Shopee Product Matching

Wee Zi Jian

Agenda

1. Problem Statement
2. Exploratory Data Analysis
3. Image Embedding
4. Text Embedding
5. Combined Predictions
6. Conclusions and Recommendations

1. Problem Statement

Goal

- To provide competitive prices to customers, retail companies like Shopee use product matching to **identify identical products** on their platform
- Given a list of product images and titles, **predict which products are the same**

Prediction Format

If [A, B, C, D, E] is the set of all unique products and we predict [A, B, C] to be the same and [D, E] to be the same:

Product	Prediction
A	A, B, C
B	B, A, C
C	C, A, B
D	D, E
E	E, D

Scoring Metric

Predictions will be scored row-wise then averaged using F1 score / Dice Metric:

$$\frac{2 * |X \cap Y|}{|X| + |Y|}$$

where X is the set of our predictions and Y is the set of actual matches

Scoring Metric

If [A, B] and [C, D, E] are the actual matching sets:

Product	Prediction	Actual Matches	F1 Score
A	A, B, C	A, B	0.8
B	B, A, C	B, A	0.8
C	C, A, B	C, D, E	0.33
D	D, E	D, C, E	0.8
E	E, D	E, C, D	0.8
		Average Score	**0.71**

Approach

- Generate image feature embeddings
- Generate text feature embeddings
- Search embedding space for nearest neighbours
- Combine image and text predictions

Challenges

- Precision vs recall
- Noisy data
- Combining image and text predictions

2. Exploratory Data Analysis

Train

34,250

rows

5

columns:

- posting_id
- image
- image_phash
- title
- label_group

Test

70,000+

rows

4

columns:

- posting_id
- image
- image_phash
- title

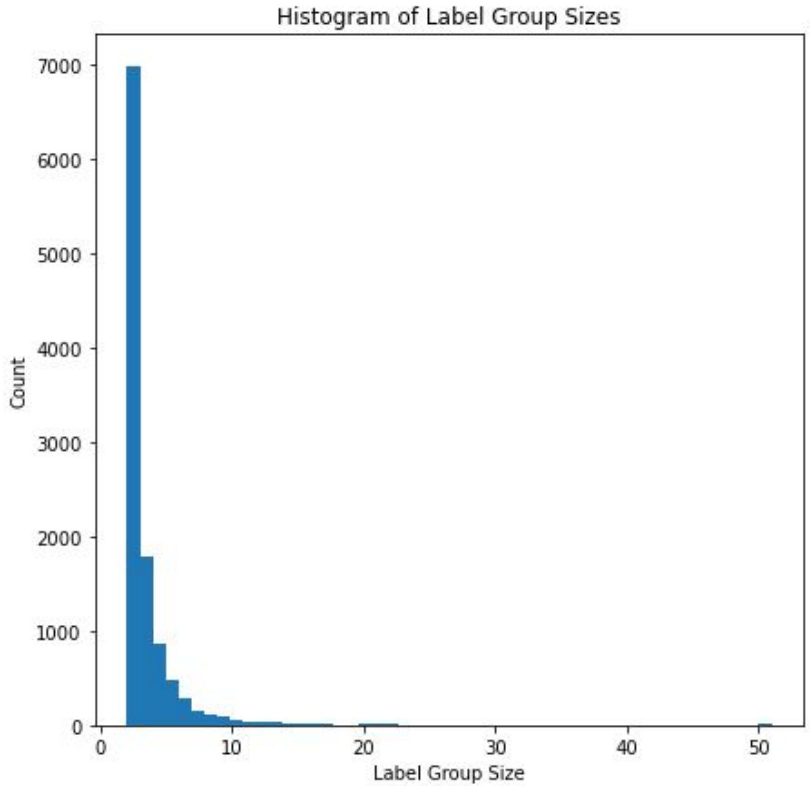
Data Dictionary

posting_id	Unique identification number of a product
image	Image file name
image_phash	Image perceptual hash
title	Product title
label_group	Identification number for actual matches **Products are the same if they have the same label group**

	posting_id	image	image_phash	title	label_group
0	train_129225211	0000a68812bc7e98c42888dfb1c07da0.jpg	94974f937d4c2433	Paper Bag Victoria Secret	249114794
1	train_3386243561	00039780dfc94d01db8676fe789ecd05.jpg	af3f9460c2838f0f	Double Tape 3M VHB 12 mm x 4,5 m ORIGINAL / DO...	2937985045
2	train_2288590299	000a190fdd715a2a36faed16e2c65df7.jpg	b94cb00ed3e50f78	Maling TTS Canned Pork Luncheon Meat 397 gr	2395904891
3	train_2406599165	00117e4fc239b1b641ff08340b429633.jpg	8514fc58eafea283	Daster Batik Lengan pendek - Motif Acak / Camp...	4093212188
4	train_3369186413	00136d1cf4edede0203f32f05f660588.jpg	a6f319f924ad708c	Nescafe \xc3\x89clair Latte 220ml	3648931069

Label Groups

- 11,014 label groups in train data
- Sizes range from 2 to 51
- **63% size 2**, 16% size 3, 8% size 4
- Test data:
  - Sizes range from 2 to 51
  - Unknown number/distribution of label groups



Images

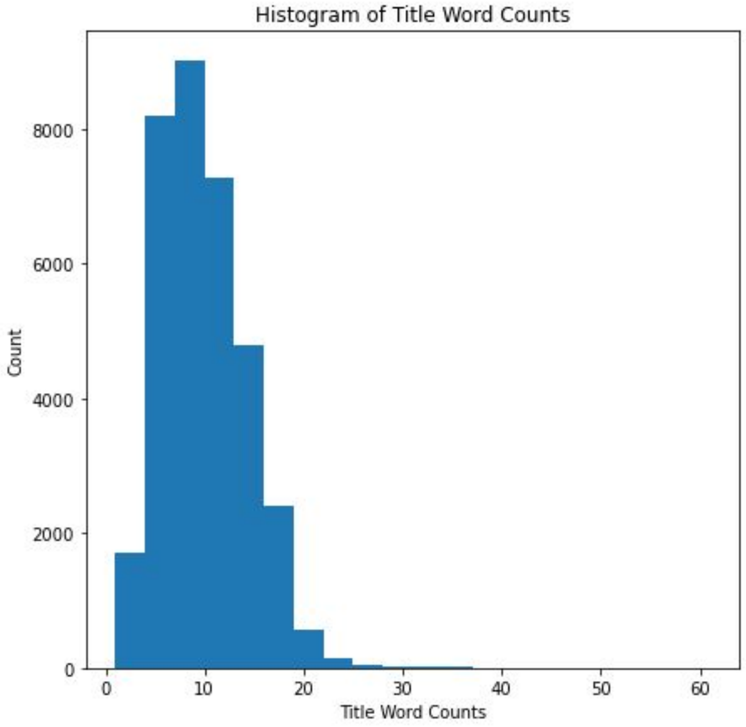
Images differ in:

- Image size
- Image phash
- Overlaying of words/branding
- Positioning/size of the product within the image
- Background colour

Sample images from same label group

Titles

- Titles comprise about 10 words on average
- Top tokens include English/Indonesian words and non-alphanumeric characters



Top title tokens

Token	Count
/	8616
-	5221
anak	1916
wanita	1820
‘‘	1706
original	1681
1	1400
murah	1363
tas	1192
dan	1157

Titles

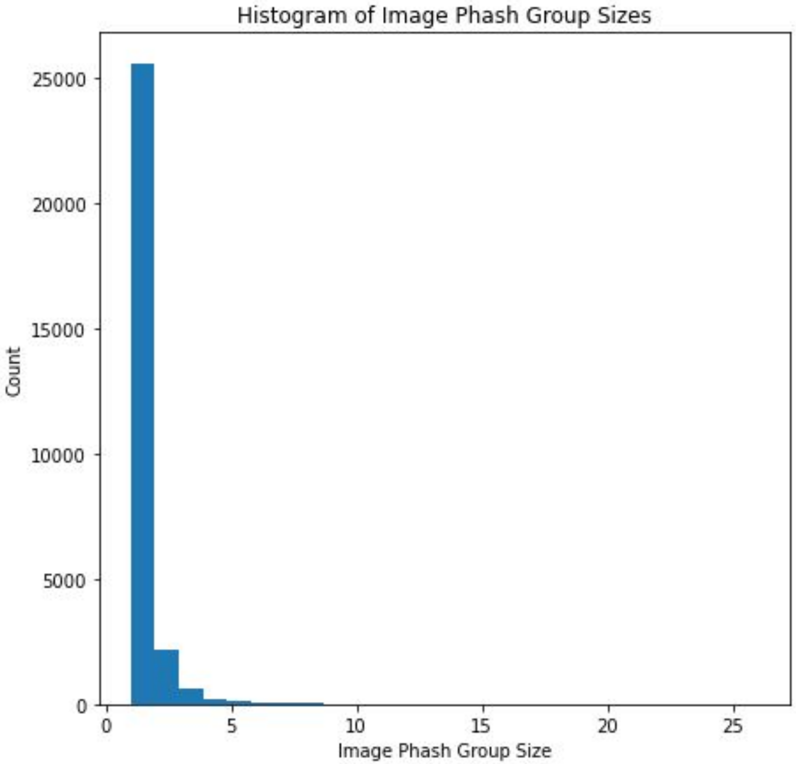
- Mix of English and Indonesian
- Common stop words e.g. 'original', ‘new’, ‘ready’
- Same keywords in same label groups
- Non-alphanumeric, UTF-8 characters e.g. ‘/’, ‘-’, ‘\\xc3’

Sample titles from same label groups

```
SCARLETT SERUM - BEBAS PILIH
SCARLETT WHITENING ACNE SERUM
SERUM SCARLETT ACNE & BRIGHTLY- SERUM WAJAH SCARLETT ORIGINAL BPOM
Scarlett Whitening Brightly Ever After Serum
SCARLETT serum wajah NEW
SERUM SCARLETT ACNE & BRIGHTLY- FACE SERUM ORIGINAL BPOM
SCARLETT Whitening Brightly Ever After Serum / Whitening Acne Serum
SCARLETT WHITENING BRIGHTLY EVER AFTER SERUM
SCARLETT ACNE SERUM & BRIGHTLY EVER AFTER SERUM
Scarlett Serum
SERUM SCARLETT 15ML BRIGHTLY / ACNE PILIH SALAH SATU
SCARLETT ACNE SERUM / BRIGHTLY EVER AFTER SERUM
SCARLETT WHITENING SERUM
[READY] Scarlett Whitening Serum Acne/ Brightly Ever After by felicia
Scarlett Whitening Serum Brightly Ever After Serum / Acne Serum

['Paper Bag Victoria Secret']
['PAPER BAG VICTORIA SECRET']

['Double Tape 3M VHB 12 mm x 4,5 m ORIGINAL / DOUBLE FOAM TAPE']
['Double Tape VHB 3M ORIGINAL 12mm x 4.5mm Busa Perekat']

['Maling TTS Canned Pork Luncheon Meat 397 gr']
['Maling Ham Pork Luncheon Meat TTS 397gr']

['Daster Batik Lengan pendek - Motif Acak / Campur - Leher Kancing
['DASTER PIYAMA KATUN JEPANG(TIDAK BISA PILIH MOTIF & WARNA)']

['Nescafe \\xc3\\x89clair Latte 220ml']
['Nescafe Eclair Latte Pet 220 Ml']
```

Image Perceptual Hash

- 28,735 unique image phash in train data
- Size ranges from 1 to 26
- **89% size 1**, 8% size 2, 2% size 3
- Image phash is mostly unique
- **May not be useful** for predicting matching products

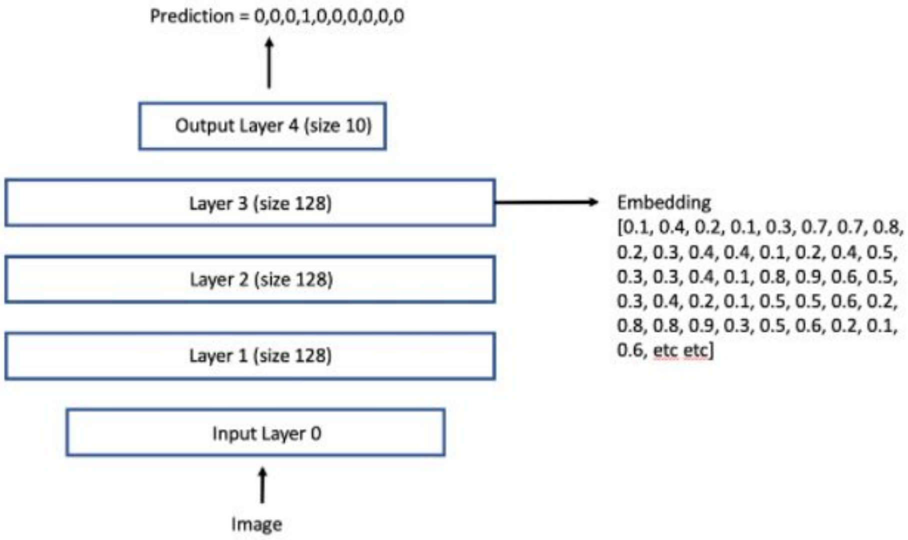


3. Image Embedding

Feature Embedding with EfficientNet

- EfficientNetB4: convolutional neural network pre-trained on ImageNet
- Generate embeddings from 2nd last layer for each image:

34,250 images x 1,792 features

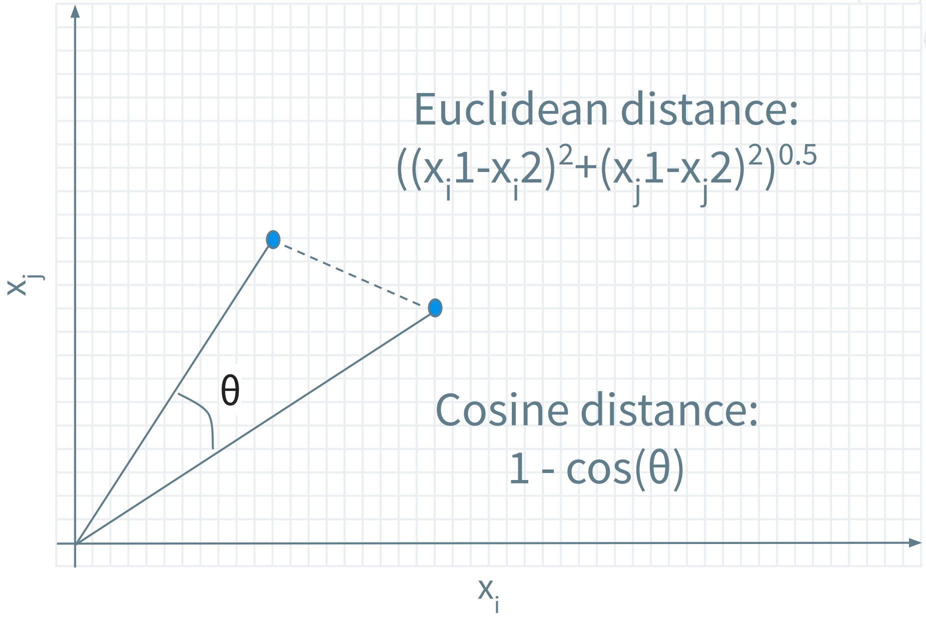


Source: https://www.kaggle.com/c/shopee-product-matching/discussion/226279

sklearn / cuML Nearest Neighbours

- sklearn.NearestNeighbors (CPU)
- cuML.NearestNeighbors (GPU)
- From (34250, 1792) feature matrix, calculate **nearest 51 neighbours** of each point using either Euclidean or cosine distance
- Returns **distances** and **indices** of nearest neighbours of each point: (34250, 51) matrices

Distance calculation of 2 points in 2 dimensions



Making Predictions

- From **distances** and **indices** of nearest 50 neighbours, set **distance threshold** to obtain predictions
- For example below, all indices below 0.2 distance will be assigned as predictions

Product	Distances	Indices	Threshold	Prediction
A	[**0, 0, 0.1**, 0.4, 0.5]	[**A, B, C**, D, E]		A, B, C
B	[**0, 0, 0.15**, 0.3, 0.6]	[**B, A, C**, D, E]		B, A, C
C	[**0, 0.1, 0.15**, 0.5, 0.6]	[**C, A, B**, D, E]	0.2	C, A, B
D	[**0, 0.1**, 0.3,0.4, 0.5]	[**D, E**, B, A, C]		D, E
E	[**0, 0.1**, 0.5, 0.6,0.6]	[**E, D**, A, B, C]		E, D

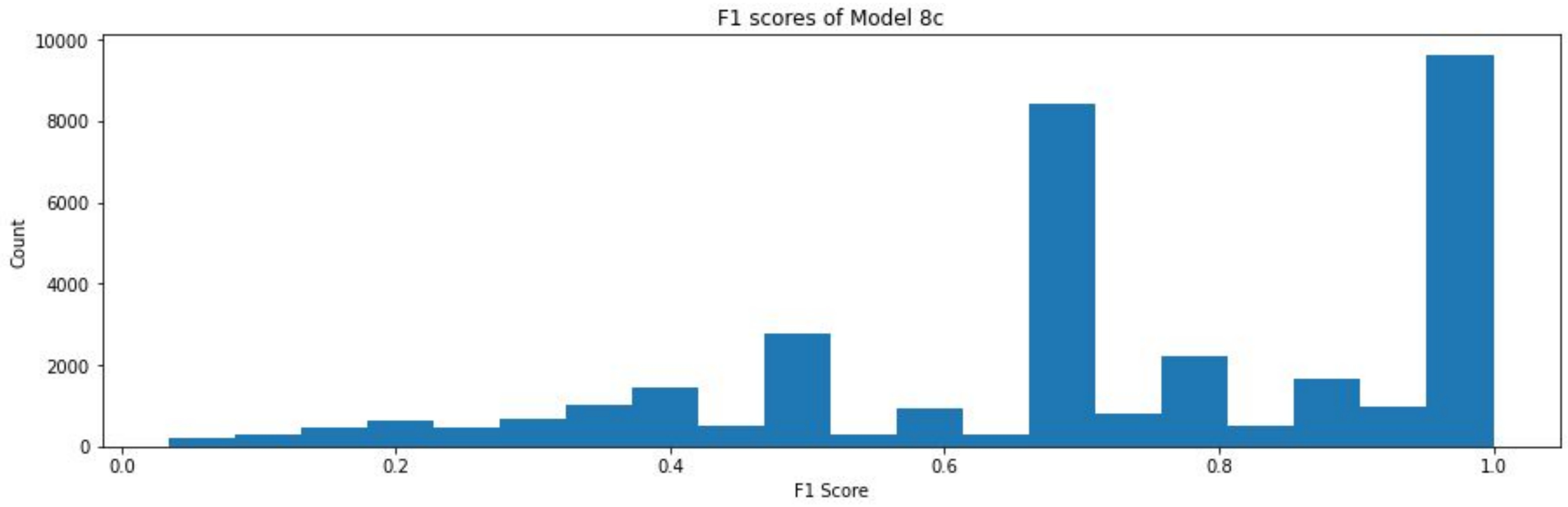
Results

- Default and re-trained EfficientNetB4 were used to generate image embeddings
- From feature matrix, obtain nearest neighbours and **grid search** various **distance thresholds** to find best score

No	Model	Euclidean Threshold	F1 Score	Cosine Threshold	F1 Score
0	Default EfficientNetB4	5	0.634	0.167	0.652
1	B4 with top_conv retrained (batch:32, epochs:3)	11	0.656	0.167	0.673
2	B4 with block7b_project_conv, top_conv retrained (batch:32, epochs:3)	13	0.654	0.2	0.685
3	B4 with block7b, top_conv retrained (batch:32, epochs:3)	11	0.665	0.2	0.681
4	Entire B4 retrained (batch:8, epochs:10)	8	0.637	0.167	0.634
5	Entire B4 retrained (batch:8, epochs:5)	8	0.632	0.133	0.629
6	B4 with block7a, block7b, top_conv retrained (batch:32, epochs:3)	18	0.684	0.2	0.701
7	B4 with block7b, top_conv retrained (batch:32, epochs:6)	11	0.662	0.233	0.679
8	B4 with block7a, block7b, top_conv retrained (batch:32, epochs:6)	20	0.675	**0.233**	**0.710**

Results

Distribution of scores for best image model:



Results

Sample predictions of best image model:

Product

Actual matches

Score:
$(2 \times 2) / (2 + 6) = 0.5$

4. Text Embedding

TF-IDF Vectorizer

- Remove English + Indonesian stop words from titles
- Lower case all words
- Regex tokenizer: [a-zA-Z0-9]+
- Fit transform using sklearn TfidfVectorizer
- **34,250 x 25,023 sparse feature matrix**

Google LaBSE

- Language-agnostic BERT Sentence Embedding
- Semantic embedding of multilingual sentence inputs
- BERT tokenizer:

```
['paper', 'bag', 'victoria', 'secret']
['double', 'tape', '3m', 'v', '##hb', '12', '4', '5', 'double', 'foam', 'tape']
['maling', 'tts', 'canne', '##d', 'pork', 'lunch', '##eon', 'meat', '397']
```

- **34,250 x 768 feature matrix**

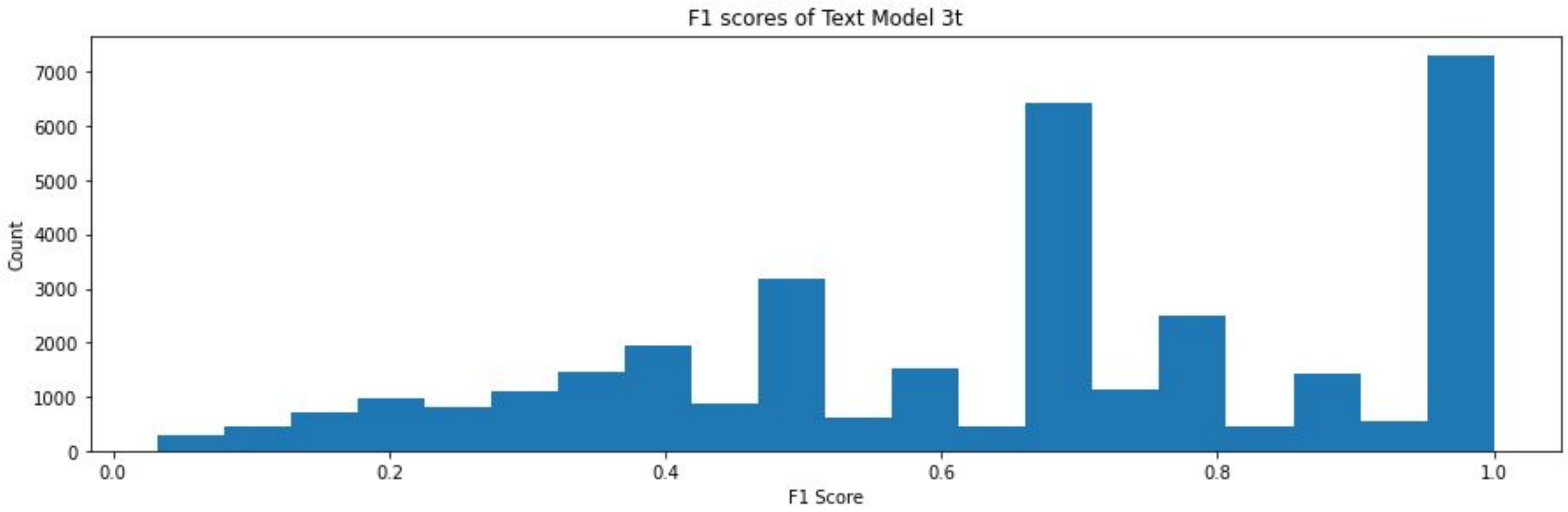
Modelling / Results

- Apply NearestNeighbors algorithm on feature matrix
- Grid search various distance thresholds

No	Tokens	Embedding Model	Distance Threshold	F1 Score
1	All stop words, [a-zA-Z0-9]+ regex tokenizer	TF-IDF vectorizer	Cosine 0.433	0.6467
2	All stop words, [a-zA-Z0-9]+ regex tokenizer	LaBSE	Cosine 0.2	0.6229
3	No stop words, [a-zA-Z0-9]+ regex tokenizer	TF-IDF vectorizer	**Cosine 0.433**	**0.6489**
4	No stop words, [a-zA-Z0-9]+ regex tokenizer	LaBSE	Cosine 0.2	0.6210

Results

Distribution of scores for best text model:



Results

Sample predictions of best text model:

```
Showing predictions of model 3t of posting_id train_3386243561
Titles:
Double Tape 3M VHB 12 mm x 4,5 m ORIGINAL / DOUBLE FOAM TAPE
Double Tape VHB 3M ORIGINAL 12mm x 4.5mm Busa Perekat
DOUBLE TAPE BUSA 3M Pe Foam Tape 24mm x 4M ORIGINAL
DOUBLE TAPE BUSA 3M Pe Foam Tape 24mm x 4M ORIGINAL
DOUBLE TAPE BUSA 3M Pe Foam Tape 24mm x 4M ORIGINAL
Double Tip / Double Sided Tape Perekat 2 Sisi Joyko 6 mm x 15 yard Double Tape

Tokens:
double tape 3m vhb 12 mm x 4 5 m original double foam tape
double tape vhb 3m original 12mm x 4 5mm busa perekat
double tape busa 3m pe foam tape 24mm x 4m original
double tape busa 3m pe foam tape 24mm x 4m original
double tape busa 3m pe foam tape 24mm x 4m original
double tip double sided tape perekat 2 sisi joyko 6 mm x 15 yard double tape
```

Product

Actual matches

Score:
(2 x 2) / (2 + 6) = 0.5

Results

Sample predictions of best text model (images):

Product

Actual matches

Score:
$(2 \times 2) / (2 + 6) = 0.5$

5. Combined Predictions

Set Union of Image and Text Predictions

Image prediction:

Score: 0.667

Text prediction:

Paper Bag Victoria Secret

PAPER BAG VICTORIA SECRET

Score: 1.0

Set union:

Score: 1.0

Combined Embeddings

- Concatenate feature matrices
- Apply Standard Scaler
- Search nearest neighbours from **combined matrix**

Image Embedding Matrix	TF-IDF Embedding Matrix	LaBSE Embedding Matrix	Combined Embedding Matrix
34,250 x 1,792	34,250 x 25,023	-	34,250 x 26,815
34,250 x 1,792	-	34,250 x 768	**34,250 x 2,560**
34,250 x 1,792	34,250 x 25,023	34,250 x 768	34,250 x 27,583

Soft Distance Threshold

Instead of fixed distance threshold for all predictions, use a **ratio of average distance** of nearest neighbours (e.g. 0.5)

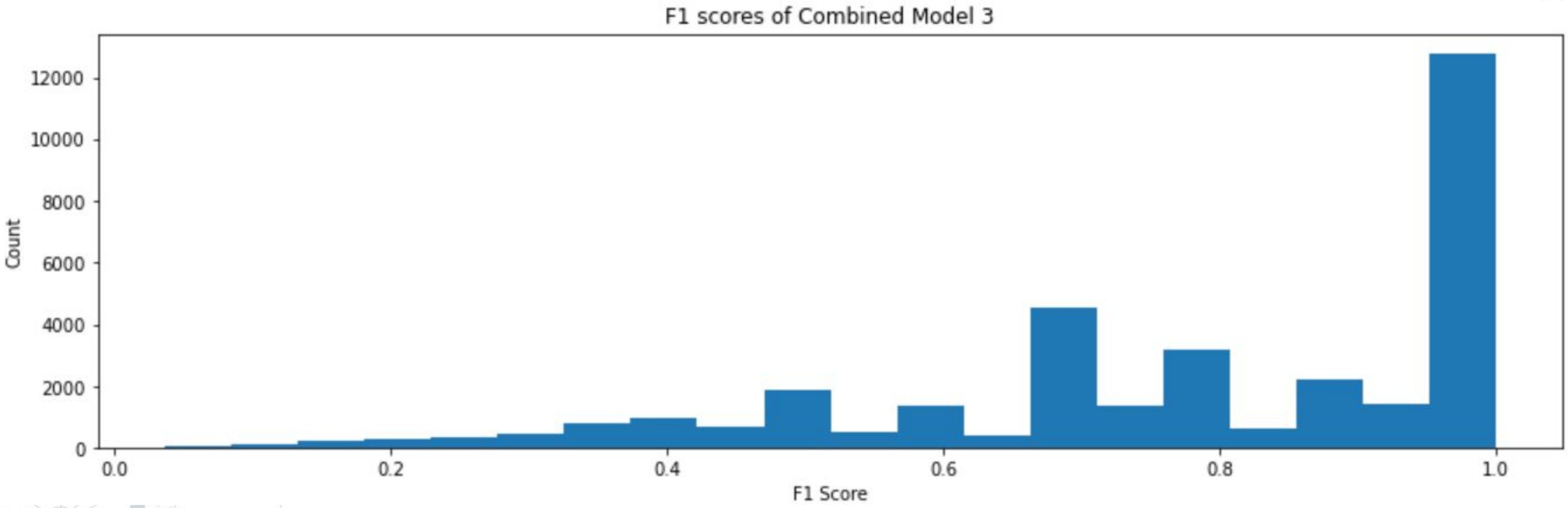
Product	Distances	Indices	Fixed Threshold	Average Distance	**Soft Threshold** (0.5 x average distance)	New Predictions	Score
A	[**0, 0**, 0.1, 0.4, 0.5]	[**A, B**, C, D, E]	0.2	0.2	0.1	A, B	1.0
B	[**0, 0**, 0.15, 0.3, 0.6]	[**B, A**, C, D, E]		0.21	0.105	B, A	1.0
C	[**0, 0.1**, 0.15, 0.5, 0.6]	[**C, A**, B, D, E]		0.27	0.135	C, A	0.4
D	[**0, 0.1**, 0.3,0.4, 0.5]	[**D, E**, B, A, C]		0.26	0.13	D, E	1.0
E	[**0, 0.1**, 0.5, 0.6,0.6]	[**E, D**, A, B, C]		0.36	0.18	E, D	1.0
						Average Score	**0.88**

Final Scores

No	Combination	Soft Distance Threshold (Ratio of average distances)	Train F1 Score	Kaggle Test F1 Score
1	Set union of best image + best TF-IDF models	Image: 0.62 TF-IDF: 0.56	0.773	0.724
2	Combined embeddings of best image + best LaBSE models	0.72	0.766	0.707
3	Set union of best image + best TF-IDF + model 2	Image: 0.5 TF-IDF: 0.45 Model 2: 0.7	**0.783**	**0.728**
4	Set union of best image + best TF-IDF + best LaBSE models	Image: 0.6 TF-IDF: 0.5 LaBSE: 0.4	0.775	0.726

Final Scores

Distribution of scores for best combined model:



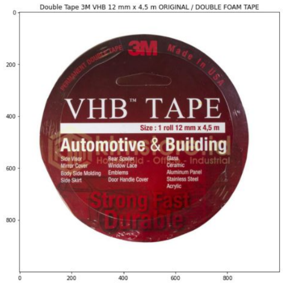
Sample Predictions

Matches:

Predictions:

Sample Predictions

Matches:



Predictions:

6. Conclusions and Recommendations

Model Analysis

- Identical or highly similar images/titles can be identified by the models
- Challenge lies in **combining image and text predictions** and setting the **distance heuristic** to make the right number of predictions
- Setting a **soft distance threshold** reduces excessive predictions and improves precision
- Using **ensembles** of different image and text embeddings improves recall

Challenges

Models perform poorly where product matches have **dissimilar images and titles**

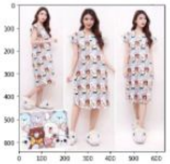
Matches:

```
Showing matches of posting_id train_2406599165
Titles:
Daster Batik Lengan pendek - Motif Acak / Campur - Leher Kancing (DPT001-00) Batik karakter Alhadi
DASTER PIYAMA KATUN JEPANG(TIDAK BISA PILIH MOTIF & WARNA)
```

Text predictions:

```
Showing predictions of model 3t of posting_id train_2406599165
Titles:
Daster Batik Lengan pendek - Motif Acak / Campur - Leher Kancing (DPT001-00) Batik karakter Alhadi
Daster Batik Lengan pendek - Motif Acak / Campur - Leher Kancing (DPT001-00) Batik karakter Alhadi
Daster Batik Lengan pendek - Motif Acak / Campur - Leher Kancing (DPT001-00) Batik karakter busui
Daster Batik Lengan pendek - Motif Acak / Campur - Leher Kancing (DPT001-00) Batik karakter IKHLAS
Daster Batik Bali Lengan pendek - Motif Acak / Campur - Leher Kancing BUSUI - BUMIL - Batik Alhadi
Daster payung klok motif acak/campur leher kancing busui bumil (DPT001-00) Batik FA
Daster Payung Bali JUMBO XXL, Motif Acak / Campur, Leher Kancing Bumil Busui (DPT005), Batik Alhadi
```

Image predictions:



Recommendations

- More data/metadata in addition to images and titles would be useful to generate a **more complete feature space**
  - Do same products come from the same shop, same supplier, same location etc?
- Product matching could also involve **other approaches** in addition to machine learning, e.g. user behavior
  - If users that view product A always view product B, it could be that A and B are the same

Thank You

Any questions?

Slides credit: Slides Carnival