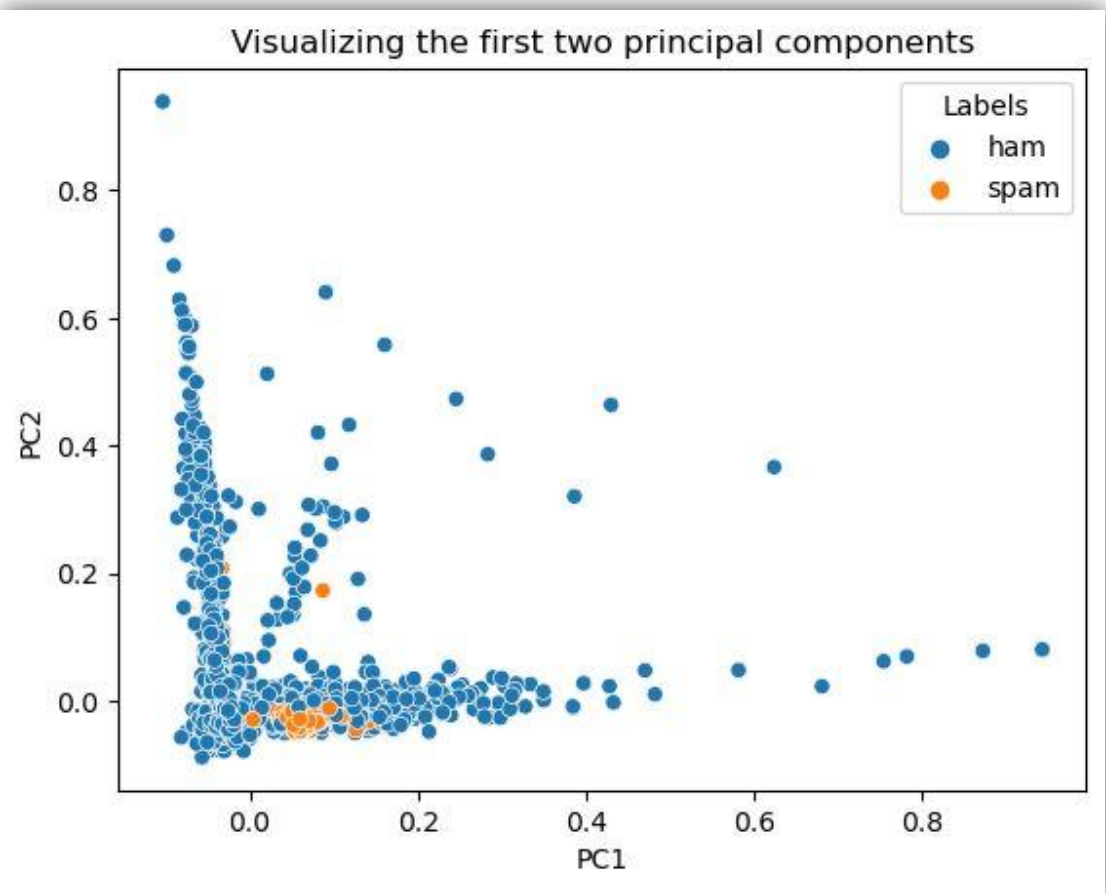
TF VECTORS

Using PCA SELECTED 1796 PC's WHICH EXPLAINED 95% OF VARIANCE



VISUALIZATION

First Two Principal Components

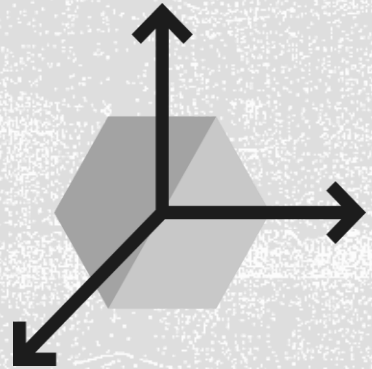


- Used PCA (Principal Component Analysis) on TF-IDF Vectors.
- First Two Principal Components for Plotting the 2D Graph

DIMENSION REDUCTION

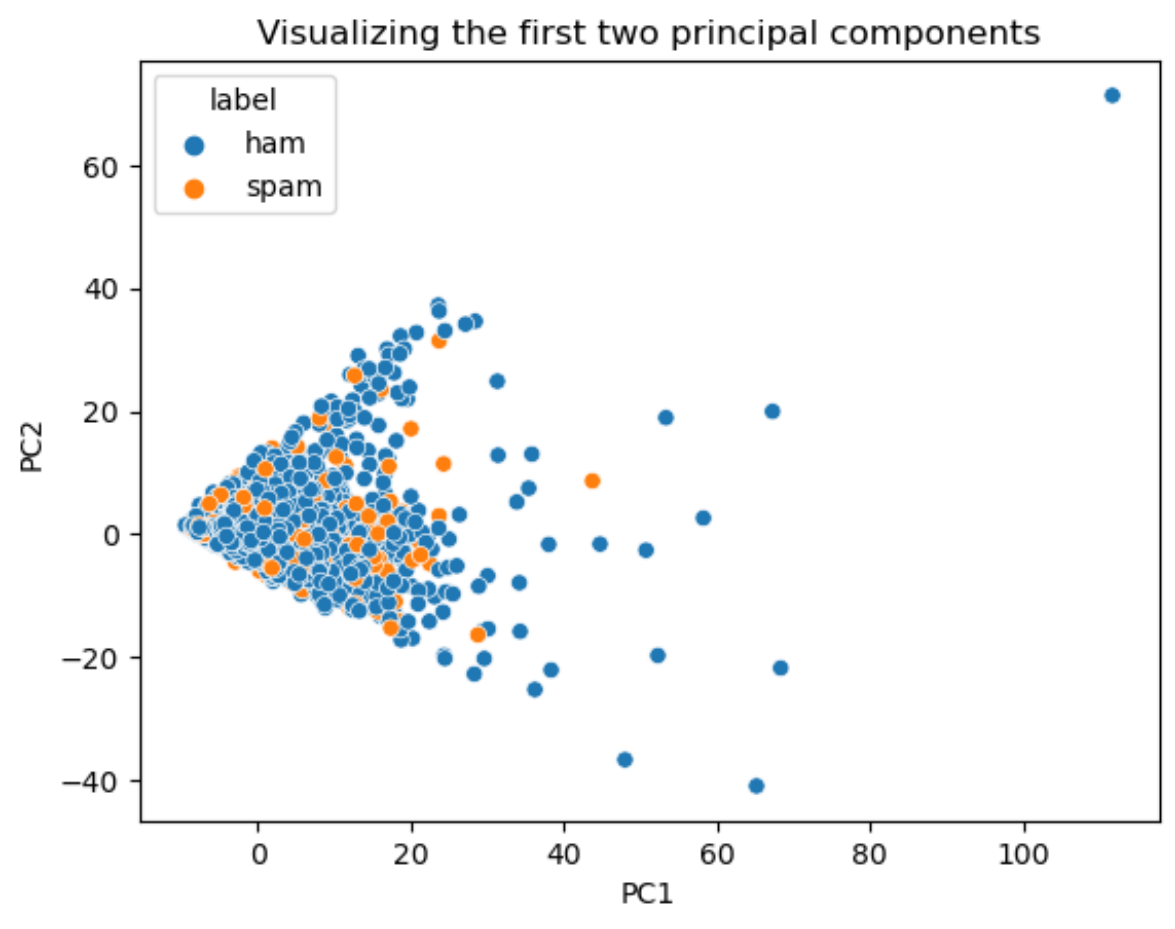
TF-IDF VECTORS

Using PCA SELECTED 2591 PC's WHICH EXPLAINED 95% OF VARIANCE



VISUALIZATION

First Two Principal Components

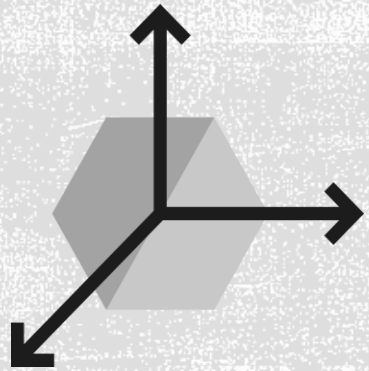


- Used PCA (Principal Component Analysis) on Doc2Vec.
- First Two Principal Components for Plotting the 2D Graph

DIMENSION REDUCTION

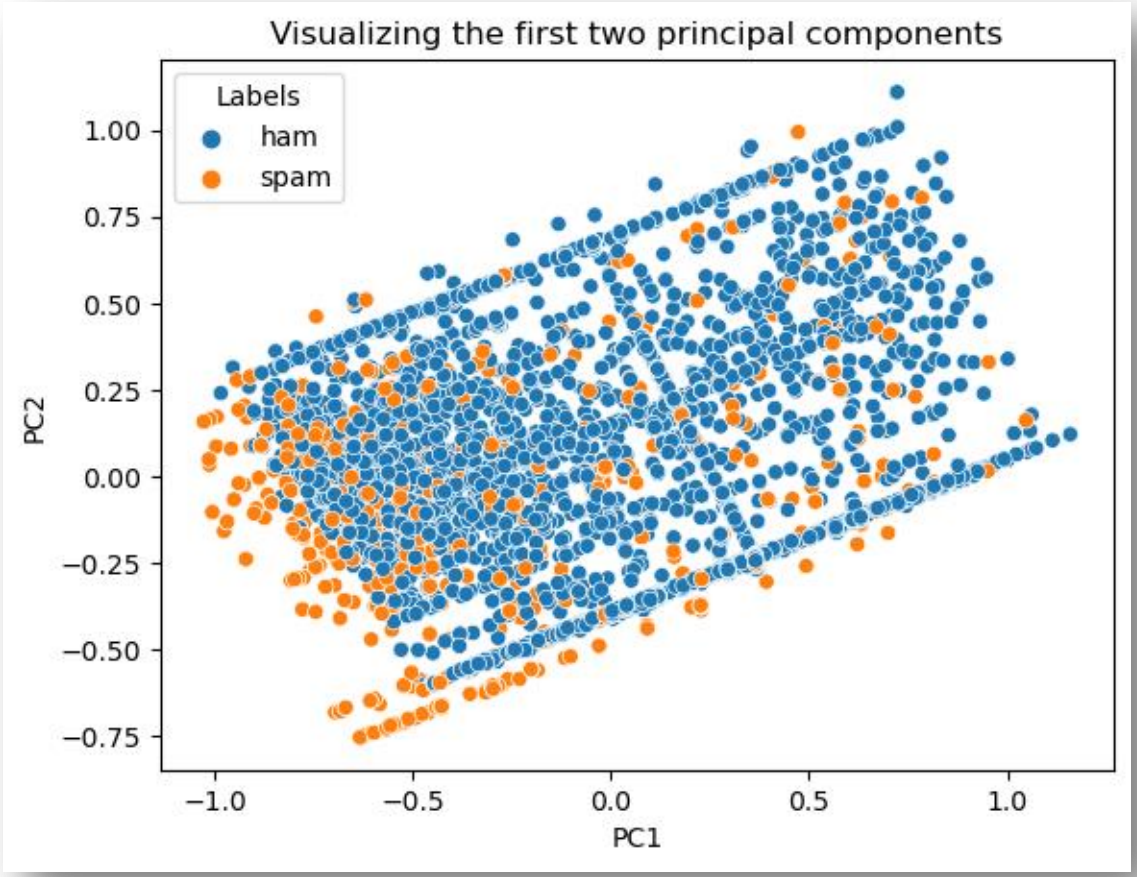
DOC2VEC

Using PCA SELECTED 180 PC's WHICH EXPLAINED 95% OF VARIANCE



VISUALIZATION

First Two Principal Components



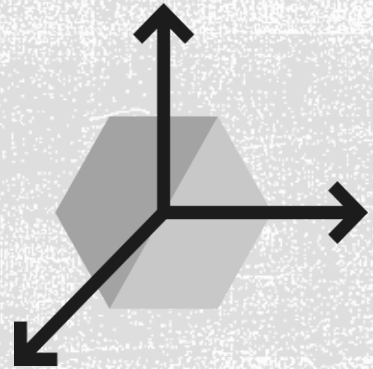
- Used PCA (Principal Component Analysis) on Heuristic.
- First Two Principal Components for Plotting the 2D Graph

DIMENSION REDUCTION

HEURISTIC

Using PCA Using PCA SELECTED

3 PC's WHICH EXPLAINED 95% OF VARIANCE



MODEL 1:

LOGISTIC REGRESSION ON TF VECTOR MATRIX

ERROR METRICS	RESULTS
ACCURACY	0.98
F1-SCORE	0.92
RECALL	0.86
PRECISION	0.98

AIM:

To Classify Text Messages as Ham or Spam

EXPERIMENT:

Fitted Logistic Regression Model on Term Frequency Vector Matrix.

MODEL 2:

LOGISTIC REGRESSION ON TF VECTOR MATRIX WITH DIMENSION REDUCTION USING PCA

ERROR METRICS	RESULTS
ACCURACY	0.97
F1-SCORE	0.91
RECALL	0.85
PRECISION	0.90

AIM:

To Classify Text Messages as Ham or Spam

EXPERIMENT:

Fitted Logistic Regression Model on Term Frequency Vector Matrix After Dimensionality Reduction Using PCA.

MODEL 3:

LOGISTIC REGRESSION ON TF VECTOR MATRIX WITH DIMENSION REDUCTION USING 'p' %

ERROR METRICS	RESULTS
ACCURACY	0.97
F1-SCORE	0.90
RECALL	0.84
PRECISION	0.98

AIM:

To Classify Text Messages as Ham or Spam

EXPERIMENT:

Fitted Logistic Regression Model on Term Frequency Vector Matrix After Dimensionality Reduction Using 'p' percentage.

1 % is Fixed as Threshold , Words appeared less than 1% in the Corpus is Removed.

SELECTED FEATURES - 200

MODEL 4:

RANDOM FOREST ON TF-IDF VECTOR MATRIX

ERROR METRICS	RESULTS
ACCURACY	0.97
F1-SCORE	0.90
RECALL	0.82
PRECISION	1.00

AIM:

To Classify Text Messages as Ham or Spam

EXPERIMENT:

Fitted Random Forest Model on Inverse Document Term Frequency Vector Matrix.

MODEL 5:

ADABOOST ON TF-IDF VECTOR MATRIX WITH DIMENSIONALITY REDUCTION USING PCA

ERROR METRICS	RESULTS
ACCURACY	0.97
F1-SCORE	0.89
RECALL	0.88
PRECISION	0.90

AIM:

To Classify Text Messages as Ham or Spam

EXPERIMENT:

Fitted Ada-boost Model on Inverse Document Term Frequency Vector Matrix After Dimensionality Reduction Using PCA.

MODEL 6:

RANDOM FOREST ON TF-IDF VECTOR MATRIX WITH DIMENSIONALITY REDUCTION USING 'p' %

ERROR METRICS	RESULTS
ACCURACY	0.93
F1-SCORE	0.72
RECALL	0.65
PRECISION	0.81

AIM:

To Classify Text Messages as Ham or Spam

EXPERIMENT:

Fitted Random Forest Model on Inverse Document Term Frequency Vector Matrix After Dimensionality Reduction Using 'p' percentage.

1 % is Fixed as Threshold , Words appeared less than 1% in the Corpus is Removed.

SELECTED FEATURES – 15

MODEL 7:

DECISION TREE ON DOC2VEC METHOD

ERROR METRICS	RESULTS
ACCURACY	0.76
F1-SCORE	0.18
RECALL	0.18
PRECISION	0.17

AIM:

To Classify Text Messages as Ham or Spam

EXPERIMENT:

Fitted Decision Tree Model with Depth= 5 on Doc2Vec Method.

MODEL 8:

LOGISTIC REGRESSION ON DOC2VEC WITH DIMENSIONALITY REDUCTION USING PCA

ERROR METRICS	RESULTS
ACCURACY	0.99
F1-SCORE	0.97
RECALL	0.96
PRECISION	0.99

AIM:

To Classify Text Messages as Ham or Spam

EXPERIMENT:

Fitted Logistic Regression on Doc2Vec After Dimensionality Reduction Using PCA.

MODEL 9:

RANDOM FOREST ON HEURISTIC FEATURES

ERROR METRICS	RESULTS
ACCURACY	0.94
F1-SCORE	0.78
RECALL	0.67
PRECISION	0.93

AIM:

To Classify Text Messages as Ham or Spam

EXPERIMENT:

Fitted Random Forest on Heuristic Features.

MODEL 10:

BAGGED DECISION TREE ON HEURISTIC FEATURES

ERROR METRICS	RESULTS
ACCURACY	0.93
F1-SCORE	0.75
RECALL	0.65
PRECISION	0.89

AIM:

To Classify Text Messages as Ham or Spam

EXPERIMENT:

Fitted Bagged Decision Tree on Heuristic Features.

CONCLUSION:

- We observed the performance of the different machine learning models with 4 different vectorization techniques along with dimensionality reduction techniques like PCA and ‘p’ percent.
- The performance of machine learning models with Term Frequency Vectorization Method performs well when compared to other 3 Vectorization methods.
- The Doc2Vec Vectorization Method gives worst performance with all machine learning models.
- Machine learning model like Logistic Regression and Random Forest Performs well in all Vectorization approach.
- The Heuristic approach performs better than Doc2vec approach and it performs well with only 7 features.
- We observe that there is a great opportunity to develop more heuristic features which classifies the target accurately and are relatively easier to interpret than other vectorization methods

THANK YOU