Shopee Product Matching

Business Problem:

Shopee is a leading online e-commerce platform which enables users to buy and sell products online. It focuses on e-commerce and operates its business mainly in South Asian countries. Customers appreciate its easy, secure, and fast online shopping experience tailored to their region. The company also provides strong payment and logistical support along with a 'Lowest Price Guaranteed' feature on thousands of Shopee's listed products. In this competition they open sourced images of products with descriptions and expect Machine Learning practitioners to build models that identify similar products based on images and descriptions.

Data Exploration:

train.csv

	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

Posting_id: unique id for each posting

Image: file name of the image

Image_phash: perceptual hash of the image

Title: Description of that image

Label_group: group code for which product belongs to.

```
Unique posting_id: 34250
Unique images: 32412
Unique image_phash: 28735
Unique title: 33117
Unique label_group: 11014
```

There are 34250 postings in the train and some postings have similar image files, phash and titles. There are 11014 label groups. Products belonging to a label_group are similar. For an example,

	posting_id	image	image_phash	title	label_group
5	train_2464356923	0013e7355ffc5ff8fb1ccad3e42d92fe.jpg	bbd097a7870f4a50	CELANA WANITA (BB 45-84 KG)Harem wanita (bisa...	2660605217
1442	train_2753295474	0b5b2b2b3f84721140a7e18c2e43a4ff.jpg	bbd097a7870f4a50	grosir_solo ||PROMO!!! CELANA BAGGY PANT XL_BO...	2660605217
29305	train_305884580	db8df2aded514a77bc47eadf9e0e9acb.jpg	ac079338b33e1d2d	Baggy Pants JOGER AMERICAN DRILL/ CARGO PANTS ...	2660605217

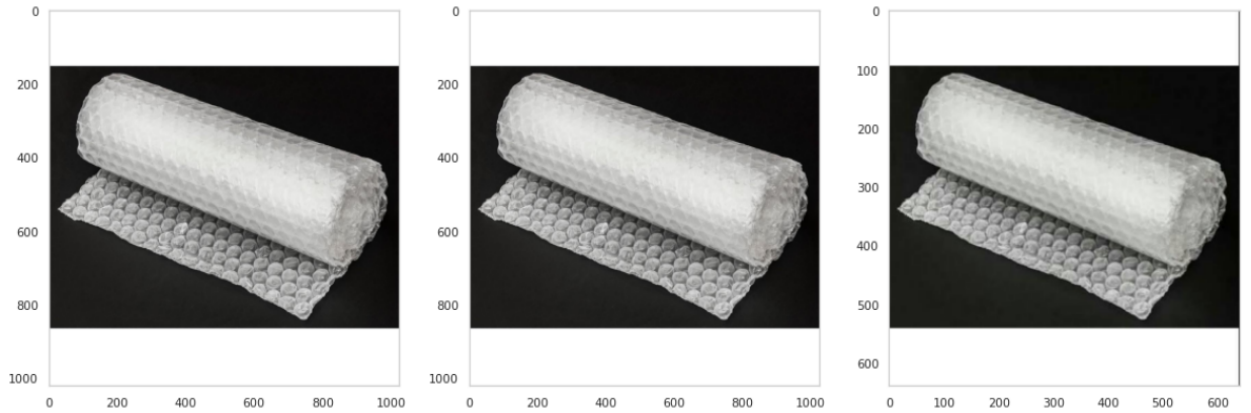
These three different products which belong to a label group are similar and it's evident by images.

Can label_group be considered as ground reality or ground truth? **No.**

Let's consider Image_phash. We can observe that below data points have the same image phash but they belong to different label groups.

	posting_id	image	image_phash	title	label_group
1651	train_3068759534	0cca4afba97e106abd0843ce72881ca4.jpg	d0c0ea37bd9acce0	BUBBLE PACK UNTUK PACKING TAMBAHAN 1BUBBLE UNT...	4198148727
1652	train_1049463374	0cca4afba97e106abd0843ce72881ca4.jpg	d0c0ea37bd9acce0	BUBBLE WARP	2403374241
5641	train_3095376889	2a8c24726ee9a1446a65325d65f66659.jpg	d0c0ea37bd9acce0	Packing Tambahan Bubble Wrap/Kardus Bekas	1960893869

When I checked the images they are similar products.



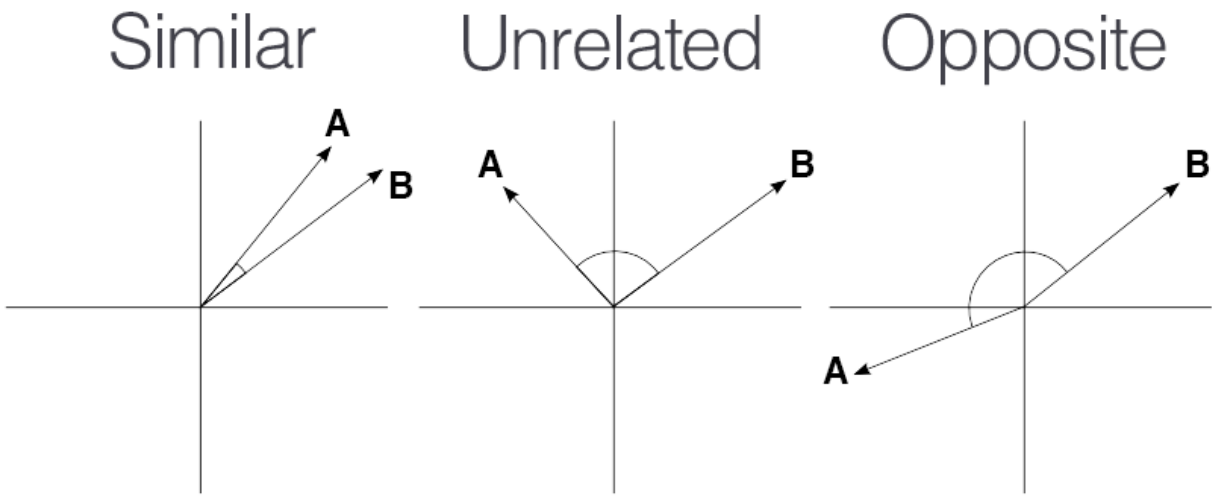
We can clearly observe that the above 3 products are similar and they belong to different label groups. So can image phash be considered as ground truth? **No.**

https://www.kaggle.com/c/shopee-product-matching/discussion/225257

From the above discussion on kaggle it's evident that dis-similar products can have the same image phash. Also, there are postings which have same/similar images or descriptions but belong to different label groups or have different image phash.

ML problem formulation:

This problem doesn't have a ground truth and it should be solved in an **Unsupervised** approach. Using description we can vectorize them using basic approaches like TFIDF, Word2Vec/Glove and can also use BERT and different variations of BERT(DistilBERT, SentenceBERT etc.,) to get the embeddings. After getting the embeddings we can make vectors as Unit Vectors and compute the DOT product(**Cosine Similarity**) with other vectors/embeddings. Making vectors/embeddings to unit length makes sure that DOT product(Cosine Similarity) lies between +1 and -1 as it only depends on angle between the vectors/embeddings now. If Cosine Similarity is +1 then the angle of separation between vectors is 0 degrees(Cos(0)=+1) which inturn mean they are identical vectors and if Cosine Similarity is -1 then the angle of separation between vectors is 180 degrees(Cos(180)=-1) which inturn mean they are most dis-similar vectors.



Similarly, Image embeddings can also be taken using **Convolutional Neural Networks** and the process of computing similarity is same as above. We can also use **Euclidean Distance** as a similarity measure but as I am already making vectors/embeddings as Unit Vectors the Euclidean Distance is proportional to Cosine Distance. We need to set a decision threshold on

Cosine Similarity/DOT product to get the matches for a product and that optimum threshold can be figured out while working on the problem.

Performance Metric:

The performance metric of this problem is F1-score averaged over postings. Which means the F1-score is computed on each product using its matches and the average of all products is taken.

First Cut Approach:

1. I will start with using only text(Product Description) for finding matches. Vectorizers like TFIDF, Word2Vec and BERT can be used and different variations of BERT can also be used.
2. Similarly I will use only images to find matches and CNN can be used for embeddings. Various architectures like ResNet, AlexNet, EfficientNet etc., can be used for embeddings. Keras have the build models which have their weights trained on ImageNet dataset.
3. ArcFace loss, which is a modification of SoftMax can be used to train the models. Training with ArcFace will update weights in such a way that embeddings belonging to a class label are more Cosine Similar. To do this we can assume label groups as class labels and this becomes a multi-class approach as we have 11014 different label groups.
4. Finally we can combine the matches by best models/embedders of both text based and image based and check the performance.