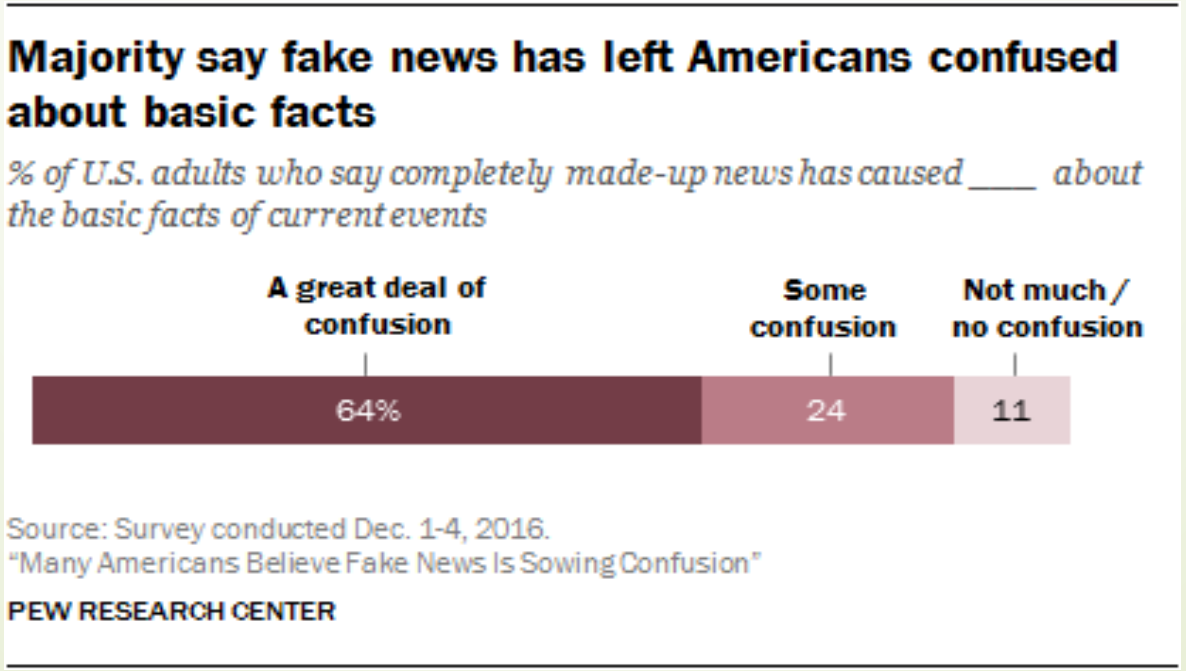
Fake News Identification and Classification

Prepared By: Kundan Singh [S4723435]

Why & its Implications

- Prevalence of fake news on social media
- Emerging research area in Machine Learning & NLP



Data Set & Exploration Tools

- **Data Source**: Kaggle https://www.kaggle.com/c/fake-news/data

- **train.csv**: A full training dataset with the following attributes:

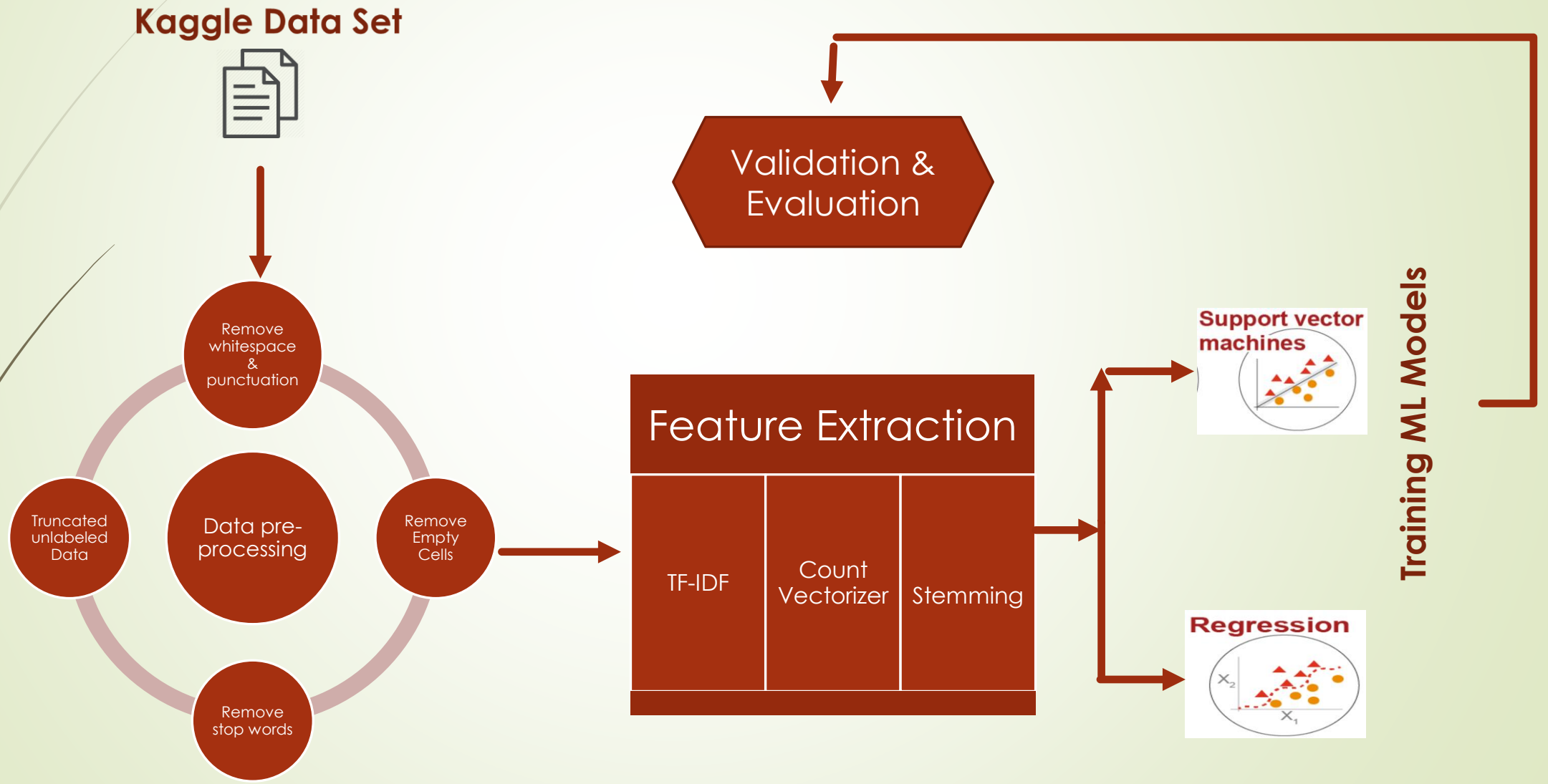
  id, title, author, text, label [1:Unreliable, 0: Reliable]

- **test.csv**: A testing training dataset with all the same attributes at train.csv without the label.

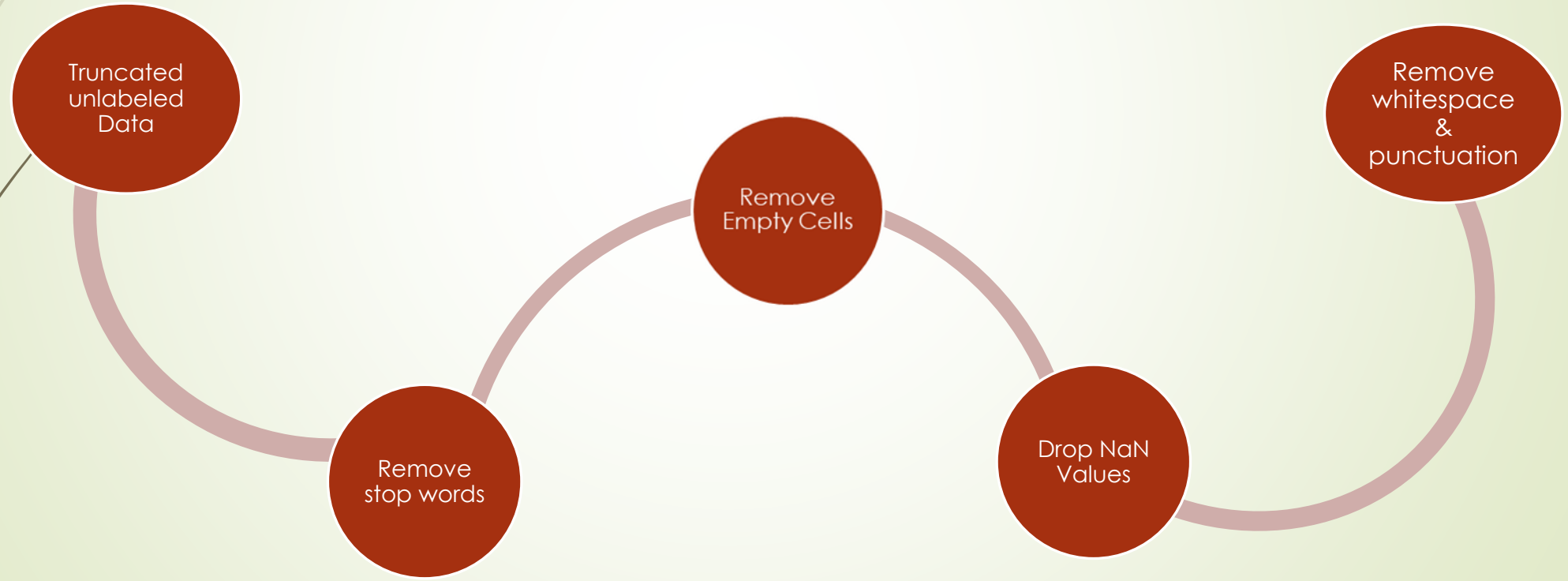
- **Tools & Dependencies Libraries:**

  python 3.6+, numpy, pandas, Sklearn, matplotlib,nltk

Work Flow

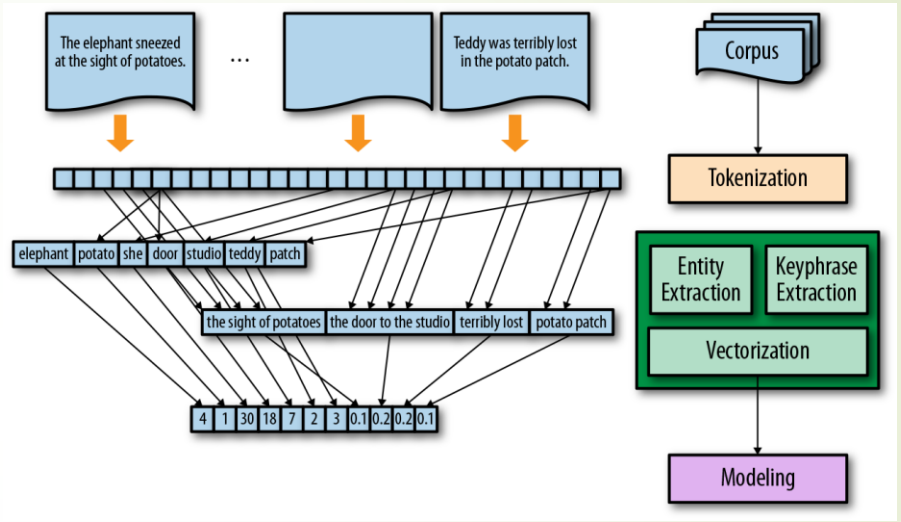

Kaggle Data Set
Validation & Evaluation
Remove whitespace & punctuation
Truncated unlabeled Data
Data pre-processing
Remove Empty Cells
Remove stop words
Feature Extraction
TF-IDF
Count Vectorizer
Stemming
Support vector machines
Regression
Training ML Models

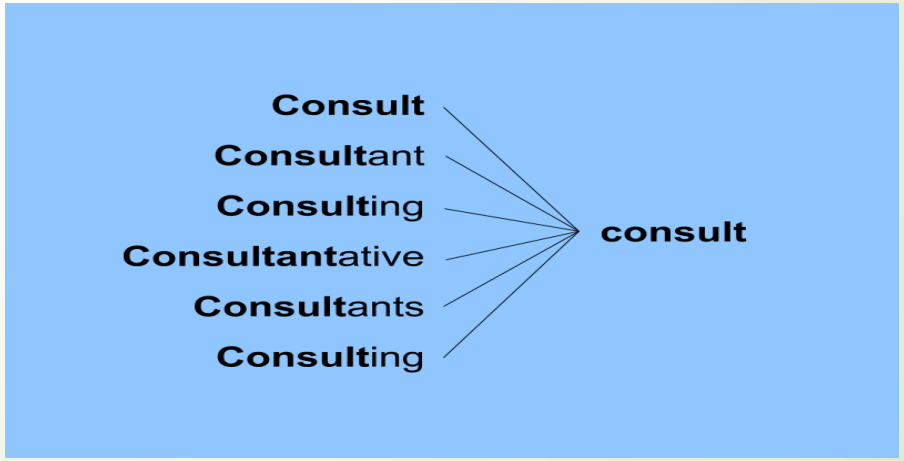
Data Pre-processing



Feature Extraction

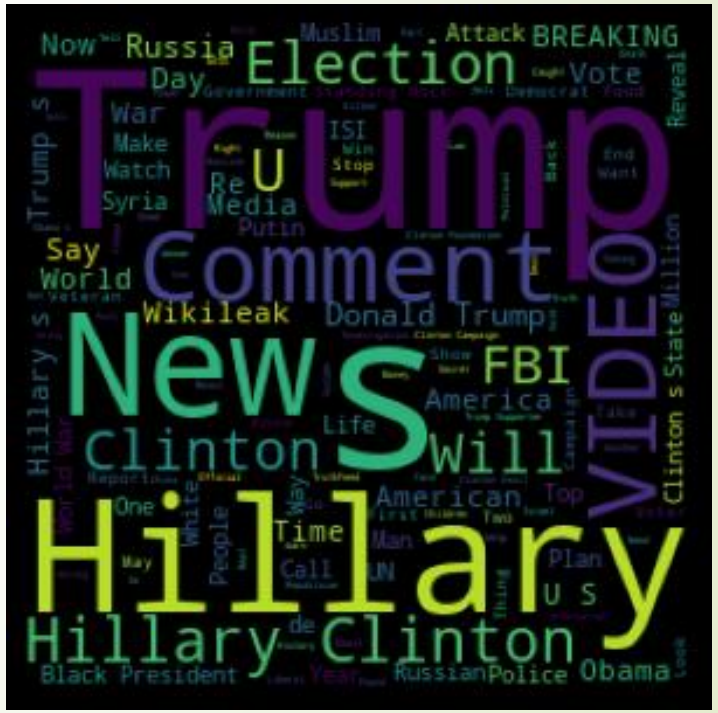
- TF-IDF
- Count Vectorizer
- Stemming
  - Porter Stemming
  - nltk





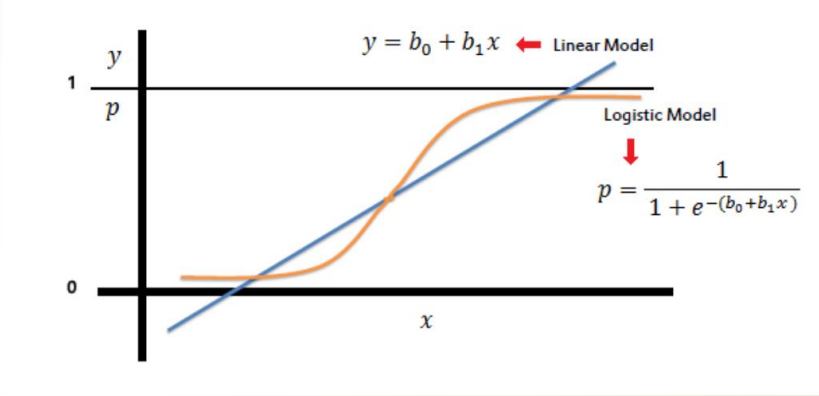
Feature Extraction

Real **Word Cloud** **Fake**

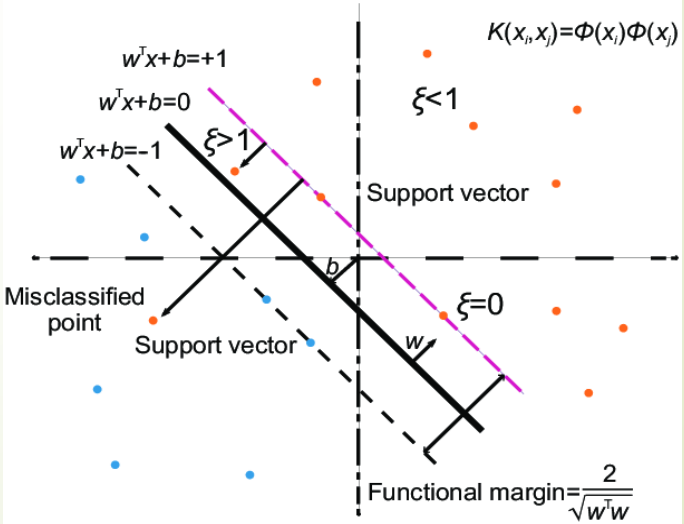


Training a Model

- Logistic Regression
  - Gradient Descent
  - Sigmoid function

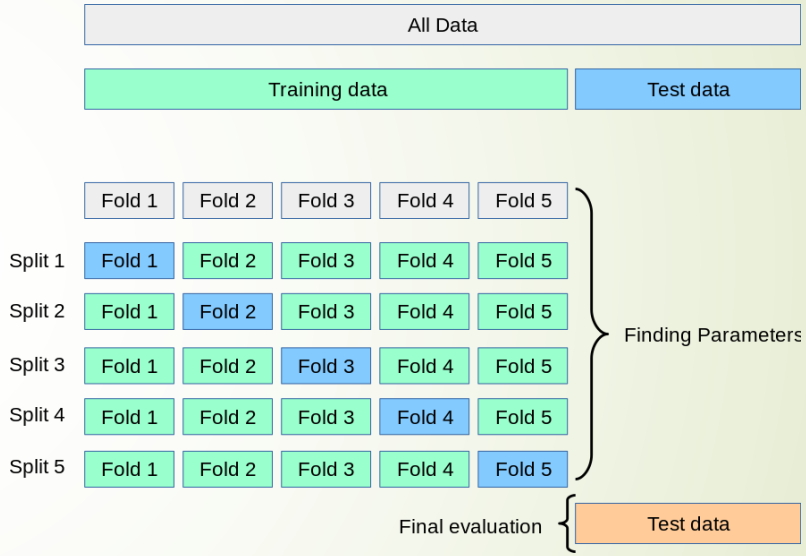


- SVM
  - linear SV Classifier
  - robust



Cross Validation

- 20 % of Data Set
- CV = 10
- GridSearchCV

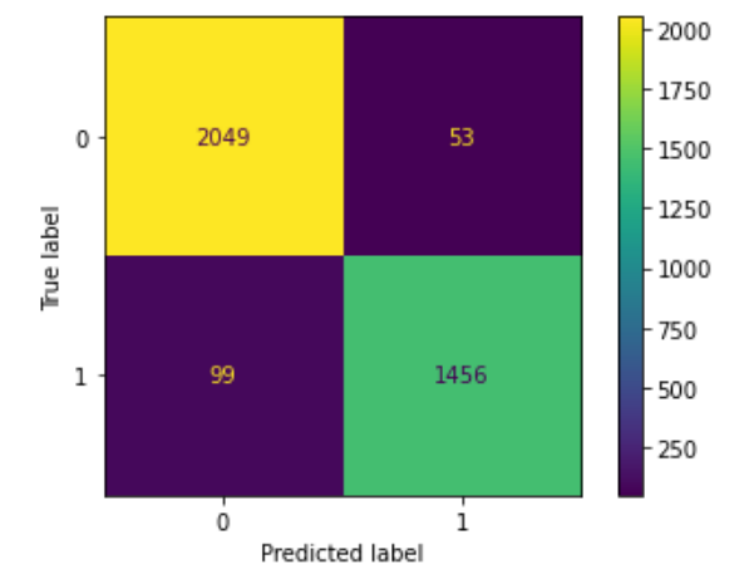


Best training accuracy :

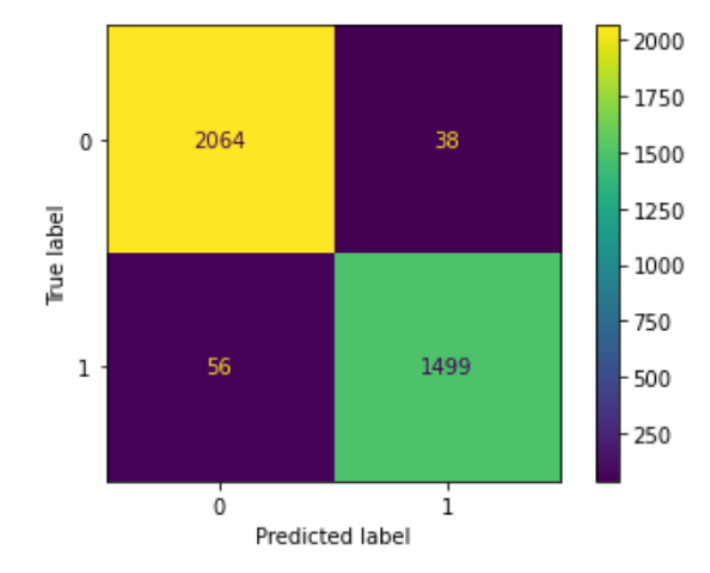
Logistic Regression: 95.4%
SVM: 97.3%

Confusion Matrices

Logistic Regression



SVM



Evaluation Metrics Results & Summary

Logistic Regression

CLassification F1/Precision/Recall:

	precision	recall	f1-score	support
0	0.95	0.97	0.96	2102
1	0.96	0.94	0.95	1555
accuracy			0.96	3657
macro avg	0.96	0.96	0.96	3657
weighted avg	0.96	0.96	0.96	3657

Accuracy: 95.84 %

SVM

CLassification F1/Precision/Recall:

	precision	recall	f1-score	support
0	0.97	0.98	0.98	2102
1	0.98	0.96	0.97	1555
accuracy			0.97	3657
macro avg	0.97	0.97	0.97	3657
weighted avg	0.97	0.97	0.97	3657

97.43%

Classifier with Best test accuracy : SVM

Future Improvement

- Use of Word embedding techniques word2vec
- Can use CNN, LSTM for better accuracy
- Use of Pre-trained vectors

THANK YOU