SENTIMENT ANALYSIS OF ELECTRONIC PRODUCTS IN Lazada

FINAL PROJECT
GROUP 2

Meet Our Teams

Members:
Denis Irham
Bashir Ammar Hakim

Mentor:
Azizur Rachman

TABLE OF CONTENTS

01 INTRODUCTION

Backgorund – Goals – Benefits – Tools- Timeline

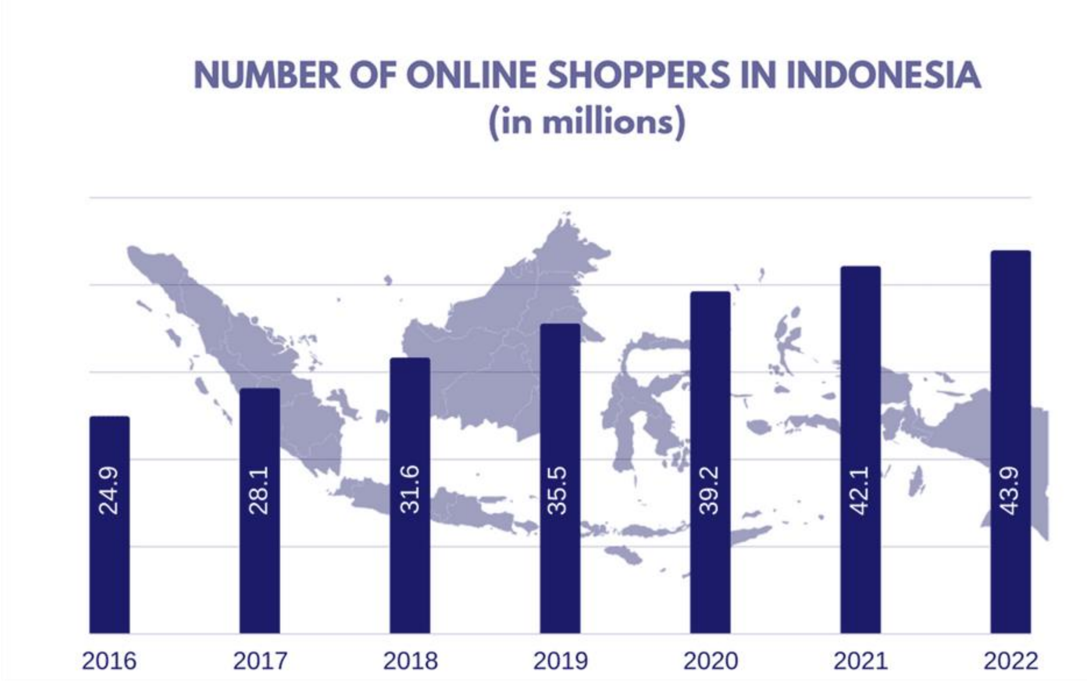
02 METHODOLOGY

Data Understanding – Data Preprocessing - EDA – NLP - Modelling

03 CONCLUSION

Evaluation – Conlusion - Literature

Introduction - Background



As the E-commerce markets in Indonesia is getting bigger every year, **product feedbacks** play the bigger roles in increasing the **revenue** of the sellers. This can be caused by the **change behavior** of consumer to **compare** products that they want to buy first from the **reviews** and prioritize the **services** they get from the seller. Hence, the sentiment analysis is needed to monitor and find the particular **aspects** of product from the reviews that people are expressing in positive, neutral, or negative way.

Introduction - Goals

Model

Build a sentiment analysis model that can interpret customer feedbacks by defining the feedback in some of the aspects.

Insights

Provide insights of what features of product and services that needed to tailored to meet their customer needs.

Benefits

1

A reference for sentiment analysis in other e-commerce or social media

2

Applicable for the recommendation system in e-commerce

3

Improve customer services of the product

4

Increase sales revenue

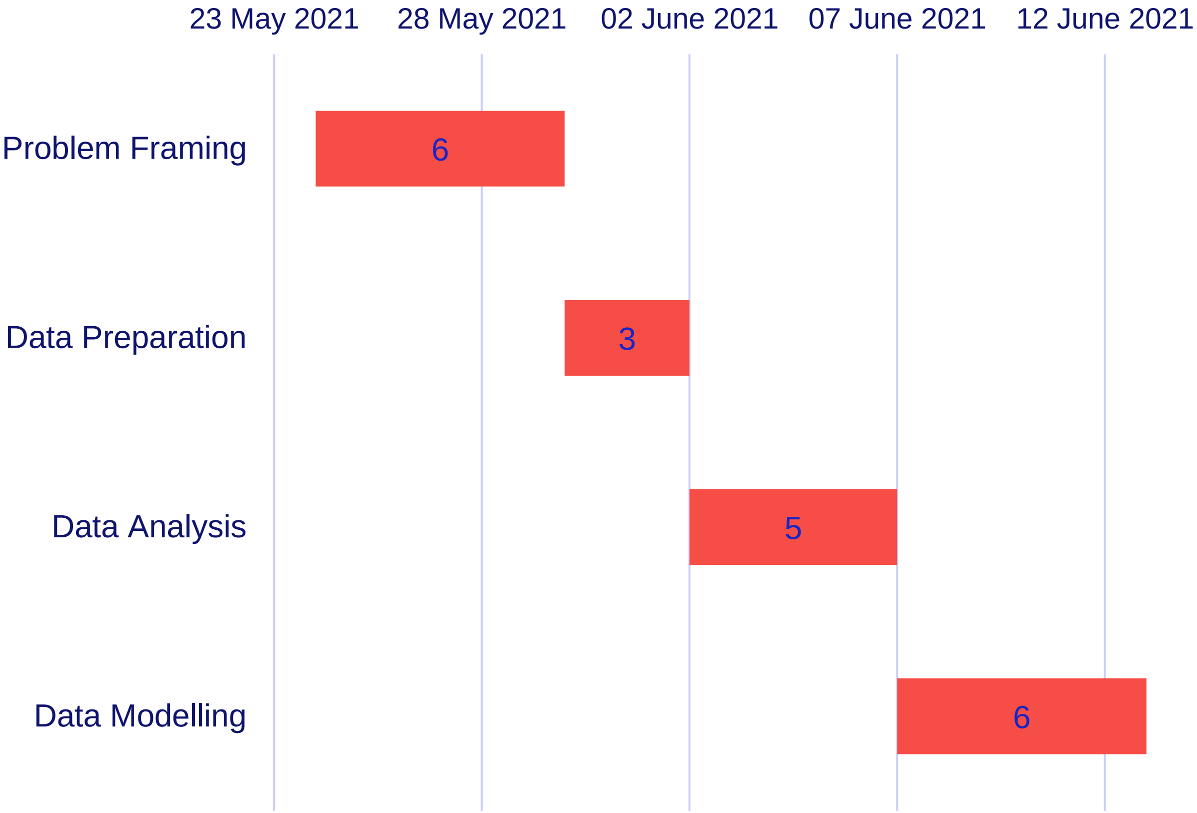
5

Develop quality products

6

Improve crisis management

Timeline


23 May 2021
28 May 2021
02 June 2021
07 June 2021
12 June 2021
Problem Framing
6
Data Preparation
3
Data Analysis
5
Data Modelling
6

Tools

Jupyter Notebook

Pandas

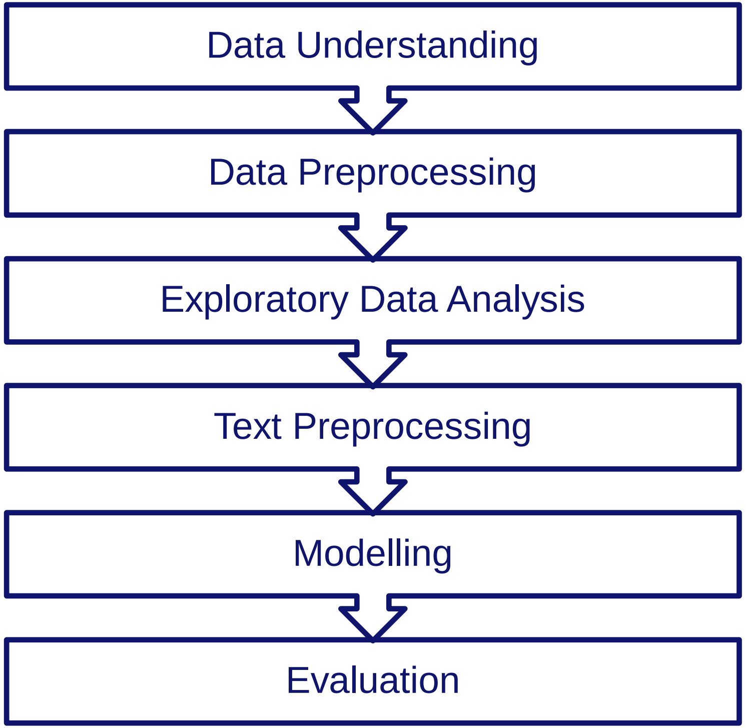
Matplotlib

Scikit-Learn

NLTK

Sastrawi

Workflow



Data Understanding

Data Preprocessing

Exploratory Data Analysis

Text Preprocessing

Modelling

Evaluation

Workflow

Data Understanding

- Explore and understanding variables of the dataset

Data Preprocessing

- Preparing the dataset for analysis and modelling

Exploratory Data Analysis

- Making the visualisation of the dataset to gain business understanding
- Investigate the shape of the dataset and number of samples

Text Preprocessing

- Cleaning the dataset by handling the null values
- Handle the text dataset by the process of NLP

Modelling

- Selecting the machine learning method suited with the dataset

Evaluation

- Predict the test dataset
- Evaluating the model with performance metrics

Data Understanding

Lazada Indonesian Reviews

Product reviews from Lazada Indonesia based on 5 categories.
The dataset were divided into 2 items:

- 20191002-items.csv
- 20191002-reviews.csv

Source: https://www.kaggle.com/grikomsn/lazada-indonesian-reviews?select=categories.txt

beli-laptop

beli-smart-tv

shop-televisi-digital

jual-flash-drives

beli-harddisk-eksternal

Dataset
Lazada Indonesian Reviews

20191002-items.csv

Column Names	Definition	Data Type
itemid	Id of item	Int64
category	Category in e-commerce	Object
name	Title of the item	Object
brandname	Brand name of item	Object
url	Link to buy the item	Object
price	Price of item	Int64
averagerating	Average rating of item	Int64
totalreviews	Total Reviews of item	Int64
retrieveddate	Date of last calculation	Object

Dataset
Lazada Indonesian Reviews

20191002-reviews.csv

Column Names	Definition	Data Type
itemID	Id of item	int64
category	Category in e-commerce	object
name	Title of the item	object
rating	Rating by user	int64
originalRating	-	float64
reviewTitle	Title of review	object
reviewContent	Content of review	object
likeCount	Total Likes	int64
upVotes	Total upVotes	int64
downVotes	Total downvotes	int64
helpful	1 : Condition is helpful 0: Condition isnt helpful	bool
relevanceScore	Relevance score	float64
boughtDate	Date of transaction	object
clientType	Type of application	object
retrievedDate	Date of calculation	object

Data Preprocessing

- Merge Dataset
  - Join two dataset with keys itemId and category resulting in 203787 rows x 22 columns. Eventhough its not related to the NLP, but it can be used for EDA

```
df =pd.merge(data, data1, how="inner",on=["itemId",'category'])
```

- Handling Missing Values
  - Drop the missing values in reviewContent for sentiment analysis

```
df=df[df['reviewContent'].notna()]
```

itemId	category	brandName	price	...	rating	...	reviewContent	...

Data Preprocessing

- Handling Outliers
  - There were some outliers in the price distribution, so we choose to investigate more about it and found the review was full of joke in rangeof price >Rp 40.000.000

```
df.reviewContent[df.price==df.price.max()]
```

Reviews

SEBUT SAJA MAWAR

“Akhirnya kebeli juga berkat jual GINJAL keponakan”

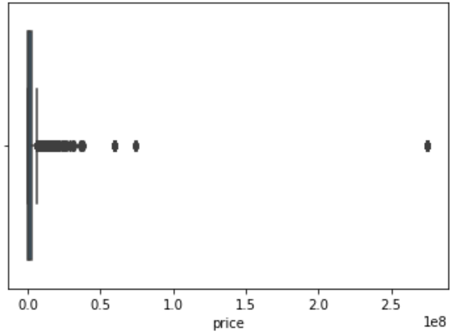
SEBUT SAJA BUDI

“Barang nya udh dtg,pesen tadi ehh nyampe nya kemaren,sekarang dipake cuma buat alas tenis meja,thxx…”

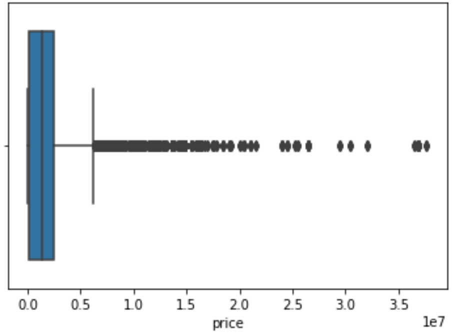
Data Preprocessing

- Handling Outliers
  - So we choose to remove the product that has the range of price > Rp 40.000.000

```
df = df[df.price < 40000000]
sns.boxplot(df.price)
```



From this



To this

Data preprocessing

- Set Samples

  We reduce the amount of dataset that will used for modelling to shorten the computer processing. The dataset became an imbalance so we set the sample to be equal in amount that is 5000

```
s1=df[df.label==1].sample(5000, replace=True)
s2=df[df.label==0].sample(5000, replace=True)

data=pd.concat([s1,s2])
```

	col	row
Label 1	1	183583
Label 0	1	20204

	col	row
Label 1	1	5000
Label 0	1	5000

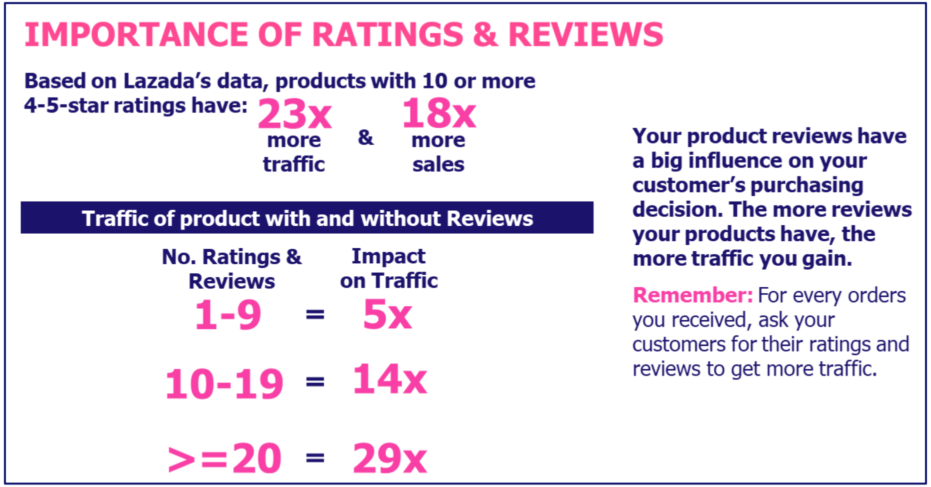
Data preprocessing

- Set label

  We set the satisfaction rate to be positive and negative. The positive label will be filled with 5 and 4 in rating and negative will be the rest.

```
label=[]
for index, row in df.iterrows():
    if row['rating']==5:
        label.append(1)
    elif row['rating']==4:
        label.append(1)
    else:
        label.append(0)

df['label']=label
df.head()
```

Data preprocessing



IMPORTANCE OF RATINGS & REVIEWS

Based on Lazada's data, products with 10 or more 4-5-star ratings have: **23x** more traffic & **18x** more sales

Traffic of product with and without Reviews

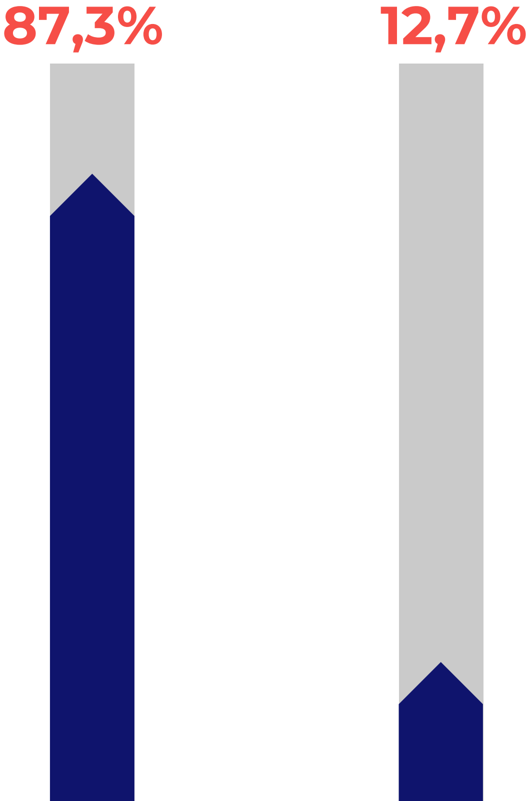
No. Ratings & Reviews		Impact on Traffic
1-9	=	5x
10-19	=	14x
>=20	=	29x

Your product reviews have a big influence on your customer's purchasing decision. The more reviews your products have, the more traffic you gain.

Remember: For every orders you received, ask your customers for their ratings and reviews to get more traffic.

- Positive Seller Rating
  - Get higher rank on Lazada's SEO
  - Chance to participated on event and campaign held by Lazada
  - Chance to join priority program
- Increase GMV of SKU
- 10 rejections = 1 bad review

Exploratory Data Analysis



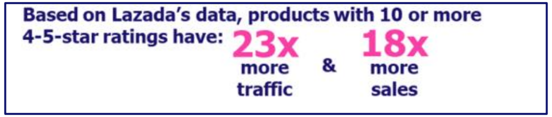
Label 1

Satisfied with the product

Label 0

Disappointed with the product

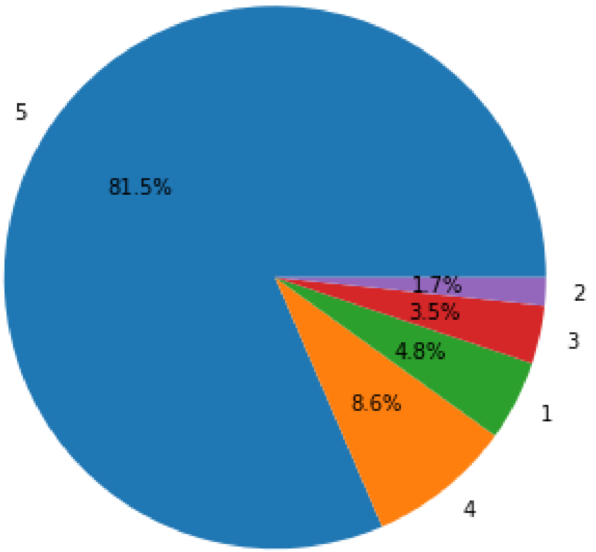
Exploratory Data Analysis



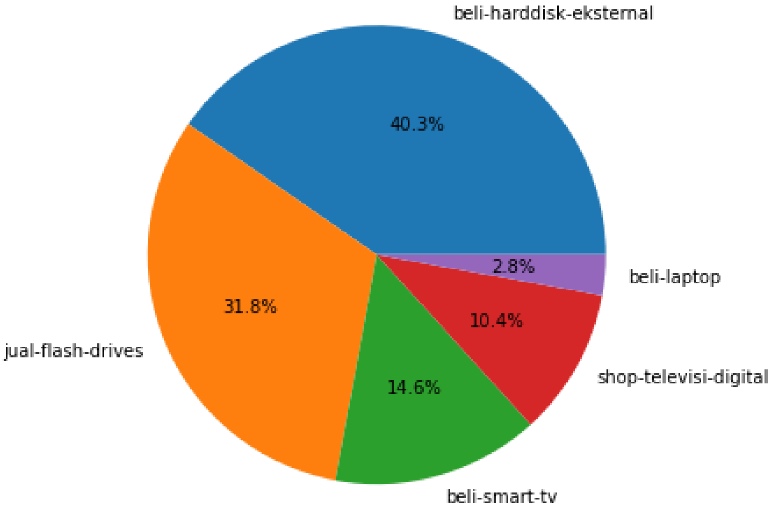
Based on Lazada's data, products with 10 or more 4-5-star ratings have: **23x** more traffic & **18x** more sales

- Positive Seller Rating
  - Get higher rank on Lazada's SEO
  - Chance to participated on event and campaign held by Lazada
  - Chance to join priority program
- Increase GMV(Gross Merchant Values) of SKU(Stock Keeping Unit)
- 10 rejections = 1 bad review

Rating pie chart



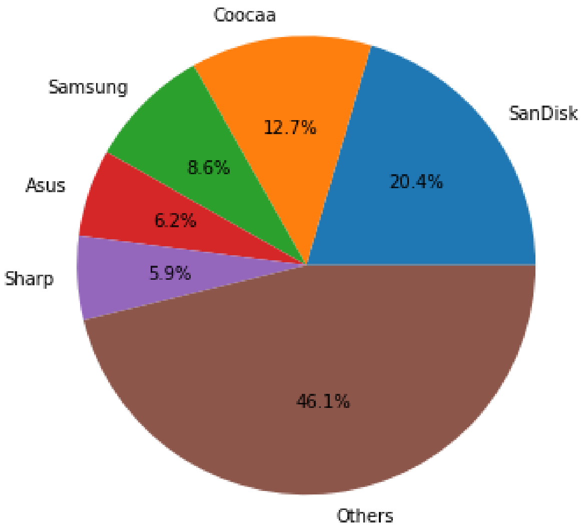
Exploratory Data Analysis



Category

Most of the categories exists in the dataset were ‘beli-harddisk-eksternal’. But it does’nt reflect the item in the dataset because there were many seller use the unrelated categories too.

Exploratory Data Analysis



Brand

There were 235 brands in the dataset and the best seller brand is SanDisk

Exploratory Data Analysis

COOCA LED TV 40 INCH

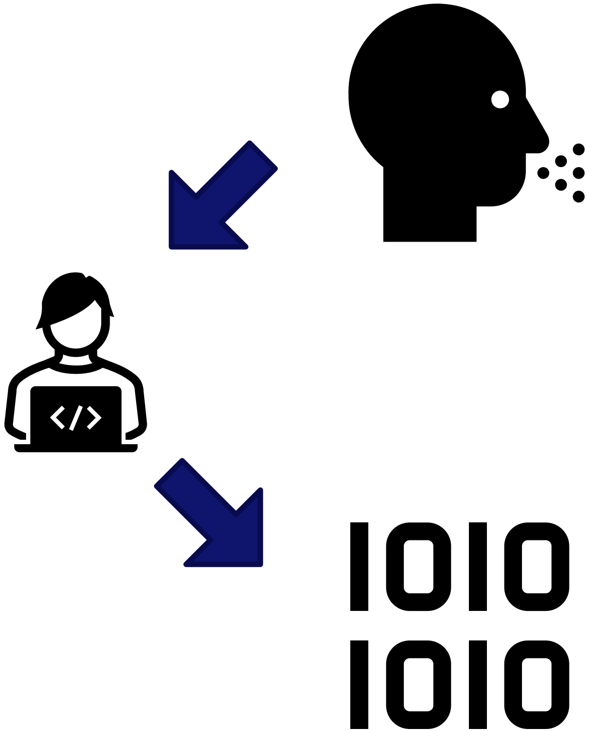
[GRATIS ONGKIR] –
FULL HD PANEL -SLIM

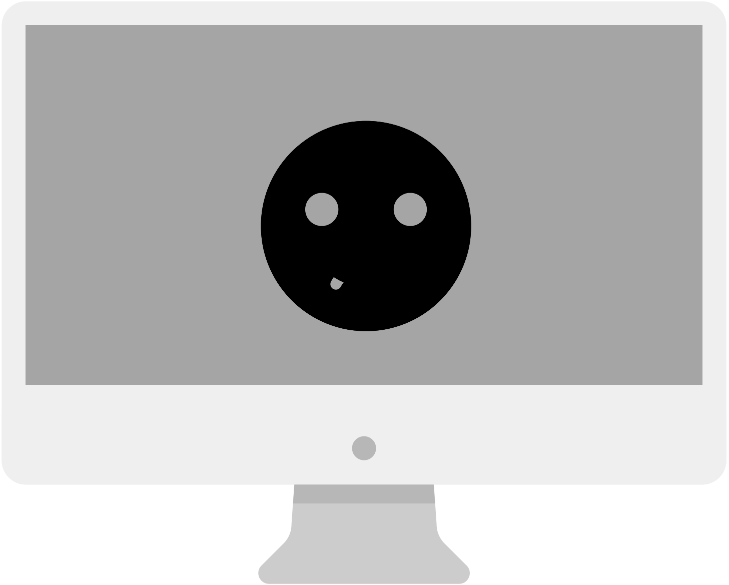
3952 units sold

COOCA LED TV 24 INCH

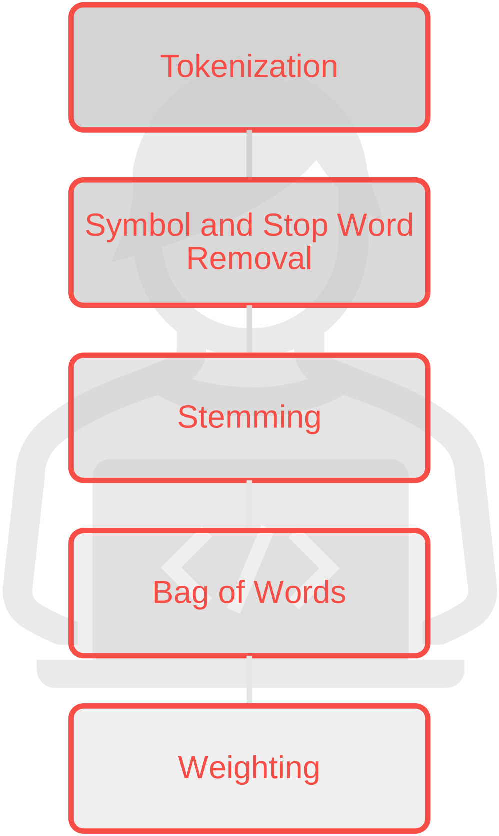
[GRATIS ONGKIR] –HD
PANEL- SLIM

3540 units sold


1010
1010



Text Preprocessing



Tokenization

Symbol and Stop Word Removal

Stemming

Bag of Words

Weighting

Text Preprocessing

- Tokenization
  - Splitting sentence into words.

```
import re
df.reviewContent=df.reviewContent.apply(lambda x: re.split(r'\s+', x))
```

Example:

```
['barang bagus banget',
'barang rusak , packing buruk banget',
'barang oke, tapi pengiriman buruk banget',
'packing rapi , pengiriman cepat dan aman']
```

↓

```
[['barang', 'bagus', 'banget'],
 ['barang', 'rusak', ',', 'packing', 'buruk', 'banget'],
 ['barang', 'oke,', 'tapi', 'pengiriman', 'buruk', 'banget'],
 ['packing', 'rapi', ',', 'pengiriman', 'cepat', 'dan', 'aman']]
```

Text Preprocessing

- Remove unimportant text

  We removed some unwanted text that unrelated to the sentiment analysis and summarizing the text.

  ○ Remove punctuation and numbers

```
df.reviewContent = df.reviewContent.apply(lambda x: x.lower())
df.reviewContent = df.reviewContent.apply(lambda x: re.sub(r'([^a-z\s]+)','',x))
```

  ○ Remove Stopwords

```
from nltk.corpus import stopwords
from string import punctuation
stop_words=stopwords.words("indonesian")+list(punctuation)
df.reviewContent=df.reviewContent.apply(lambda x: [w for w in x if not w in
                                        stop_words])
df.reviewContent=df.reviewContent.apply(lambda x: str(' '.join(x)))
```

Example:

'Paket sudah sampai, sudah dicoba dan berfungsi dengan baik, semoga awet dan tidak ada kendala.'

↓

Paket sampai, dicoba berfungsi baik, semoga awet kendala.'

Text Preprocessing

- Stemming

  Reducing the affixes and suffixes to get the root of words.

```
from Sastrawi.Stemmer.StemmerFactory import StemmerFactory
factory = StemmerFactory()
stemmer = factory.create_stemmer()
df.reviewContent=df.reviewContent.apply(lambda x:stemmer.stem(x))
```

Example:

'Paket sudah sampai, sudah dicoba dan berfungsi dengan baik, semoga awet dan tidak ada kendala.'

'Mereka Meniru-nirukannya.'

‘paket sudah sampai sudah coba dan fungsi dengan baik moga awet dan tidak ada kendala’
‘mereka tiru’

Text Preprocessing

- Train Test Split

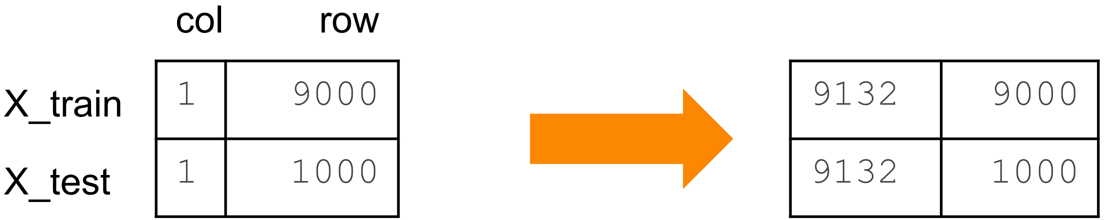
  Split the dataset to be train and test dataset

```
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(dfr.reviewContent,
                                        dfr['label'], test_size=0.1, random_state=30)
```

- Bag of Words

  The bag-of-words model is a popular and simple feature extraction technique used when we work with text. It describes the occurrence of each word within a document.

```
from sklearn.feature_extraction.text import CountVectorizer

cv = CountVectorizer()
bog= cv.fit_transform(dfr.reviewContent)
```



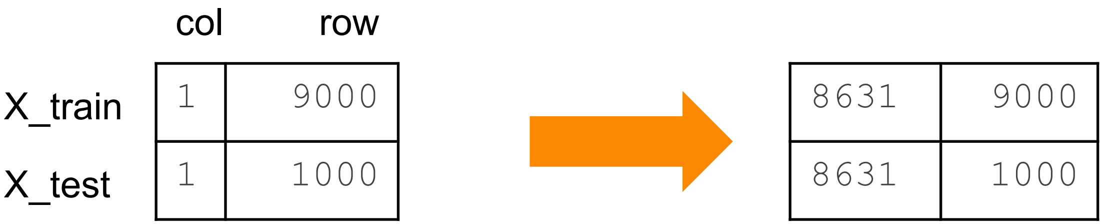
	col	row
X_train	1	9000
X_test	1	1000

9132	9000
9132	1000

Text Preprocessing

- TF-IDF
  Using TF-IDF to transform sentence into vector that can be applicable for machine learning model. In information retrieval, tf–idf, TF*IDF, or TFIDF, short for term frequency–inverse document frequency, is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus.

```
from sklearn.feature_extraction.text import TfidfVectorizer

tf = TfidfVectorizer()
X_train = tf.fit_transform(X_train)
X_test = tf.transform(X_test)
```



	col	row
X_train	1	9000
X_test	1	1000

→

8631	9000
8631	1000

Text Preprocessing

Examples of Bag of Words and TF-IDF

```
df =[
    'barang bagus banget',
    'barang rusak , packing buruk banget']
```

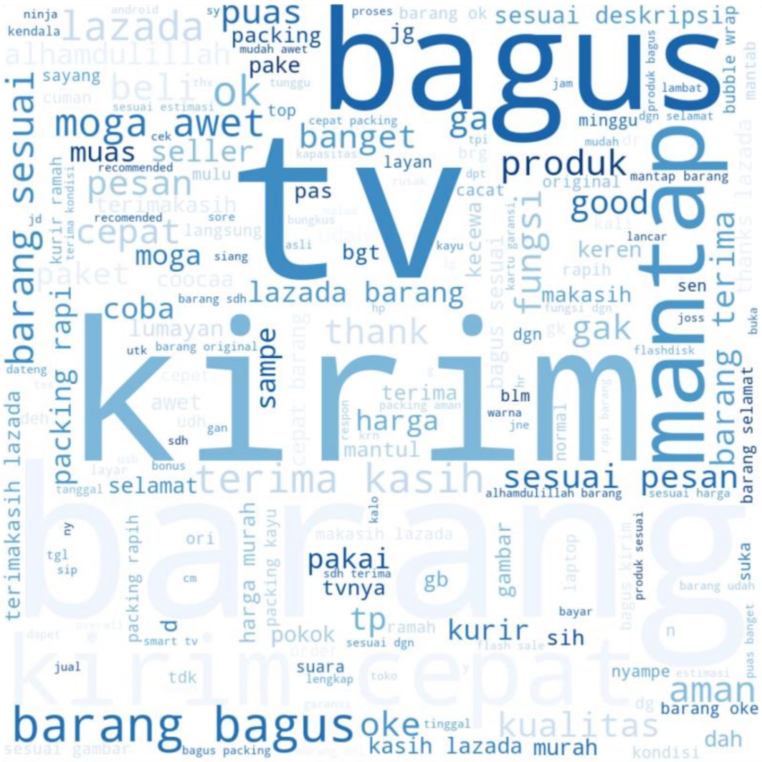
Bag of Words

```
['bagus', 'banget', 'barang', 'buruk', 'packing', 'rusak']
[[1 1 1 0 0 0]
 [0 1 1 1 1 1]]
```

TF-IDF

```
'bagus', 'banget', 'barang', 'buruk', 'packing', 'rusak']
[[0.70490949   0.50154891   0.50154891   0.           0.           0.]
 [0.           0.35520009   0.35520009   0.49922133   0.49922133   0.49922133]]
```

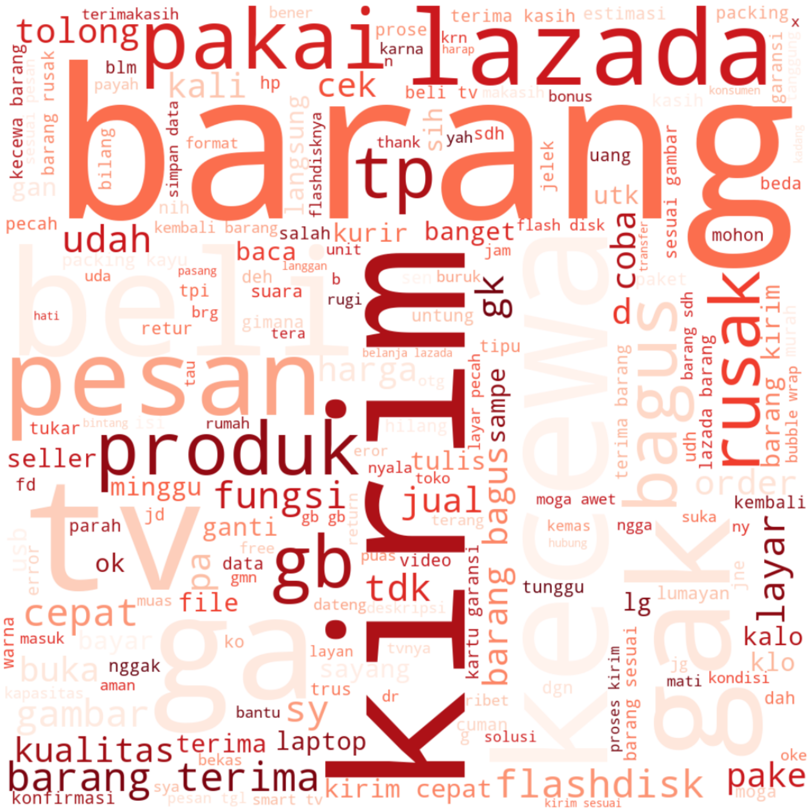
Text Preprocessing



Word Visualisation

- For label 1

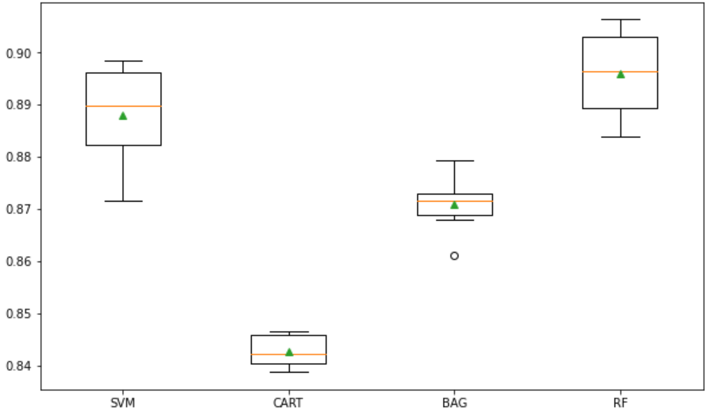
Text Preprocessing



- Word Visualisation
  - For label 0

Modelling - Comparation

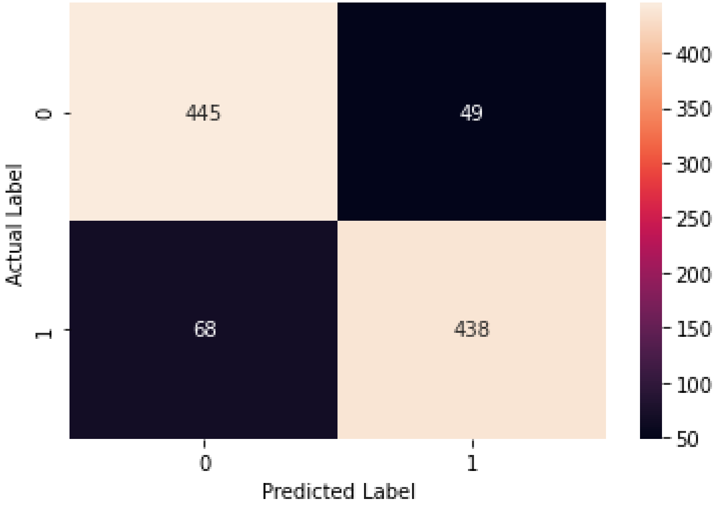
```
cv = RepeatedStratifiedKFold(n_splits=2, n_repeats=3, random_state=1)
```



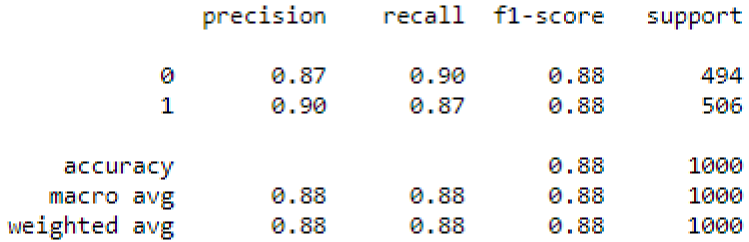
- Decision Tree
  AUC : 0.843
- SVM
  AUC : 0.888
- Bagging
  AUC : 0.871
- Random Forest
  AUC : 0.896

Modelling – Confusion Matrix

```
sns.heatmap(confusion_matrix(y_test, pred),annot=True,fmt=‘g’)
```



Modelling – Classification Report

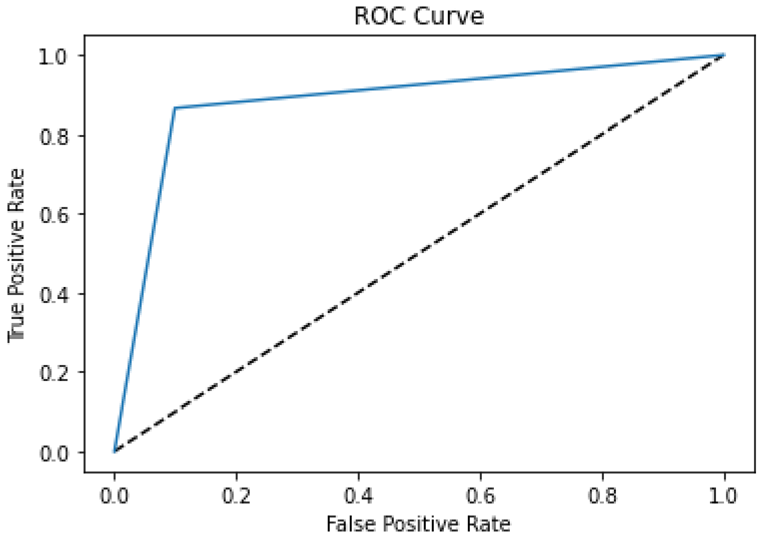


```
              precision    recall  f1-score   support

           0       0.87      0.90      0.88       494
           1       0.90      0.87      0.88       506

    accuracy                           0.88      1000
   macro avg       0.88      0.88      0.88      1000
weighted avg       0.88      0.88      0.88      1000
```

Modelling – RFC Performance

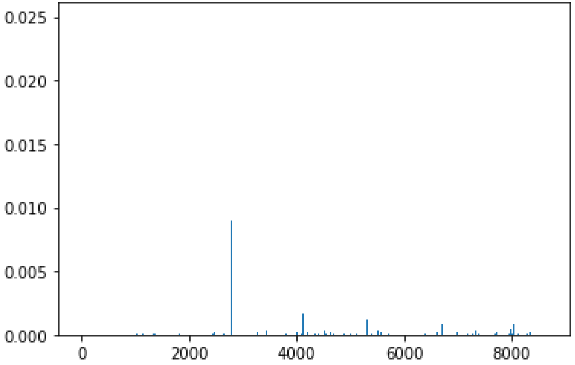
Random forest Classifier
Algorithm with high interpretability and accuracy

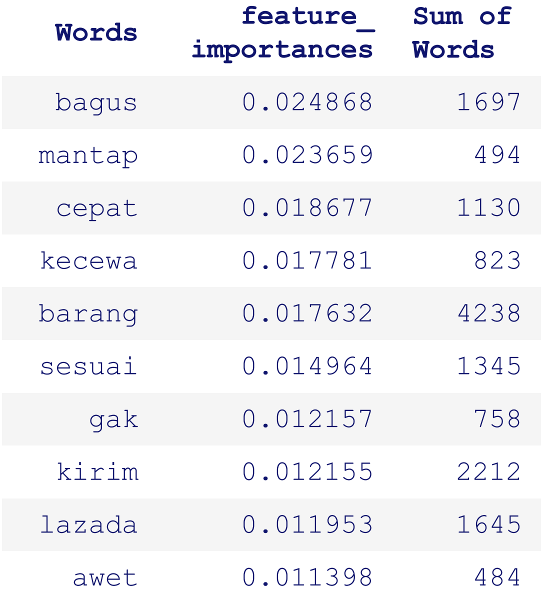
Accuracy	: 0.883
AUC	: 0.896



Modelling – Feature Importances

RFC Feature Importances





Words	feature_ importances	Sum of Words
bagus	0.024868	1697
mantap	0.023659	494
cepat	0.018677	1130
kecewa	0.017781	823
barang	0.017632	4238
sesuai	0.014964	1345
gak	0.012157	758
kirim	0.012155	2212
lazada	0.011953	1645
awet	0.011398	484

Conclusion

Sentiment Analysis model has been created with RFC model with the biggest AUC (0,895)

There are 3 ways that you can implement to increase your ratings and reviews:

REMIND

Each customer only has 30 days to leave a review. **Remind** them via "Chat" or send an "invitation card" in your parcel.

REWARD

Reward customer with vouchers for leaving a review.

RESPOND

Listen and understand your customers. **Respond** to both good and bad reviews.

Literature

- J. Wong, Natural Language Processing Workflow , 2020, (https://towardsdatascience.com/natural-language-processing-workflow-1dddf3a48ab5).
- B. Shetty, Natural Language Processing (NLP) for Machine Learning, 2018, (https://towardsdatascience.com/natural-language-processing-nlp-for-machine-learning-d44498845d5b).
- A. F. Zulfikar, Pengembangan Algoritma Stemming Bahasa Indonesia dengan Pendekatan Dictionary Base Stemming untuk Menentukan Kata Dasar dari Kata Yang Berimbuhan, Universitas Pamulang, 2017.
- https://sellercenter.lazada.com.ph/seller/helpcenter/ratings-reviews-11737.html?spm=a2a15.helpcenter-psc-article.articles-list.4.41a71da3t3hxJY
- https://sellercenter.lazada.co.id/seller/helpcenter/Apa-Yang-Dimaksud-Dengan-Positif-Seller-Rating-6070.html?spm=a2a14.helpcenter-psc-search.article.1.37cb475dS3IBs5
- https://sellercenter.lazada.co.id/seller/helpcenter/Apa-Itu-Bisnis-Analisis--6041.html?spm=a2a14.helpcenter-psc-article.articles-list.7.5b0f4af5qYlIv4

THANKS

CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon**, and infographics & images by **Freepik**

Please keep this slide for attribution