

# MacOS & Windows APIs & NLP Text Classification

Data Science Immersive Course  
Project 3

# Problem Statement

Given that

## We work for a Tech News & Media Company

that writes/talks about Computer Technology update, which usually choose variety of IT buzzing and popular topics from social media post to write/talks in our news

## Scenario

**Due to high competitiveness in this industry** where everyone can become a news reporter and publisher, Business positioning is a must!, we can't just write random article about this and that just like before.

### In response,

Our company is deciding to be more specific and subjective, whether our news is to serve Apple or Windows fanboy for our brand positioning sake!

### Mac or Windows, one way or another,

All we need is the tool to skim through posts whether each one is about Windows or Mac so we can choose post from our (yet) chosen label for upcoming articles/news correctly and conveniently !

# Our Plan

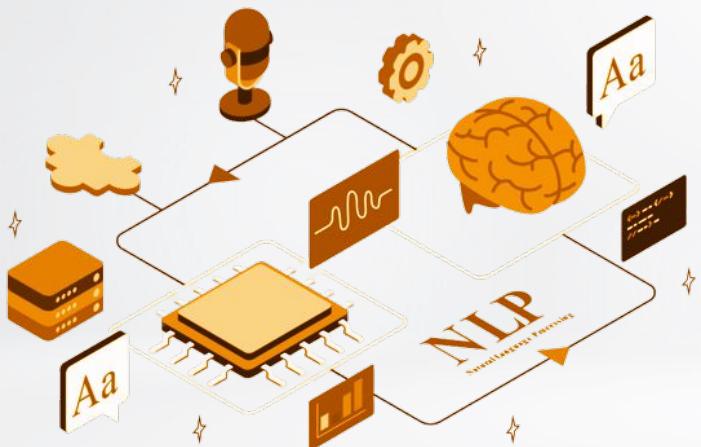
Find out ➤

1. **Most Precise Classification Model**  
to classify text from social media alike.
  
2. **Recommendation** for further studies and  
Business strategy/execution suggestion

## 1. Most Precise Classification Model

- to classify resemble text from social media.

# Text Classification Modeling



### NLP Technique for Feature engineering:

This involves transforming the raw text data into a numerical representation that can be used by the machine learning model. This will be done using a technique called bag-of-words (CountVectorizer).

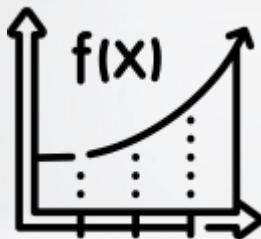
### Model selection and training:

Once the text data has been transformed into a numerical representation, a machine learning model can be selected and trained.

Our variety machine learning models that will be used for text classification are including **Logistic Regression, Decision Tree, Random Forests, KNN-neighbor and Multinomial Naive Bayes.**



Reddit



Modeling & Prediction

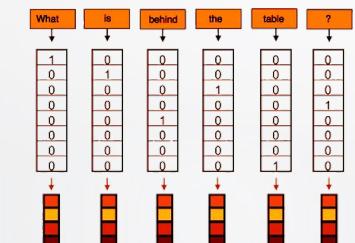
2  
Datasets

Scraped Text from Reddit  
Period: Year 2010-2022

MacOS: 19,312 rows, 4 columns

Windows: 28,209 rows, 4 columns

Vary from 2010 - 2022



Feature Engineering  
using NLP Technique:  
**Bag of words**

\*Excludes stopwords and class name (Mac,Win)

EDA

Exploratory  
Data Analysis



Data Cleansing

Data Imputation

(i.e. filling/dropping missing values, duplicates filtering out odd ones)

New Column

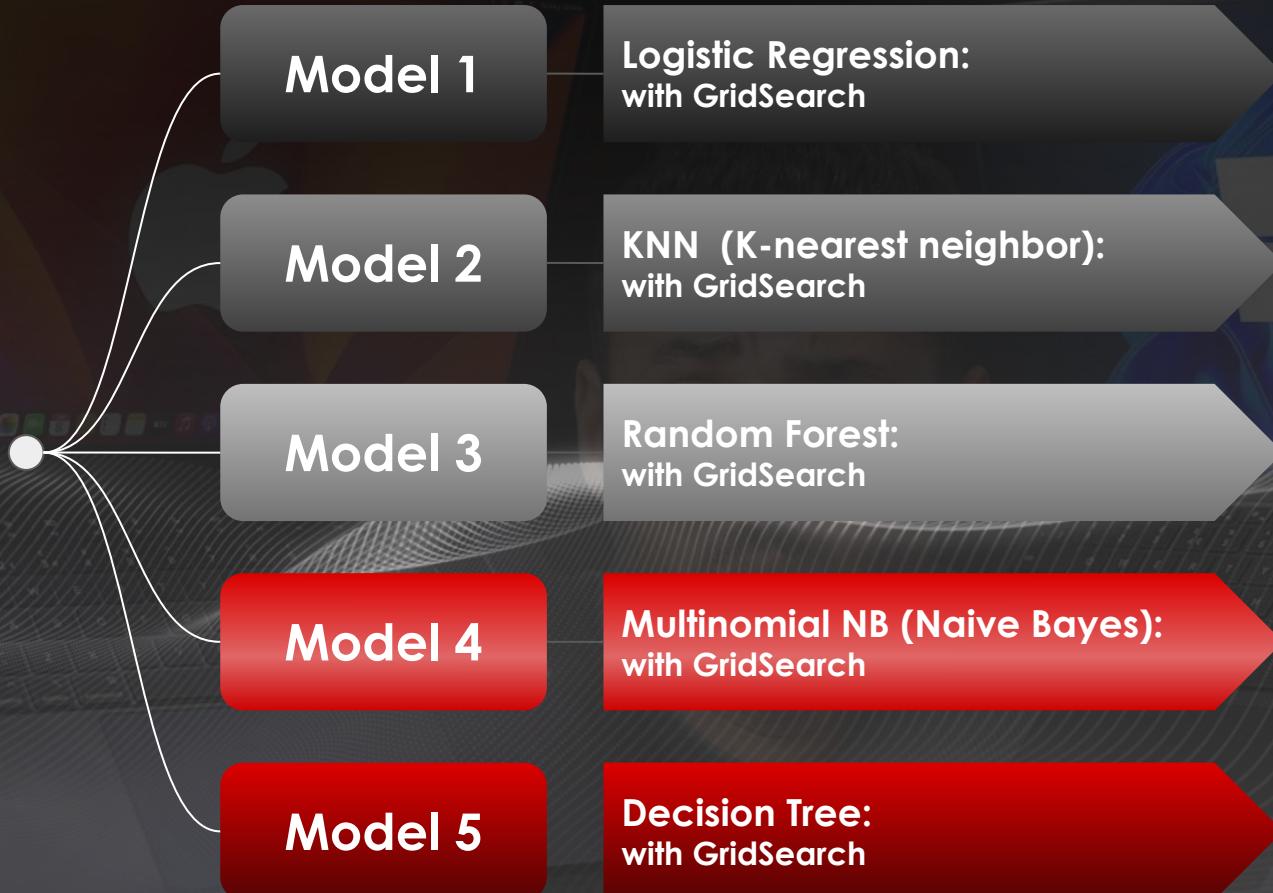
Message = Topic + Text (Body)



51:49  
Baseline

MacOS: 10,867 rows, 5 columns

Windows: 10,380 rows, 5 columns



# Model 1 : Logistic Regression

Metric	Score
Accuracy (Train Set)	0.99
Accuracy (Validation Set)	0.93
Recall (Mac)	0.92
Recall (Windows)	0.94
F1 Score*	0.93
Precision Score*	0.93

Weighted Average\*

Verdict

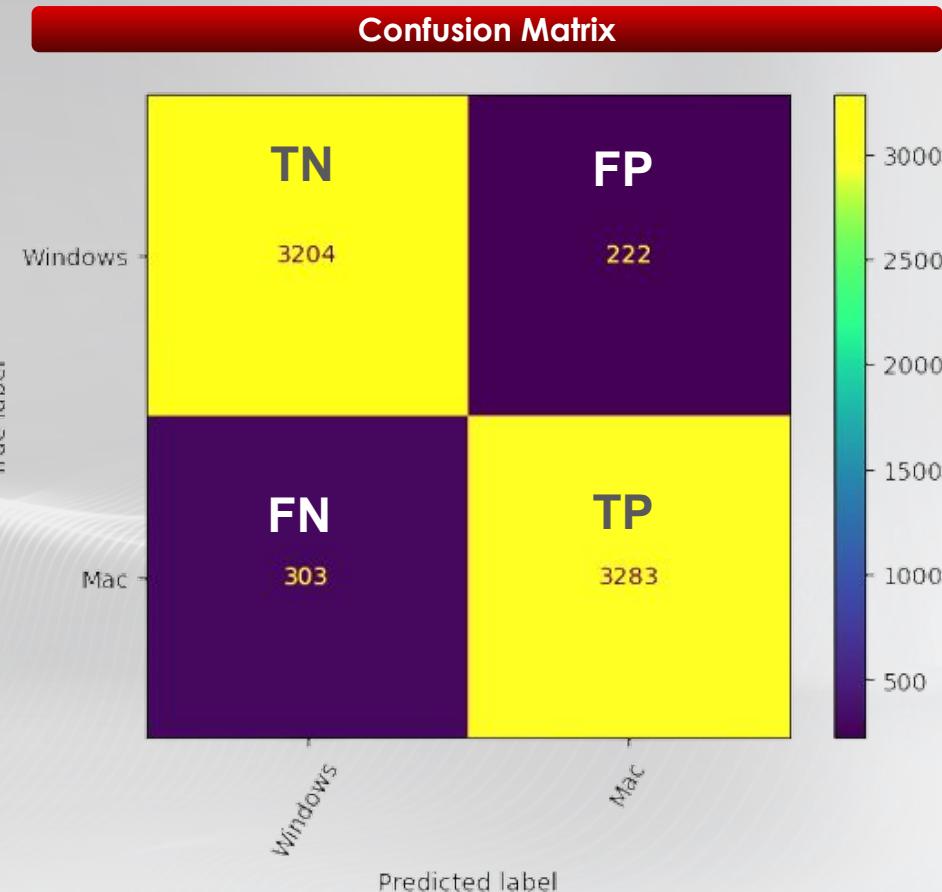
Overfit

Underfit

Just Right

Lack of precision

Confusion Matrix



# Model 2 : KNN (K-nearest neighbor)

Metric	Score
Accuracy (Train Set)	0.84
Accuracy (Validation Set)	0.76
Recall (Mac)	0.68
Recall (Windows)	0.85
F1 Score*	0.76
Precision Score*	0.77

Weighted Average\*

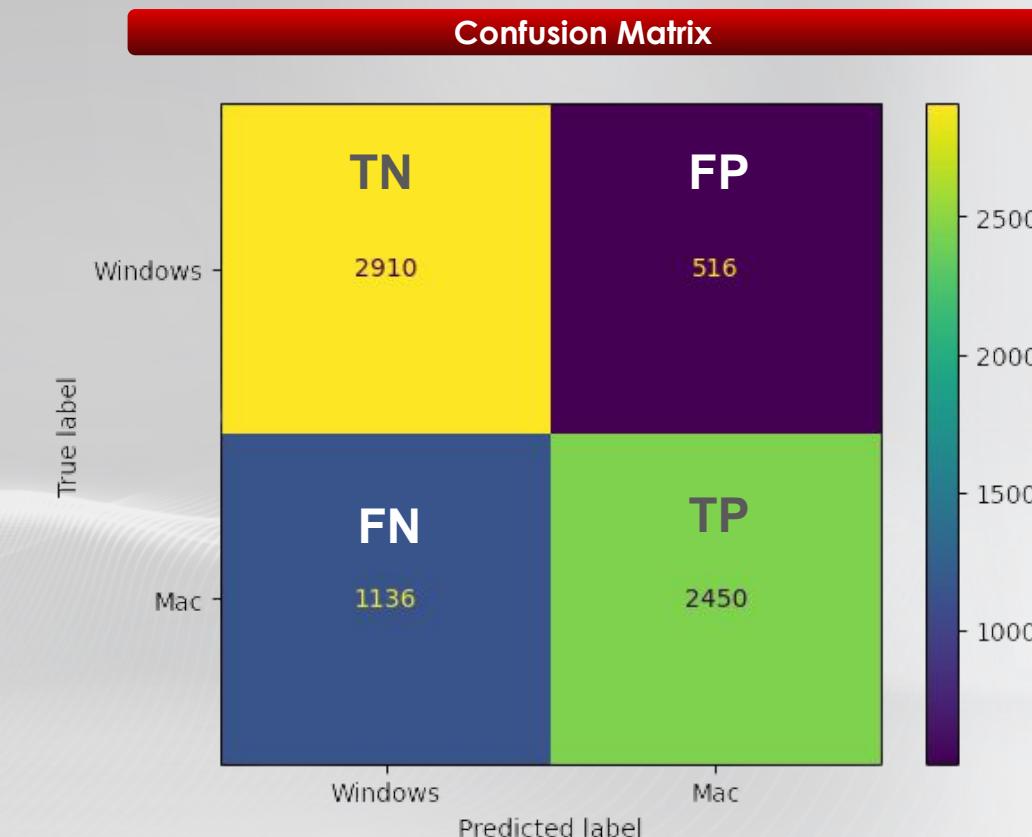
Verdict

Overfit

Underfit

Just Right

Lack of Precision



# Model 3 : Random Forest

Metric	Score
Accuracy (Train Set)	0.99
Accuracy (Validation Set)	0.93
Recall (Mac)	0.91
Recall (Windows)	0.95
F1 Score*	0.93
Precision Score*	0.93

Weighted Average\*

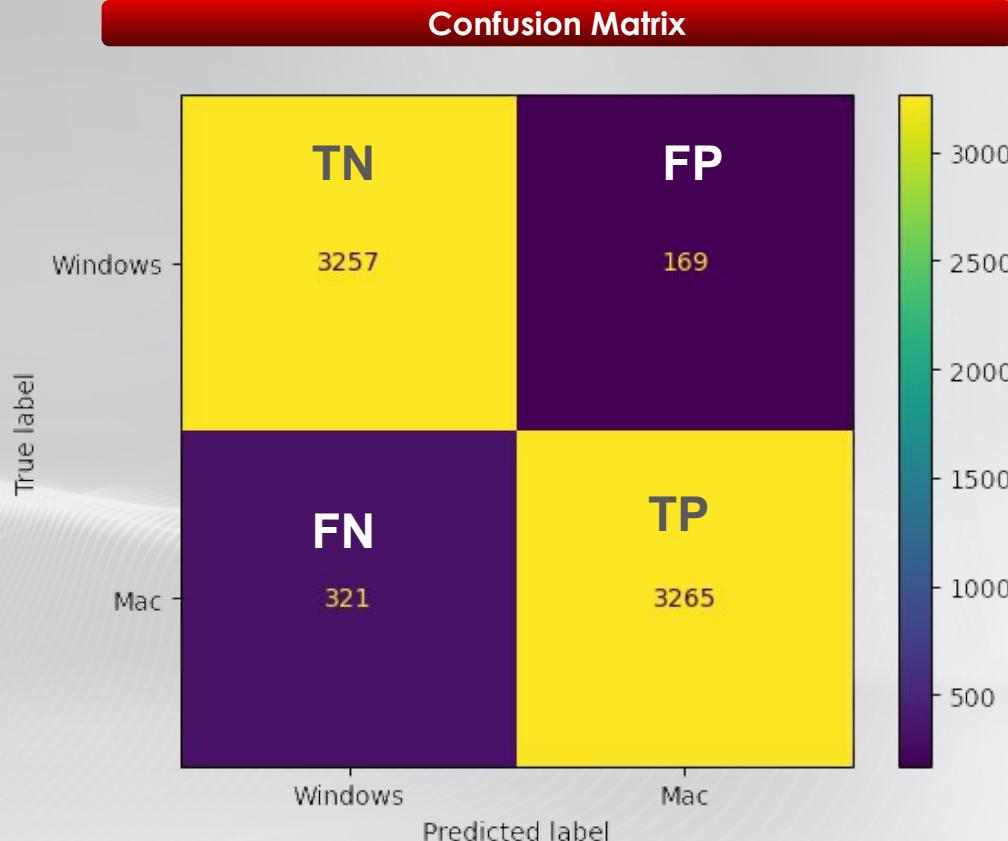
Verdict

Overfit

Underfit

Just Right

Lack of precision



# Model 4 : Multinomial NB (Naive Bayes)

Metric	Score
Accuracy (Train Set)	0.95
Accuracy (Validation Set)	0.96
Recall (Mac)	0.91
Recall (Windows)	0.96
F1 Score*	0.93
Precision Score*	0.94

Weighted Average\*

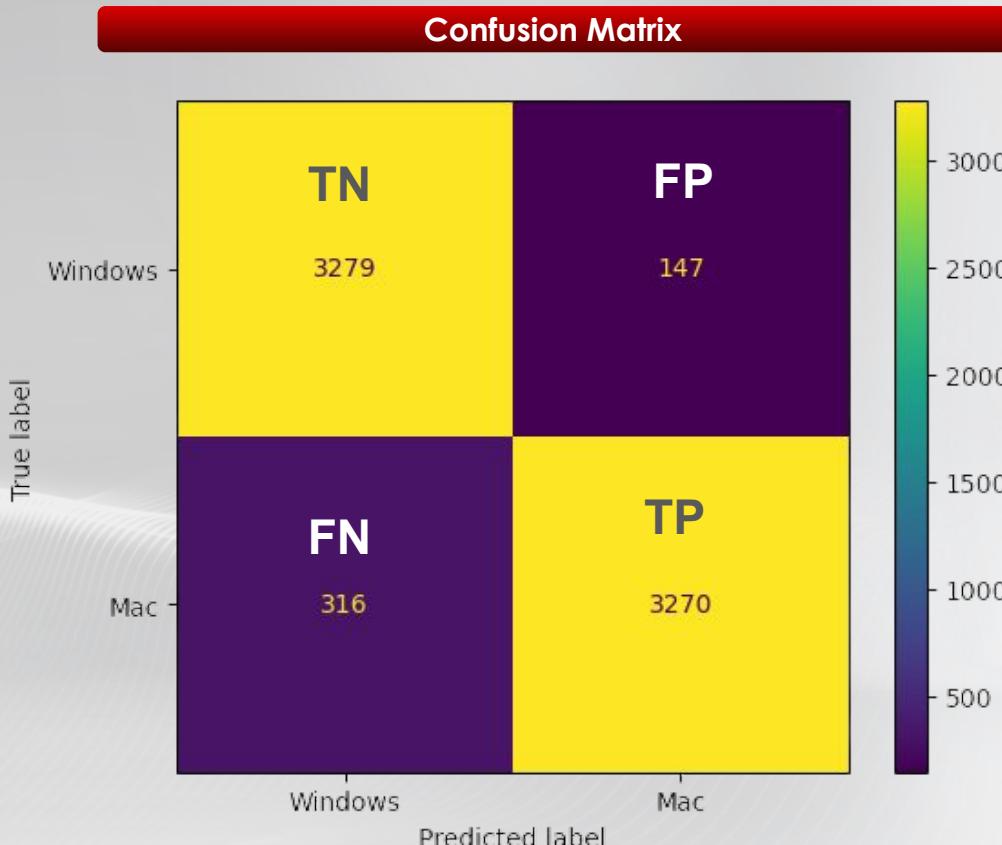
Verdict

Overfit

Underfit

Just Right

Lack of precision



# Model 5 : Decision Tree

Metric	Score
Accuracy (Train Set)	0.81
Accuracy (Validation Set)	0.80
Recall (Mac)	0.64
Recall (Windows)	0.97
F1 Score*	0.76
Precision Score*	0.84

Weighted Average\*

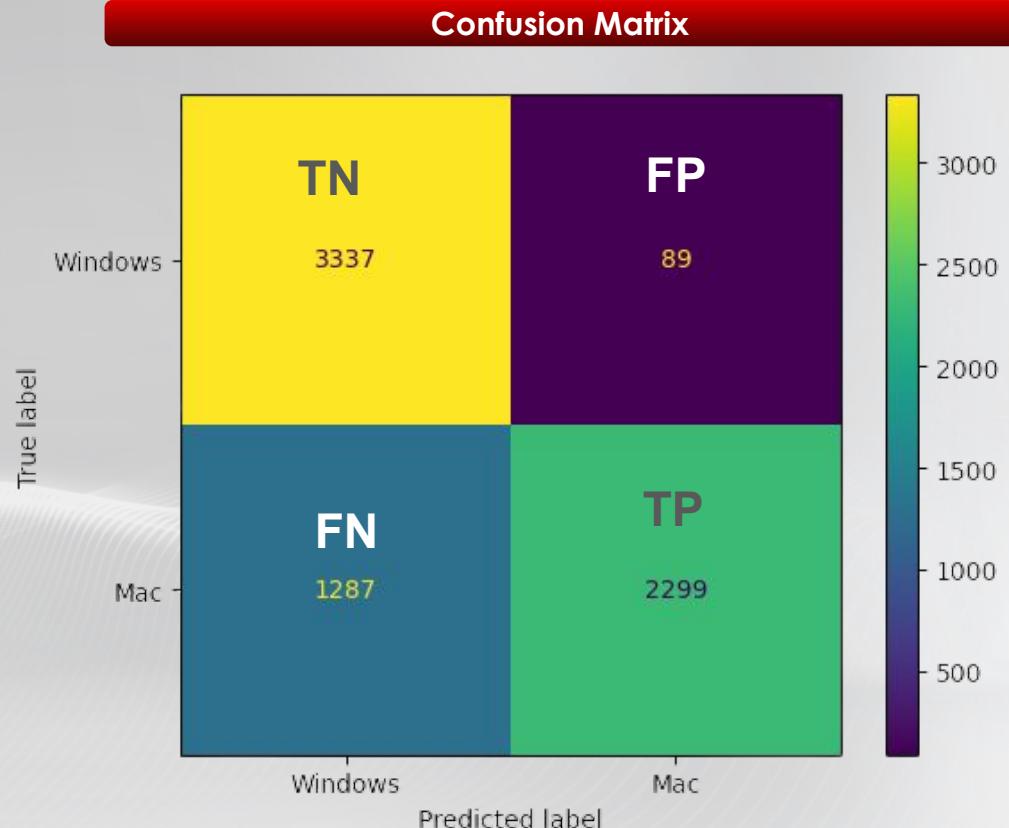
## Verdict

Overfit

Underfit

Just Right

Lack of Precision



# Model Performance

Metric	Model1 (LR)	Model2 (KNN)	Model3 (RF)	Model4 (NB)	Model5 (DT)
Accuracy (Train Set)	0.99 <small>Overfit</small>	0.84 <small>Overfit</small>	0.99 <small>Overfit</small>	0.95 <small>Underfit</small>	0.81 <small>Overfit</small>
Accuracy (Validation Set)	0.93	0.76	0.93	0.96	0.80
Recall (Mac)	0.92	0.68	0.91	0.91	0.64
Recall (Windows)	0.94	0.85	0.95	0.96	0.97
F1 Score*	0.93	0.76	0.93	0.93	0.76
Precision Score*	0.93	0.77	0.93	0.94	0.84



## Error Type I : Mentioning both Operation System in one post

Running Windows 10 along side Mo'jave? For those of you who are running Windows 10 on your Mac, do you find BootCamp to be an adequate solution or do you use Parallels?

Another question pertains to software that you use. For example, Office 365 is more fully developed with support for Access and Publisher on the Windows platform. However several applications run great on Mac. How would you recommend running Office 365 Home?

## Error Type II : User Error ( Posting on Wrong subreddit )

[Windows 7] How do you disable pointer speed buttons on mouses? Every time I buy an extra buttons mouse there's a button right under the scroll that Windows decides to change to a pointer speed/sensitivity adjusting button. I can usually assign it to something else in video games, but when I can't it becomes extremely annoying when I accidentally press it. Does anyone know how to disable this function?

## Business Recommendation

### **Multinomial Naive Bayes**

**Logistic Regression, Random Forest** are among the best model to classify such text with high prediction score.

Although, we suggest improve these 3 models to multi-classification,  
So they will be able to classify more than 2 classes.

Because audience interested in either Mac or Windows may also have other co-interests.

**(e.g. Windows users tend to also have interests in Android // Mac users also interested in iOS)**

So, we can classify more and extend topics of interests to write and talk about on our news while capturing the same audience with variety of topics while still establishing strong brand positioning presence.

## Further Studies

We could try these following to improve the model,

Neatly clean our dataset since there are slight chance of data contamination (mislabelling post)

Mix with text sources from other platform to stimulate our model with environment our unseen data could possibly come from

# Thank you