MacOS & Windows
APIs & NLP Text Classification

Data Science Immersive Course
Project 3

Problem Statement

Scenario

Given that
We work for a Tech News & Media Company
that writes/talks about Computer Technology update mostly for Mac and Windows, With short in manpower, Our Editors usually choose variety of IT buzzing and popular topics from social media post to write/talks in our news sometime just paraphrasing.

Tech News company needs to think like a tech company,
Our CEO trying to find a tool to effectively help our editors get more productivity and focus more in quality content without getting more headcount by creating tool to **auto-categorize** our news articles on our website whether it is Windows or Mac. So our editors don't have to input it manually and have time to focus more on quality of content. We are in charge of this project!

Our Plan

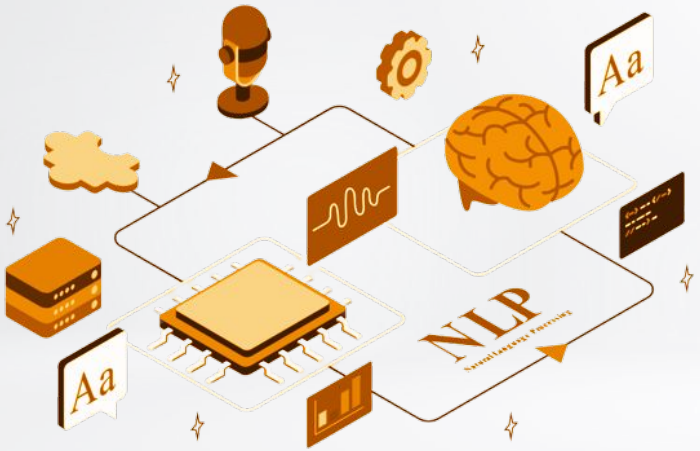
Find out

1. **Find Most Precise Classification Model** to help categorize text into ‘Mac’ or ‘Windows’

2. **Recommendation** for further studies and Business strategy/execution suggestion

1. **Most Precise Classification Model**
to help categorize text into 'Mac' or 'Windows'

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.**

Methodology Process

Reddit

2 Datasets

Scraped Text from Reddit
Period: Year 2010-2022

MacOS: 19,312 rows, 4 columns
Windows: 28,209 rows, 4 columns

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)

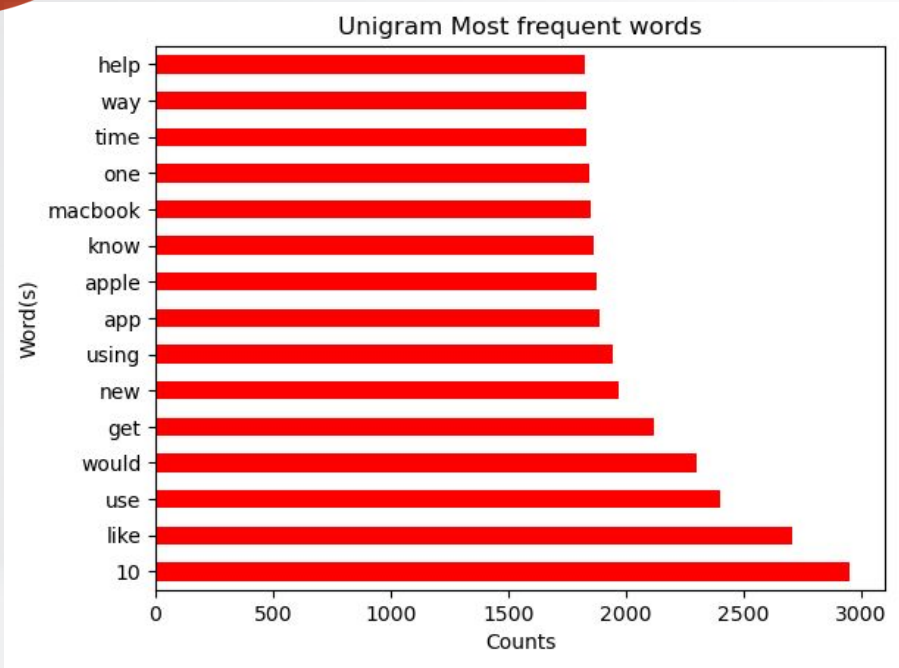
51:49
Baseline

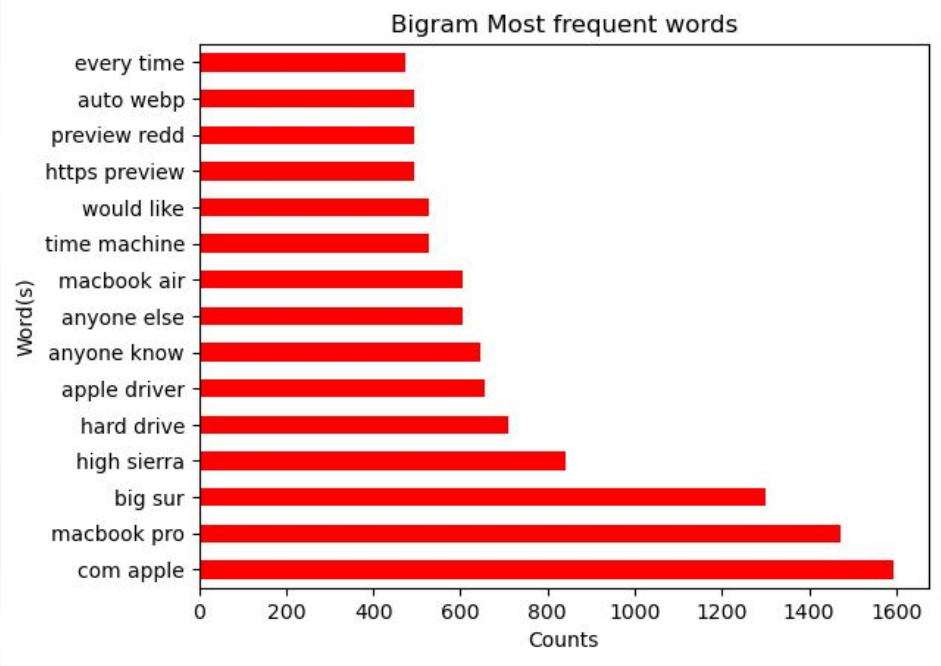
MacOS: 10,867 rows, 5 columns
Windows: 10,380 rows, 5 columns

Feature Engineering
using NLP Technique:
Bag of words
*Excludes stopwords and class name (Mac,Win)

Modeling & Prediction

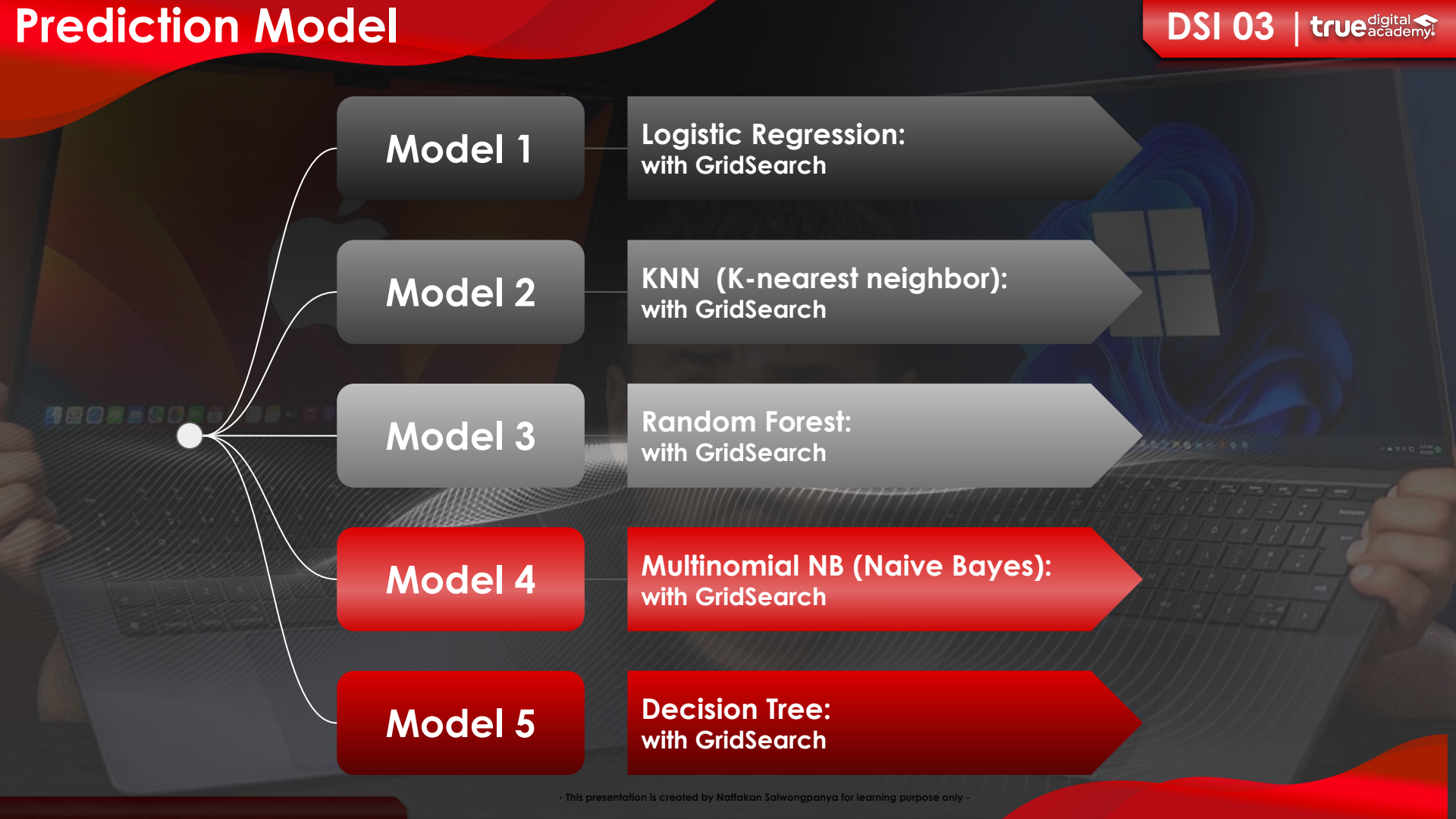
Exploratory Data Analysis





***Unigram and Bigram will be used on Bag of Words (NLP technique)**

Prediction Model



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

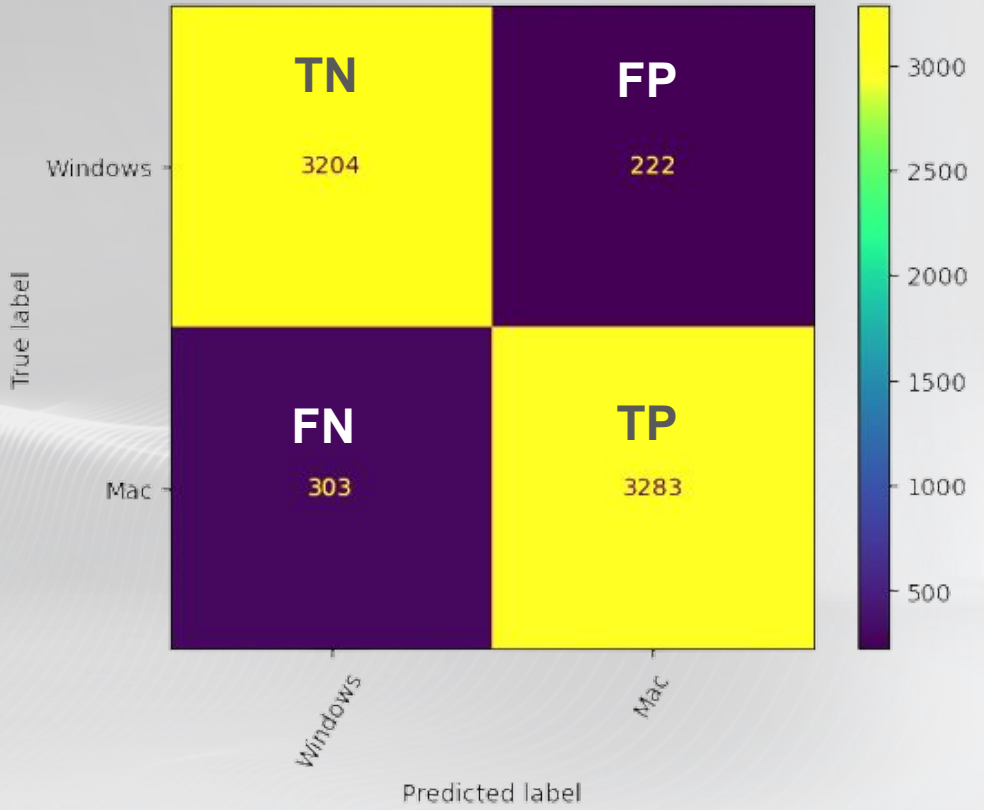
Overfit

Underfit

Just Right

Lack of precision

Confusion Matrix

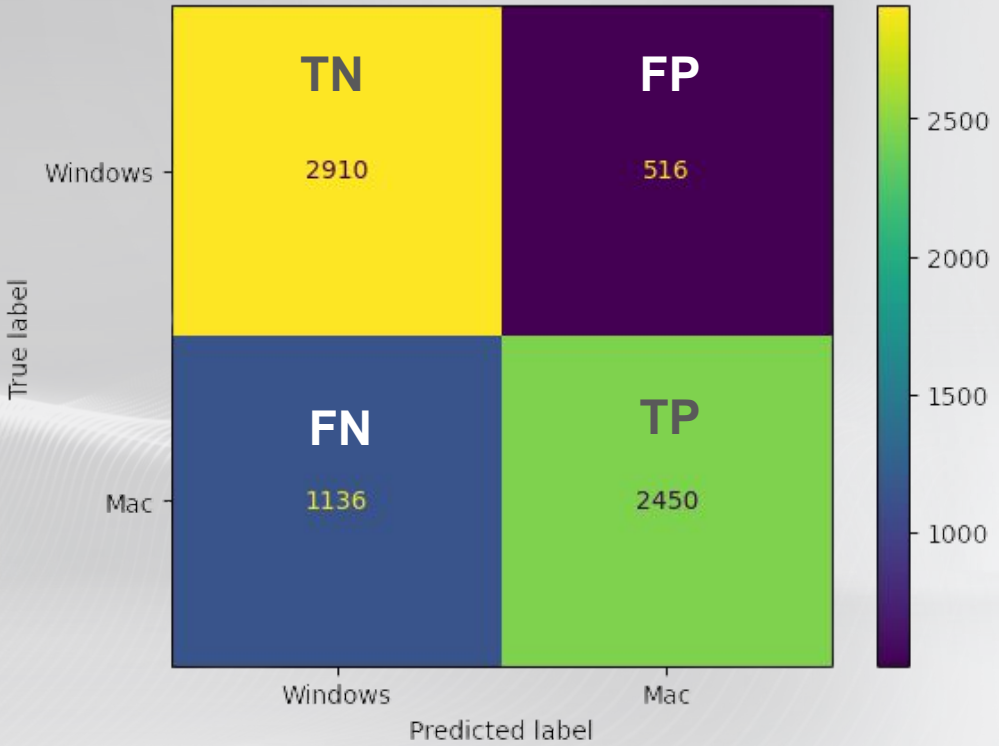


Model 2 : KNN (K-nearest neighbor)

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*

Confusion Matrix



Verdict

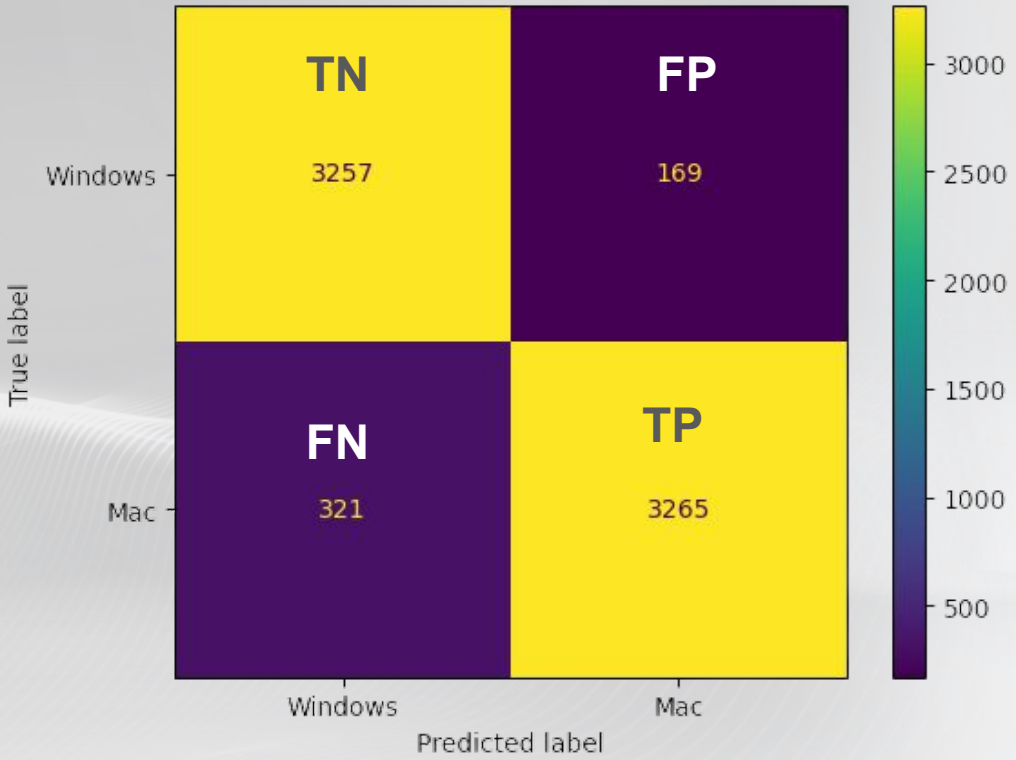
- **Overfit**
- Underfit
- Just Right
- **Lack of Precision**

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*

Confusion Matrix



Verdict

Overfit

Underfit

Just Right

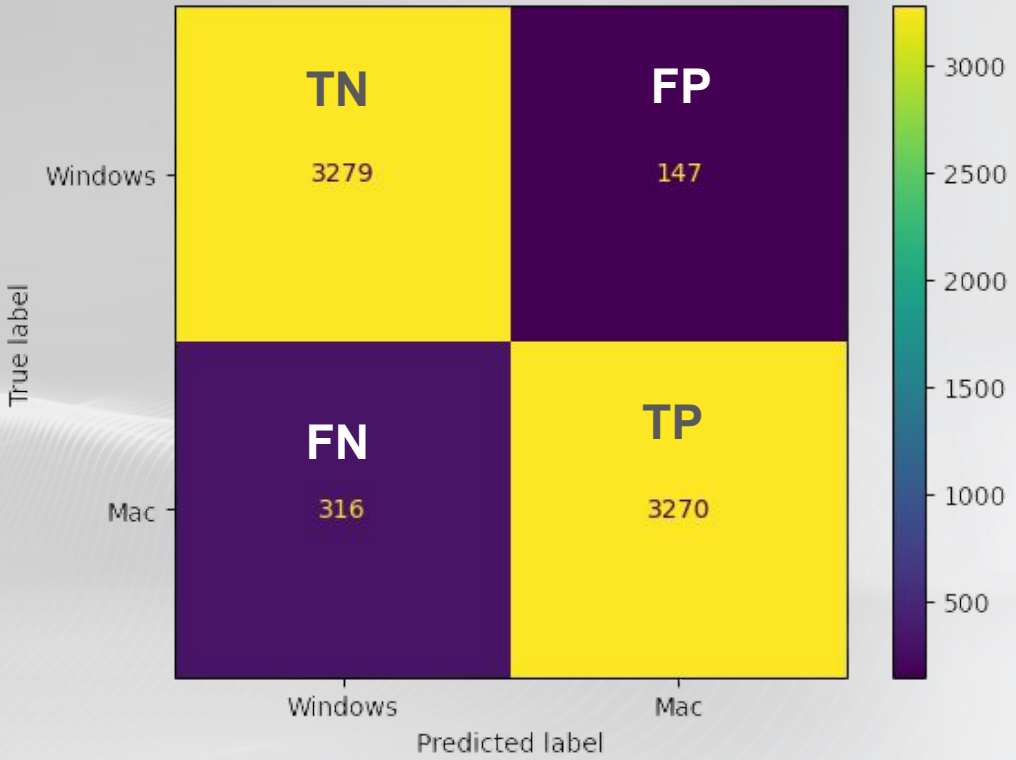
Lack of precision

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*

Confusion Matrix



Verdict

Overfit

Underfit

Just Right

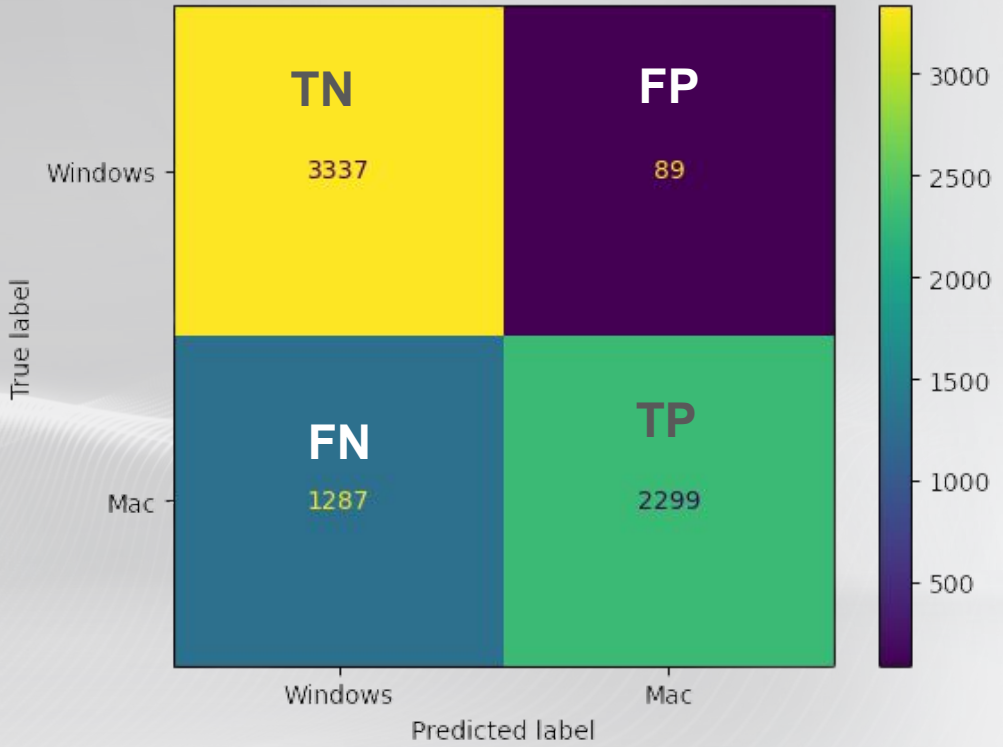
Lack of precision

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*

Confusion Matrix



Verdict

Overfit

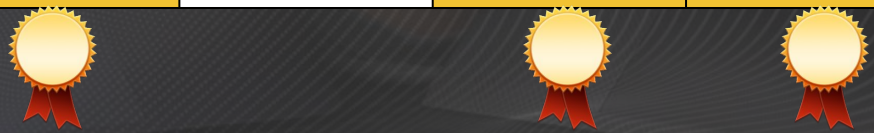
Underfit

Just Right

Lack of Precision

Model Performance

Metric	Model1 (LR)	Model2 (KNN)	Model3 (RF)	Model4 (NB)	Model5 (DT)
Accuracy (Train Set)	0.99 Slightly Overfit	0.84 Overfit	0.99 Slightly Overfit	0.95	0.81 Underfit
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 Analysis

Common Error 1: 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?

Common Error 2: 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?

(Windows context post on Mac subReddit)

Conclusion

Business Recommendation

Multinomial Naive Bayes, Logistic Regression, Random Forest are among the best models to classify such text with high accuracy score and ready to deploy its first trial.

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

Because we could expand our topics to write more than just Mac or Windows because mobile operating systems are rising its popularity these day! And this could be new business opportunity. That way, also can **auto-categorize more categories** and get broader topics of interests to write/talk more about on our news article.

Further Studies

We could try these following to improve the model,
Neatly clean our dataset since there are slightly chance of data contamination (mislabelling post)
Mix with text sources from other platform to stimulate our model with various environment for our unseen data could possibly based on and get more generalizable.

Thank you