National College of Ireland

E-commerce Product Similarity Match Detection using Product Text and Images

MSc Research Project
Data Analytics

Hari Krishnan Kannan
Student ID: x19225547

School of Computing
National College of Ireland

Supervisor: Jorge Basilio

National College of Ireland
Project Submission Sheet
School of Computing

National College of Ireland

Student Name:	Hari Krishnan Kannan
Student ID:	x19225547
Programme:	Data Analytics
Year:	2021
Module:	MSc Research Project
Supervisor:	Jorge Basilio
Submission Due Date:	16/08/2021
Project Title:	E-commerce Product Similarity Match Detection using Product Text and Images
Word Count:	
Page Count:	23

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	23rd September 2021

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Contents

E-commerce Product Similarity Match Detection using Product Text and Images

Hari Krishnan Kannan
x19225547

Abstract

The online merchants and users supply rich data everyday on e-commerce websites, this only gets bigger in this era of data growth allowing more scope for research advancements in the area of product similarity match detection. The text descriptions of two similar products in the e-commerce websites may be slightly different, but their pictures could be largely varied. The use of either only text-based comparisons or image comparison methods had been the trend for all e-commerce platforms in the market but the technology growth explore ways for combining both image and title descriptions together. This research project uses deep learning methods to detect the list of e-commerce identical products using product title and images. Residual Network called ResNet with deep layers and siamese twin is used as an approach for modelling the images, TF-IDF vectorizer is used for text modelling. Experimented through image augmentation, embedding and text processing to identify the identical products and a combination of TF-IDF + ResNet-18 is used as a base model for this research. Cross validation score is used to calculate the model accuracy and computational time is captured for model performance. The outcome of this research work will evaluate the implementation results and a comparison study is carried out using the real-world data from one of the leading e-commerce providers.

Keywords- e-commerce, ResNet, Siamese, TF-IDF, match detection

1 Introduction

I am one of the online shoppers who searches internet merchants for the best bargains on specific e-commerce products, are you too? The leading E-commerce companies around the world employ a range of strategic models for product matching to find similar products and promise customers that they will be offered with price match guarantee, special offers at market competitive prices, with improved user experiences and product recommendations on the go. A novel approach of product match detection using both product text and images leads the way of effort in this research paper. The background, motivation, research topic and question, literature justification for this research project, and the objectives and deliverables involved in project implementation and evaluation are all briefly discussed.

1.1 Background

Savvy online customers do not want to pay that extra when a better deal on the same product is up for grab. Various e-commerce giants from regions to regions worldwide utilize the E-commerce product similarity match. Some e-commerce platforms offer used items, and in business scenarios in which sellers always update custom information, this updated information may be incorrect from seller to seller even if human errors occur. For product classification, markets, and pattern descriptions, it is necessary to have a more precise product match. With this knowledge, e-commerce companies may take proactive action to deliver consumer commitments of great quality. The high-level characteristics for comparison are generally attributes containing picture, text, audio, and video format. In the description of product details, text and photos are the typical format used. Product comparisons are carried out with the title of the product, product characteristics like color, structure, type, and photos of a product. Leading global e-commerce companies devote substantial time and investment to make further innovate in these areas using accessible data on their platform.

1.2 Motivation

Increasing numbers of buyers only choose to buy new or used items if a better price is offered, given the downturn in economic owing to declared pandemic. It's not buyers time right now; you cannot buy engagement you have to build engagement. For e-commerce shoppers like one and all, the need to look out for a similar product that offers the best price and quality is the need of the hour. The amount of website hits for e-commerce giants like amazon, Shopee etc. is rising enormously due to available resources like simple Internet connection and user's ample time owing to lockdown. Referring to the examples of business scenarios, the number of people seeking a product online will quickly increase but converting the leads into sale is only when selling such items are at the best bargain for the same. I personally remember missing out some of the best deals on e-commerce sports goods which offered almost the same thing at a higher price later in time. Because the product photographs, product title and its descriptions provided by the sellers were different though they were similar items, I was only interested just so to understand the probable reasons that the product from another seller was not included in my search results. I missed a potential greatest bargain as a customer, to my dismay, which was still quite close to me. The necessity to look closer at the parallels of e-commerce items and algorithms and models used to detect the similarity motivated this research work. Any additional research Moraes et al. (2020) progress in this field covers the customer and business issue statement quickly. In comparison with the organizational advantages, customer focus benefits are more important. My own experience like yours is a proof of statement to the problems when the results of the product match do not show similar goods in vicinity.

1.3 What is Product similarity match in e-commerce?

Customers frequently compare items to similar comparable products while purchasing an item online or even if the purchase is not entirely online like at the outlet shops. Only when there are fewer direct comparison factors can comparisons between items be done easily, but in an era of huge data, comprehensive information is supplied for every product offered online. Many sellers on the retail and commercial line of business can

offer the same items online, therefore the rich information supplied by these sellers in terms of Fuchs et al. (2020) product title, Li et al. (2020) characteristics, and Rajest et al. (2021) pictures may change. A strategy for Product similarity matching is required by e-commerce platformers to comprehend the item similarities that are being offered on their own online platform or on any third-party platforms. E-commerce firms may utilize this measure to drive consumer behavior, enhance product discoverability, boost sales, create a recommendation framework, and improve shopping experience, among other things, by finding related items.

1.4 Research Question

How well can neural networks using deep layers and siamese twins along with TF-IDF vectorizer determine the similar products on shopee website [1] dataset containing identical e-commerce products with subtle varying title descriptions but broadly different images?

1.5 Academic Justification

While many ecommerce giants are currently using product similarity match detection algorithms, the space is still void for a more accurate and robust method which is why this is still a profound research topic for ecommerce companies and academics. Text or picture-based solutions are now accessible in the ecommerce industry; a mix of machine and deep learning data specific approaches evaluates similarity by analyzing text and image data independently. One of the most recent possibilities for use cases in this business Moraes et al. (2020), is integrating text and image together for product comparison; similarity may be successfully detected by combining these information parameters. It's been used in latest research studies to find identical items; two products with distinct product titles or pictures can still represent the same products if not different completely. On a wider scale, finding near identical items is a common difficulty for so called online companies; in this case, the issue is that merchants submit their own version of product pictures, titles, and product attributes, posing problems to online users. The peculiarity of our work is that instead of utilizing state-of-the-art approaches, we use the strategy of product titles and photos combined to discover comparable goods that have been frequently uploaded. In addition, the dataset comprises a wide range of e-commerce items with varying text descriptions, pictures of identical products that are vastly different, but only minor difference between related products. Allowing the investment in development of stronger consumer recommendation systems in the future, more advancement can be seen in predicting accurate or identical e-commerce product listings.

1.6 Objectives and Deliverables

The research major output is to present research artefacts on the progress made in the field of analysis to identify the identical e-commerce products in the listings. The detailed back ground, motivation, research scope and academic value is explained in section 1

The related work on the past literature study relevant to our research is discussed in section 2 and section3 briefly discuss the methodology and design where as the section 5 covers the details on the implementation and section 7 at the last concludes the research work. The list of research object is put below in bullet in points

[1]Dataset: `http://shopee.com`

- Number 1 - Identify and Declare the Research Scope
- Number 2 - Literature review of the relevant work on product similarity detection
- Number 3 - Relevant Data Selection for detecting similar products
- Number 4 - Pre-processing and feature extraction to classify the identical products
- Number 5 - Model Implementation methods for product similarity match detection
- Number 5.1 - TF-IDF Vectorizer Model Implementation and Results Evaluation
- Number 5.2 - ResNet-18 Model Implementation and Results Evaluation
- Number 5.3 - ResNet-50 Model Implementation and Results Evaluation
- Number 5.4 - Siamese ResNet-50 Model Implementation and Results Evaluation
- Number 5.5 - TF-IDF + ResNet-18 Model Implementation and Results Evaluation
- Number 6 - Model Comparison and Discussion

2 Related Work Review

In this study by Moraes et al. (2020),supplied a customer-specific e-commerce system with images, textual descriptions, and attributes needed for comparison and feedback, as well as customer-specific questions and answers When this case study is broadened to include a crowd-sourced example, it is revealed that customer-generated characteristics receive more weight than catalogue-generated features. This audience study also demonstrates that low correlation product quality attributes are ineffective for prediction, and that only high correlation qualities should be considered for machine learning applications. Specific product features and reviews may be evaluated for product similarity detection utilizing machine learning models with this demonstration of user provided attributes content significance. This research proposal is fueled by a strong desire to continue looking at relevant case studies and experiments in this direction, as well as domain applications. The important modification towards this is to use product description characteristics as input in discovering product match similarity in addition to the traditional technique of having product title and image.

2.1 Critical Review of Text Attributes Based Models

Detecting similarities between the online items in the listings on a platform like eBay is often a backend operation that uses the L2Q (Listing2Query) technique to determine the similarities based on the text semantics of the listed product names. In E-commerce any approach Fuchs et al. (2020) with Intent-Driven Similarity is a paper that employs a bidirectional RNN (recurrent neural network) to tokenize the significance weights in the Listing2Query method training. This approach shows a substantial improvement in similarity findings in a straightforward manner that even outperforms the most popular BERT method. This approach may be used to any type of search platform that involves text content, not only big online markets.

Information retrieval via item search on any e-commerce website is regarded more difficult than standard website and news portal search boxes. In a comparison research Sarvi et al. (2020) utilizing the match strategy idea for item search, it was discovered that, in terms of accurate and efficient model, ARC-I is the best model for real business applications over MV-LSTM and DRMMTKS techniques for short text matching. In addition, the performance of the text BERT technique is demonstrated to be poor, and the impact of search query characteristics such as complexity and length on selecting the best performing methods is examined.

In the online merchant sector, recommendation systems have remarkable success throughout realtime consumer interactions. The study was based on Bag of Words and Term Frequency-Inversed Document Frequency, which were obtained from Amazon product marketing API once data cleansing and text analytics has been conducted. Text-based item similarity by text vectorization approach is helpful for creating a data supportive product search function and user-like item recommendation, as the applications are heavily integrated with the textual description, which can be used with different e-commerce applications. Shrivastava and Sisodia (2019)

As more of the businesses put their operations online, the associated product details expands the range of volumes, introduces UPM, a clustering uncontrolled machine-learning method Akritidis et al. (2020) which matches comparable goods, regardless of external inputs. In the same way, headline comparison is avoided, but rather possible combinations of titles, words and scores are generated on the basis of the frequency of each length after three phases; in the first step, words are extracted out of the labels, and in the second level, they are the most suitable string of words. The post-pre-treatment check occurs at the third step, and the similarities are refined by the original clusters. UPM (Unsupervised Product Matcher) is the quickest way and has obtained a significantly greater concordance with exceptional instances.

To create a product similarity system Huang et al. (2019), and examine the function of the consumer influence prediction model in using the features of the sample file [2]. This study investigates the similarity by means of product evaluations and descriptions utilizing semantics and finds a detrimental impact on demand. The more resemblance between the text as well as the customer reviews are detected, the greater the need for the goods and suitably is made available. The major assumption out of this study is that the characteristics of product descriptions and product evaluations have concluded that the resemblance between the goods is essential. The fundamental request for identification of the field of discussion through using data from marketing material was demonstrated by this experimental investigation on actual world data.

There is a significant chance of offering the same items on several websites by the retailers with the increased number of online stores. The path models in internet is becoming a growing need for merchants and customers. Li et al. (2020), In this study the items on different on line marketplaces were compared using two criteria such as product description and product characteristics. This experiment has been done on two distinct various websites using data including hundreds of e-commerce goods. The product Title Match Model (TMM), and Attribute Match Model (AMM), is distributed over a multi-layer perception, a neural network match model is constructed. The approach proposed gives an output that considerably outperforms state-of-the-art matching models.

[2]Reference URL : `http://taobao.com`

2.2 Image Based Models for Identical Product Detection

Several e-commerce businesses suffer product Recommendation restrictions since they presently discover comparable goods with text-based methods. The suggestion for searching for similarities is developing the followers among e - commerce giants, in order to provide the client an improved viewing experience of search. K-means clustering is utilized in this published paper to identify clusters after the principal components analysis is applied to decrease dimensionality using a singular value decomposition approach. This results are utilized to compare five additional non-controlled machine learning algorithms such as K-medoid, Birch, Agglomerative, Minibatch, and Gaussian mixture model. In this research paper by Addagarla and Amalanathan (2020) the results of the study were utilized as a comparative comparison for five different uncontrolled approaches of machine learning, for example K-medoid, Birch, Agglomerate, Minibatch and Gaussian, made using 40 km of fashion models. The comparison analysis demonstrated an outstanding quality in grouping the identical models by the suggested K-means PCA-SVD approach method.

Trappey et al. (2020) Their research study involves the use of orthographic, visual and phonetic characteristics to establish trademark similarities. The classifier is constructed to check human language sentence structure similarity through orthodoxy, phonetic and picture similiarity analysis, using machine learning methodologies such as the Siamese neural network and the Convolutional neural network (CNN), with the help of a Machine Learning Technology, such as vector space algorithm. This solution is designed and evaluated with picture pairings of 270 thousand. It also offers digital content checking functionality to automatically detect potential infringements of images and text.

The possibility that all accessible e-commerce vendors will miss a fantastic bargain due to the price a seller sells for one particular product is very high. Here Zuo et al. (2020) two sorts of difficulties, almost same, colour, design, type, style and comparable items promoted by different vendors replacing things that are of common appeal to buyers. Second, recognizing similarities between millions and millions of goods that are accessible creates problems for the processing of large-scale data. The article on the Amazon data offered a customizable approach using PSS (Product similarity service) in which the deeper neural networking methods and distributed computing technologies were used to solve the obstacles. As a result, a very comparable product is produced with efficiency and scalability both in classification and in verification activities.

2.3 ResNet

Two separate data sets were employed throughout this study to determine the ResNet model's performance. The first is photos regarding healthcare information and the latter is a malicious and benign data. We conducted cancer prediction tests with the first dataset and on the second method to detect malware. ResNet, ResNet18, ResNet50, ResNet101, and ResNet152 models, which are Microsoft-owned, have been researched and evaluated. In order to approximate any incessant capability, the system of neural network models was demonstrated to be feasible in a few spaces spanning late deep neural systems (DNNs), from PC vision, speech recognition to word processing. The objective of this article is to forecast cancer illness using neural networking and to identify malware with the same ResNet model. On two separate datasets were tested for ResNet's performance. Khan et al. (2018)

Classification of cloths is a major research topic utilized by websites for e-commerce in

order to present good items to end consumers. Indian garments are widely categorized in both male and female's apparel. Traditional Asian clothing like 'Kurta' and 'Saree,' like shirts and denim, are worn significantly different from western attire. In addition, ethnic clothing styles and designs are extremely distinct from western costumes. Thus, ethnic clothes fail the models that are trained with regular ethnic data. The biggest ethnic data collection with over 0.1 million pictures in 15 distinct categories to classify ethnic clothing in fine grains using ResNet structure with 18,50 and 101 layers deep. Image Augmentation such as Jitter and Flip is used to improve the model prediction and performance. Model accuracy reaches 88.43 percent and to promote research into many algorithms, such as the categorization of garments and landmarks, particularly for ethnic outfits. Rajput and Aneja (2021)

2.4 Siamese Network

In the context of e-commerce businesses, recommendations systems play an essential part in helping customers discover what they want at the appropriate time as they optimize the user experience. This article focuses on the identification of the complimentary connection between the items using just the product names with 85 percent accuracy. The study propose to discover complementary products using Siamese Networks, a content-based approach of recommendation (SNN). We use the two versions Siamese CNN and Siamese LSTM to develop and evaluate. In addition, the SNN method is being extended to handle millions of items in only a couple of seconds and training time complexity is reduced by half. Angelovska et al. (2021)

Deep neural network is proposed for understanding the integration of pictures in this paper for the evolution of the notion of ocular affinity. We demonstrate the profound structure of Siamese that teaches how to integrate items appropriately into the categorization of visual similarities while training in beneficial and harmful combinations of photos. We often create a new loss measurement method depending on the requirements of the situation. The integrated descriptions of embedding at low and high levels was its ultimate picture embedding. The fraction distance matrix was also utilized to determine the distance between investigated embeddings with in n-dimensional space. Comparison of our design is done with a lot of different contemporary architectures and demonstrate that the image recovery is superior in this method typically showing how our suggested networks are strengthened by learning how to integrate them in an optimum way as compared to standard deep CNNs. Rajest et al. (2021)

2.5 Other Related Work

Considering our data collection is made up of text and pictures,by Sezavar et al. (2019), it is possible to identify similitudes between word and based on image matches. However, it is not important to look at the knowledge image recovery technique. The breadth of the search for close identical photos from the descriptions might be achievable with a description of the product and its relevant images. CNN and a scarce representation are utilized to extract the deep features of goods from the most recently published research article on knowledge picture recovery. Compared to state-of-the-art techniques, the suggested method employing the CNN led to good speed and exactness trials.

For embedding pictures, the deep CNN neural technique Rajest et al. (2021) is recommended. Siamese twin architecture is used to train and to illustrate how the images

closeness is right. To assess the design according to the description of issue, a system of loss computation is utilized. The spacing between images datatypes is calculated by a fraction of a distance. Current outcomes with Amazon's 16 thousand fashion goods have been beneficial for performance and visual improvement. Comparative research reveals that a deep convolutional network has importance over other traditional image embedded approaches.

Tsagkias et al. (2021), presented the newest scope for exploration prospects in the field of e-commerce. This article has recognized the problems and emphasized the areas in which the researchers may further immerse themselves in order to maximize customer experience in the field of e-commerce. This document leverages Amazon [3] and eBay[4] references to address consumer involvement and activity, identifies the resemblance of the product, and converts the results into sales. Improving the discovery of the products for consumers is essential to our research drive in a number of fields.

When listed on the web platform, e-commerce vendors usually pick a product category. Because of the diverse range of product categories and low degree of organized hierarchy refinement, it is quite likely that sellers would mistake. Hasson et al. (2021) This affects the items sold to consumers and the listings shown to the buyers in the case of the selling the product. Submitted suggestion group system to handle this problem in various e-commerce business scenarios. A convolution sequence-to-sequence technique is utilized to recognize the respective category and to evaluate its performance using new measures that demonstrate both efficiency and effectiveness in actual world data.

In the goods they are interested to acquire, customers typically search for attractive bargains. Niu et al. (2017) Many vendors might offer a product on the same e-commerce platform, which means the search results must include the listing of all items. They E-commerce search behavior predictive analytics for conversion. The client behavior in e-commerce was investigated in order to establish search parameters and buy predictions using Walmart data. Random forest is the method for predictive analysis and the outcomes of this prediction indicate a greater rates of accuracy. Different measurement methods and publications were suggested to evaluate consumer behavior in web search scenarios.

With critical review of all the related work in the past, the best feasible approach of the research project about it's implementation is detailed accordingly. The state-of-the-art algorithms used in this domain for product similarity match detection uses the product descriptions and images together are reviewed, the selected models expected to provide best performance. Advanced version of CNN algorithms with introduction of deep layers upto 50 and siamese architecture enhances the model, the model implementation is carried out by TF-IDF vectorizer + ResNet-18 to derive the results that is expected to outperform the existing methods.

3 Research Methodology

The details related to specific methodology used for this research project will be addressed in this below section briefly relating to the steps followed under the design specification and implementation sections. Through supporting papers and justified reasons about existing works in this research area, analysis is clearly carried out to explain why certain

[3]Business use case: `http://amzzon.com`

[4]Business use case: `http://ebay.com`

models, methods and process are chosen over others. Two main research methodologies namely KDD and CRISP-DM were considered to approach kick start this research work, In the event of relating to data mining research and achieving better results, this approach Fayyad et al. (1996) is utilized. The below following figure 1 shows the different stages involved in a graphical representation of the research methodology used

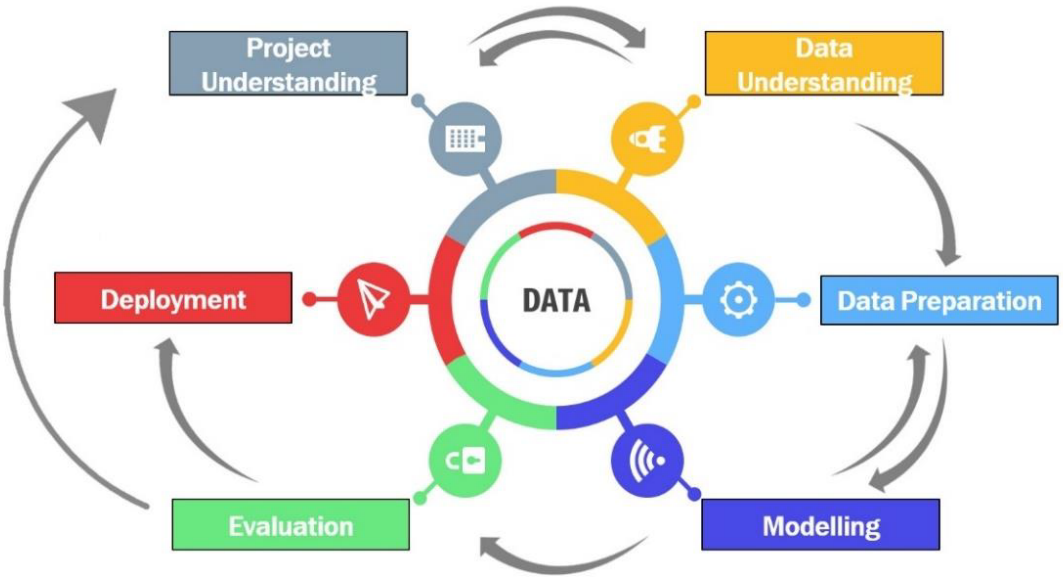

Figure 1: Research Methodology

3.1 Understanding the Project and Application domain

The first phase involved is to better understand the research area, topic, and research question under scope. In the consideration, e-commerce product similarity match detection has advantages for better guaranteed product price match, similar products from search listings and so on. The need to gather domain knowledge is important to detail the business use case where this research will have an impact. Various data models based on CNN is used as base to target the completion of this project.

3.2 Understanding the Data

The significant stage of the research work is to address the data requirements, the attribute of the dataset is thoroughly drill down in understanding the nature and behavior of the data. The data is searched around the available public websites and finally was sourced from Kaggle. From here on, this data was completely understood in terms of its size, attributes so that the next set of research steps such as preparing the data and choosing the required models can be addressed.

3.3 Data Pre-Processing and Transformation

To achieve effective results, the most important phase of the methodology is Data Preparation. Before applying any models, it is rather necessary to pre-process the data and techniques such as data embedding, image resizing, and normalization were used to ensure data consistency is met. Also, attempts were taken to visualize the data using certain libraries from python and transform the data as required.

3.4 Data Modelling

With the kind of data used, a various set of models are selected. The data is first split into train and test sets according to the dataset size and fed into the selected models for classification. The below following is the list of CNN based models that are implemented to predict the similarity match in e-commerce products as reference to the identified research question.

3.4.1 TF-IDF Vectorizer

TF-IDF is called as Term Frequency Inverse Document Frequency, the vector based model is a statistic model used to express the importance of retrieving information, aiming to extract the importance of a particular word in a document which itself is part of collection of documents called corpus.This is done by understanding the number of times a word that appears in the document, in our project the product title is that attribute.

3.4.2 ResNet-18

The Residual Network called ResNet came into existence after the evolution of Convolutional Neural Network (CNN). The need for deep neural network showed efficiency towards better performance. Additional layers up to 18 were introduced for solving complex problems with improved results Targ et al. (2016). So, ResNet-18 is a CNN based deep neural data model that is exactly 18 layers deep. The problem of training deep layers resulted to Residual Network or Reset.

3.4.3 ResNet-50

ResNet-18 is again CNN based neural network that is exactly 50 layers deep. Shen and Savvides (2020) The need for more layers is to alleviate the model training issue, the significance of using ResNet-50 is that it comes along with a version pre-trained network from ImageNet that contains a set of more than million images. There are many categories of images of objects such as computer accessories, stationery objects, fashion apparels, animals and so on that can classify images up to around 1K objects.

3.4.4 Siamese ResNet-50

A Siamese neural network can be defined as a neural architecture containing two identical neural networks Han et al. (2019), each taking one of the two input data for comparison. A function called as contrastive loss is used to compare the similarity score between each pair of input image data, the last layers of two identical network is fed into this function. Siamese ResNet-50 is again a CNN based neural network containing 2 identical neural network each with 50 layers deep.

3.5 Results and Evaluation

In this stage, the results achieved from the models are evaluated just before final implementation plan. It is needed to evaluate the model against certain metrics to be confident in measuring the model performance, CV score is one evaluation metric used here. The analysis of results is carried out to ultimately target and meet the research goals.

3.6 Research Deployment

The last step involved in the research delivery and methodology is to deploy the implemented models and interpret it to gain knowledgeable insights. Extra care is provided at this step to monitor if the research study done is meant to be an essential part of day-to-day life. The identified action varies from different types of systems where implementation it is performed.

4 Design Specification

An architectural framework [5]is developed to enable efficient design in classifying whether the product is similar to other products or not. The different methods, techniques and tools are used to design the architecture for a deep learning model which contains presentation layer where insights are gathered through visualization of the achieved results and business logic layer where the data sources, data processing steps and models are carried out before in a two-tier architecture . Below figure 2 is the design specification architecture implemented for product match detection based on images and text data as shown.

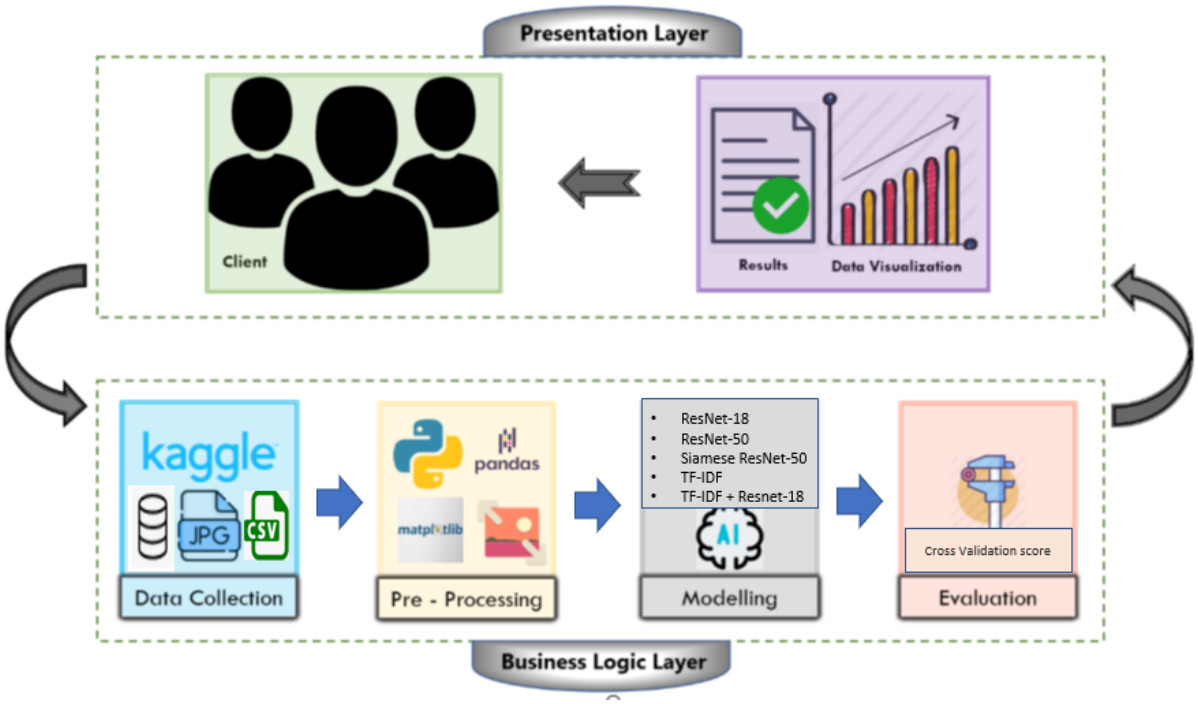

Figure 2: Design Approach

[5] architecture: https://towardsdatascience.com/retrieving-similar-e-commerce-images-using-deep-learning-6d43ed05f46b

In this research, significant importance on the design work is given to the business logic layer. The implementation of this tier of the architecture is explained in detail under section 5 but the key highlights of it is given in the below table 1. Every effort is taken to keep the implementation of the design in an iterative and experimentation approach and to conclude with the best derived results.

Table 1: Business Logic Layer

Data Collection	**Pre-Processing**	**Modelling**	**Evaluation**
Shopee DataSet	Vectorization	TF-IDF	Validation Loss
Sourced at Shopee website	Embedding	ResNet-18	CV Score
Product Title and Images	Augmentation	ResNet-50	Experiments
JPG and CSV format	Remove Duplicates	Siamese ResNet	Model Comparison

Enough information about the configuration details related to the hardware and software components, specifications to carry out the design implementation are listed in the configuration manual document submitted along with this research paper.

5 Implementation

The implementation is built in a way to identify which product images and product text together contain the same e-commerce products. The result of this is achieved through various implementation challenges and some of which are highlighted below:

- To find the near duplicate products and just not only the images
- The impact of erasing the area surrounding the products, the background area of the product in particular
- By using the descriptors of the product images or the product title

5.1 Data Collection

The desired framework for Product Similarity Match Detection using Ecommerce Images and text is implemented and evaluated using the Shopee dataset published by the Shopee Ecommerce organization. This dataset was made available in Kaggle as competition publicly during 2020. The data is not directly fetched from kaggle rather a link available at kaggle is used to fetch the data directly from shopee website. The data set includes 32,400 images and texts of the ecommerce products from various product categories. The analysis of this data is used to derive output of identical products among them by relevant training and testing. The main attributes present in the dataset as seen in figure 3 includes product title, product image and image phash. Label group and PostingId is also equally important. Phash is a perceptual hash technique that is used to group the images by generating random number based on hash algorithm. A sample dataset as how it looks like can be referred to below figure 3. The next primary step involved is to analyse the data structure of the image and text separately. From here on the data should be tried to categorise into two sets of classes which can be named as match detected and no match detected by grouping them.

posting_id	image	image_phash	title	label_group
train_129225211	../input/shopee-product-matching/train_images/...	94974f937d4c2433	Paper Bag Victoria Secret	249114794
train_3386243561	../input/shopee-product-matching/train_images/...	af3f9460c2838f0f	Double Tape 3M VHB 12 mm x 4,5 m ORIGINAL / DO...	2937985045
train_2288590299	../input/shopee-product-matching/train_images/...	b94cb00ed3e50f78	Maling TTS Canned Pork Luncheon Meat 397 gr	2395904891
train_2406599165	../input/shopee-product-matching/train_images/...	8514fc58eafea283	Daster Batik Lengan pendek - Motif Acak / Camp...	4093212188

Figure 3: Sample Dataset Description

5.2 Data Preprocessing

Initially, data is loaded into data frame then the text attribute which is the title column from figure 3 is assigned with suitable weights through vectorization. KeyedVectors function is used for text to vector conversion. To work with images, the basic image pre-processing step should be to transform the image into its suitable RGB format. The better the images are consistent, the comparison is made easier. This calls for equal image resizing as part of pre-processing step, the image resolutions are evenly poised by converting the image resolution as 512*512. Using Pytorch libraries the image pre-processing is carried out in simple steps, the image normalization is an important process as it ensures that each input pixel parameter has equal data distribution as it makes convergence faster during model training. This is achieved by transforming the images using torch library functions pretrained on Imagenet where per channel mean and standard deviation are calculated, by setting the values of mean = [0.485, 0.456, 0.406] and standard deviation = [0.229, 0.224, 0.225] ".

5.3 Experiments

The title and image of the products are separately experimented through various techniques and then execute the model to find the best performing behaviour. The different ways to approach the image and text data are as below.

5.3.1 Text Preparation

Although, vectorization is the major deal in text processing the need for other data preparation steps could prove it's worth. We'll have to tweak it a little bit, so the important insights about the data can be gained afterwards as close to accurate and possibly enhancing the model performance. The check for any special characters available and removing punctuation, stop words, white spaces and covert all text to lower case is done alright as seen in figure 4, I prefer not to take off the numbers as the details about number of product quantities may miss out, these numbers in turn produce high prediction rate. All these are done through relevant pandas' library functions 4.

```
Before: Aroma Terapi Pengharum Ruangan Merek Josmine / Aroma Pengharum R
uangan

Lower case: aroma terapi pengharum ruangan merek josmine / aroma penghar
um ruangan
Remove punctuation: aroma terapi pengharum ruangan merek josmine  aroma
pengharum ruangan
Remove whitespaces: aroma terapi pengharum ruangan merek josmine  aroma
```

Figure 4: Different Text Processing

5.3.2 Image Embedding

To try improving the model performing at run time, it is now important to carry out PCA analysis on the image attributes. The need to reduce the dimension of the input image data is evident as the resolution is set of 512*512. Using Image embedding technique, the input images are further converted into low-dimensional vectors that can be more easily processed by the models. The embeddings created are the abstract representation of the images, the input of an image for three channels with size as 224*224 is described as [3, 256, 256] , the output delivered is an array of 1k items representing the exact structure of the input seen in below figure 5

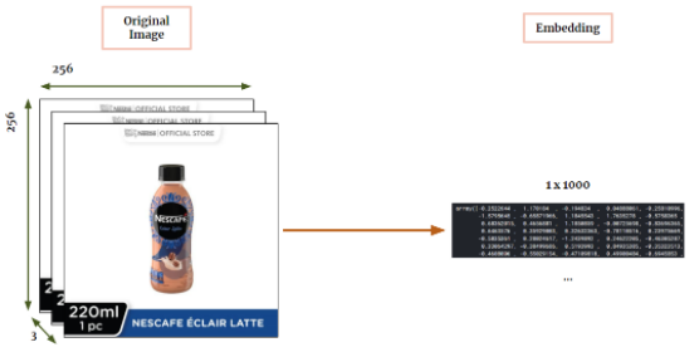

Figure 5: Image Embedding approach

5.3.3 Image Augmentation

Image Augmentation is another aspect that needed consideration, the existing data need to be altered to see which augmentation technique best suits this kind of data. Rotate, Crop, Blur, Flip, Invert, Gamma shown in figure 6 are the augmentation techniques generally used. The factual reason to use these augmentations is that they neither change the original colour nor the texture rather only display the products in different directions and angles.Iterations with each augmented data frame is used for selective model trails to clearly declare the best resulting approach to be further utilised.

- **Why Image Augmentation is important?**

As a trial and error approach, each augmentation is carried out and fed into the model to understand the the best performing augmentations for this type of problem

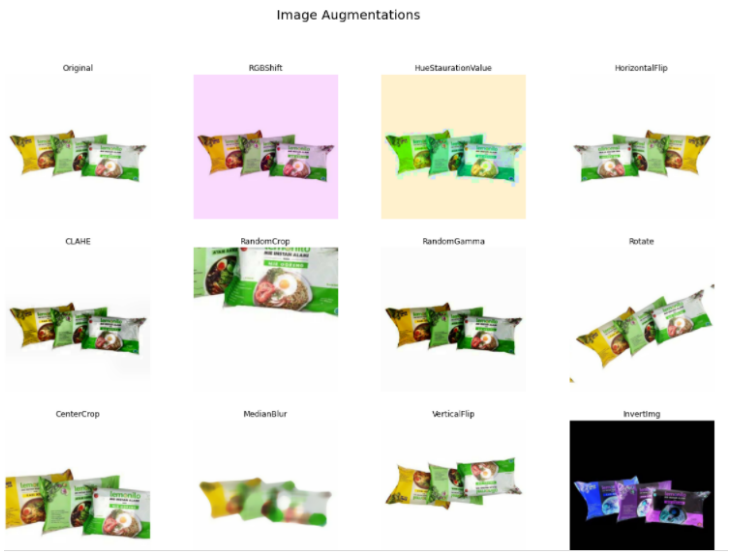

Figure 6: Image Augmentation Experiment

scope. Among all the image augmentation methods, Resnet-50(C+F) is best, where (C+F)= Crop and Flip. For example, the class Women Kurta is often confused with Gowns, Mojari Men with Women since these pair of classes are similar to each other.

5.3.4 Clean Duplicates

Figure 7: Duplicate Image Example from Shopee Dataset

The overall data have duplicates, there are 1,246 images that have 2 or more apparitions from figure 9. The title differs for most of them and the label group is usually the same, but there are a few cases where it differs as well. The duplicates are analysed by understanding the deep differences by looking at the picture where the names of the images are identical. The wordings of the text descriptions can be different though it refers the same images, similarly the image group ID varies in certain cases even though the images are identical. A sample of duplicates in seen below figure 7. This only means the title description is the attribute which indicates the image category it belongs to. The input in each experiments for different iteration is fed into the models created to test and train for evaluation.

5.3.5 Text Feature Extraction

The text attribute related to the input data is only title column, it is experimented to see if any other features can be extracted from here. The below figure 8 details the the relevant extractions carried out, the word count feature is extracted based on number of words in a sentence, the char count is again number of characters in a sentence, The average length of word in a sentence, the stop words, the counts number of stop words present in a sentence, number count is exact numerics in a sentence. The feature is then extracted and the results are experimented based on the best text feature extracted.

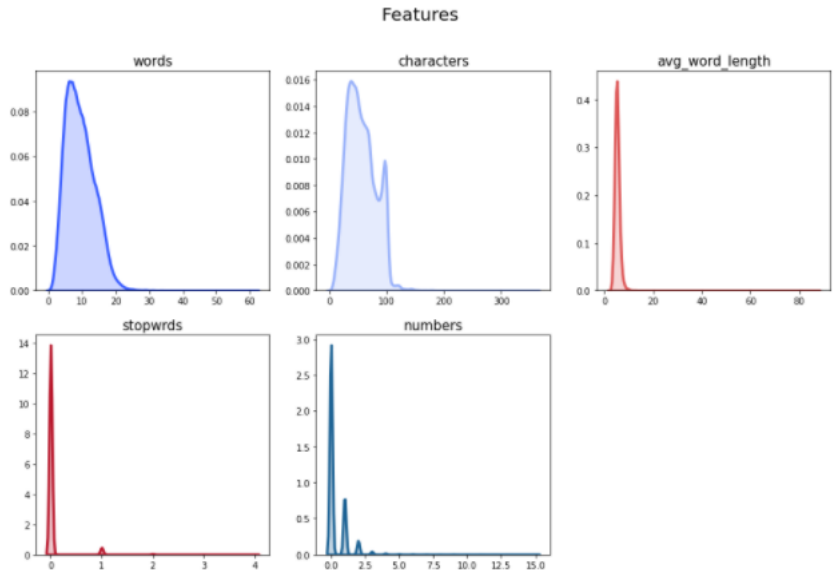

Figure 8: Features Extracted from Title

- **Why Removing Duplicates and Text processing is important?**

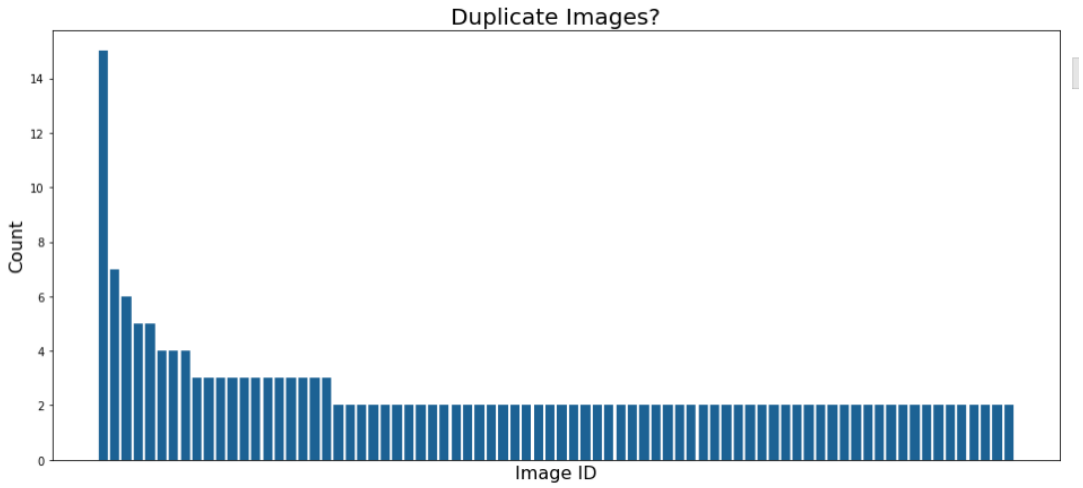

Figure 9: Number of Duplicate Images based on Image ID

The above image augmentation performed well when the duplicates in the data were removed and with epochs made consistent for each iteration for upto 10, the performance of model increased with respect to computational time. The clearer text processing only adds up to better implementation with only duplicates being removed.

5.4 Hyperparameters

All the experiments are run on Google Colab and Cloud Platform by using required infrastructure to support our requirements. We use Cross Validation loss score and better learning rate to determine the best performing model. The hyperparameter setting for ResNet is taken from the experiment with best performing test and train accuracy. For consecutive 10 epochs to 20 epochs, when validation loss remains almost constant and consistent the parameters are tuned with the hyperparameter values maximum learning rate = 0.0001, nesterov momentum = 0.5 and weight decay = 0.01 achieving the highest accuracy.

5.5 Model Results and Evaluation

The model results are extracted on iteration basis through data processing of image and text separately and running the models. To achieve the desired result and evaluate the research study, the focus is on the comparison of various deep learning models such as ResNet-18, ResNet-50, Siamese ResNet-50, TF-IDF + ResNet-18 in order to predict the product similarity detection and so the hyperparameters passed are made constant for all models. The learning rate value is 0.001 and the batch size is maintained consistent at 40.

5.5.1 Iteration 1

With the initial data pre-processing approach, the data is fed into the models in train set and then validate it with test samples. The performance of the model was against par on the overall time involved to run the data models, one reason could be due to heavy image size and the resultant of the model ran was below par based on the evaluation metrics conducted. The initial iteration of the model run finds way to fine tune the model further and look out for other data pre-processing methods that could increase the model accuracy and its performance.

5.5.2 Iteration 2

For better performance, further data preparation for the text and image data are separately handled through image augmentation and text processing as detailed in section 5. The result of the data preparation is again fed into the model, significant improvement in the computational time and model performance was evident during this iteration.

- **TF-IDF Vectorizer**

Through text vectorization the word to vector conversion is done using fitting to the model. Since the stop words are already removed as part of text processing, during TF-IDF vectorizer model parameters are set as true for Binary, None for stop words and maximum feature are given as 55k.The model is then shaped into an array embeddings,

torch matrix multiplier function is used to calculate cosine similarties in chunks of 1024*4 each. Cross Validation score is then calculated for predicting similar titles.

Table 2: TF-IDF

Model	**Metric**	**Value**
TF-IDF Vectorizer	CV Score	0.613

- **ResNet-18**

ResNet has pretrained version images from Imagenet and the test images can be classified using this model containing 17 CNN layers and 1 softmax layer is used to achieve the classification. The CNN layers uses 3*3 filters and the output feature is same size as containing an average pooling layer 2D used to retain features through minimum sampling and a fully connected softmax layer is configured at the end. No further information loss helps prevent the over-fitting issue and the input of which is fed into the ResNet 18 model to produce the output to classify whether the products are identical or not.

Table 3: ResNet-18

Model	**Metric**	**Value**
ResNet-18	CV Score	0.652

- **ResNet-50**

Model ResNet-50 contains an advanced version of ResNet architecture that contains 48 CNN layers and 1 average pool layer along with 1 maxpool layer. Also with five stages each a convolutional block consists of 3 CNN layers and Identity block also consists of 3 CNN layers. By configuring GPU for TensorFlow with memory limit of 1024 in the configured virtual machine. Again matrix multiplier function is used to create an array and consine similarities are used to calculate the distance of near identical items using indexes.

Table 4: ResNet-50

Model	**Metric**	**Value**
ResNet-50	CV Score	0.663

- **Siamese ResNet-50**

The Siamese Network created with test and train loops calculates the triplet loss value by the use of 3 embeddings. The weights are considered from Imagenet and pooling is set to average pool where as the activation function used is linear. A distance layer is built which is responsible to compute distance between anchor embedding and positive and similarly distance between anchor embedding and negative. The output of the compute loss network is a tuple that holds the distance determining the identical e-commerce products.

Table 5: Siamese ResNet-50

Model	Metric	Value
Siamese ResNet-50	CV Score	0.712

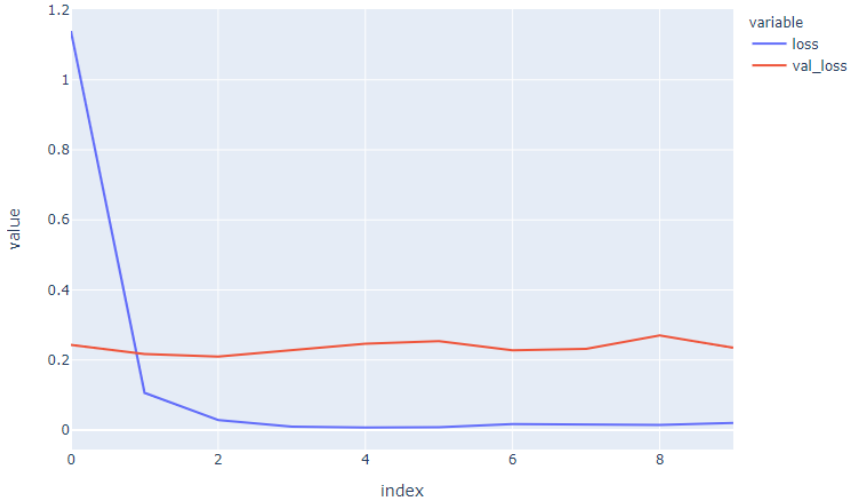

Figure 10: Siamese validation loss graph

5.5.3 Iteration 3

The scope of this research work is finally delivered in this iteration, the need to prove the prediction improvement when the combination of text and image both used together is under test, there is no further data preparation carried this time as discussed in section 5. In this rather the exact models used are replicated for text which is TF-IDF vectorizer and for image, ResNet-18 is used for calculating the cross validation score.

- **TF-IDF Vectorizer + ResNet18**

Using the image Phash input data, a temporary image hash is created in the dictionary by doing a group by on the unique values of image phash. A base cross validation score is calculated on the training data by mean on the f1 score which is 0.553 . The positing id is generated based on the group labels for the image, text and baseline model separately, this posting id of all is then concatenated using a concatenate function. The CV score is then computed by finding the match applied on all the three models, trained separately based on the mean f1 score. The unique values that do not have a match are retained to represent the unmatched products that is retrieved through index value as false, the location of the products are traced back through the hash values located at the dictionary.

Table 6: TF-IDF Vectorizer + ResNet-18

Model	Metric	Value
TF-IDF Vectorizer + ResNet-18	CV Score	0.734

6 Discussion and Model Comparison

The model evaluation and comparison is needed to study the results gathered along the project implementation. Evaluation metrics is used to evaluate the model classification accuracy using cross validation score in finding the similarity match in e-commerce products. The research done here is mostly performed using the multiclass data. On the basis of prior research, the several selected models examined for the study proved the results as anticipated . The research results were pretty satisfactory and considerably expected the similarities in the e-commerce items. The following is a comparison of the many models utilized in this study as shown in figure 11. The algorithm with the highest validation accuracy of 0.75 cross validation score was Siamese ResNet-50. The discussion about the use of Siamese ResNet-50 which is just a Siamese ResNet upgraded version, the method was validated against the dataset resulting with high validation score among all only on image attribute. With decent computational time 52:25 mins as compared to others but the combination of text and image model has an improved CV score.

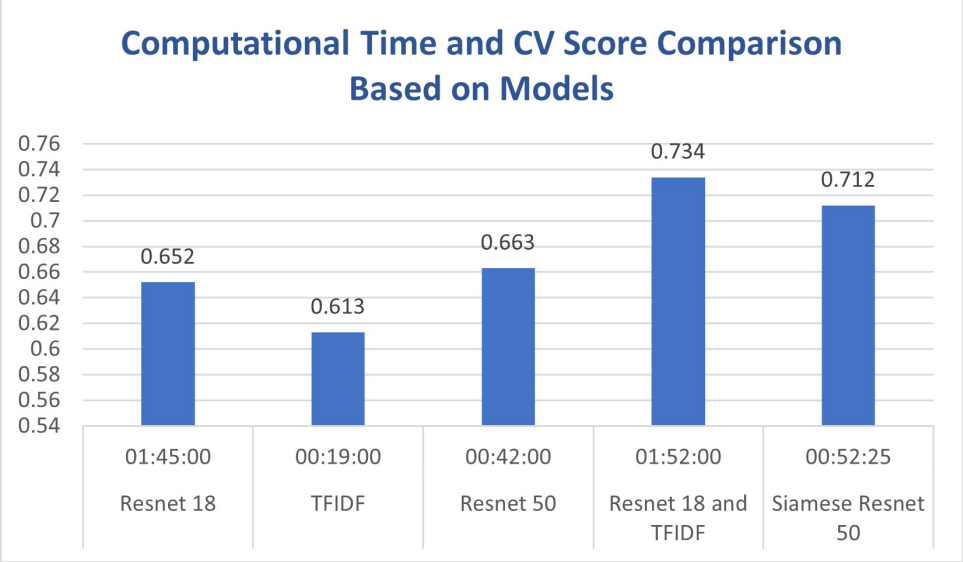

Figure 11: Comparison of Models based on CV Score and computational time

Table 7: Model Comparison with best augmentation type for only image models
The table illustrates the effect of applying different image augmentations. The best performing image only models notated as R=ResNet, (C)= Crop augmentation, (F)=Flip, and (C+F)=Crop and Flip together, (NA)=No Augmentation applied

#	**Model**	**CV score**
1	ResNet-18(NA)	0.652
2	ResNet-18(C)	0.612
3	ResNet-18(F)	0.691
4	ResNet-18(C+F)	0.701
5	ResNet-50(C+F)	0.712
6	Siamese ResNet-50(C+F)	0.722

.

7 Conclusion and Future Work

The research paper here discussed the highlights 1.4 of the neural networks using deep layers and siamese twins to find the similar products using images and TF-IDF vectorizer for title description on shopee website [6] dataset containing identical e-commerce products with subtle varying title descriptions but broadly different images. The models and methods employed have been based on a rigorous examination of several previous research 2 related to e-commerce product similarity match detection. A detailed comparison of the results obtained from ResNet-18, ResNet-50, Siamese ResNet-50 and TF-IDF + ResNet-18 models and its performance is discussed. The model with the highest cross validation score leads the way towards identifying more accurate deep learning method to be implemented for similar business use case. Along with the more accurate model, the computational time metric is used to evaluate the best performing model on time taken. Although we see a good accuracy towards using Siamese ResNet-50, the computational time for it is high for 10 epochs and TF-IDF + Resnet18 performs well in finding more similar items using both text and images and have reasonable computational time as well.

With the use of other relevant data from open sources that are publicly available in online platforms, the research can be further carried out to predict the similar products in the e-commerce website. The need for tuning the model with other advanced parameters and metrics can be taken ahead to derive any further better results. The similar combination technique can be applied on ResNet-50 and Siamese ResNet-50. Even though we achieve 0.73 cross validation score, more research towards developing better classification models can incentivize products similarity identification. The dataset can be improved by balancing the number of product group images per category. In future, we can plan to extend the dataset by introducing more categories and large product groups from other e-commerce data. Also, the list of different augmentation methods can be tried to validate similarly in table 7 which best fits to each business use cases in identical product detection in e-commerce.

- **Research Acknowledgement**

I am thankful for the unwavering support and feedback during the research study from my supervisor Jorge Basilio. He was really helpful and knowledgeable. With the necessary support for the paperwork, he directed me towards the right direction in completing the masters project.

References

Addagarla, S. K. and Amalanathan, A. (2020). Probabilistic unsupervised machine learning approach for a similar image recommender system for e-commerce, *Symmetry* **12**(11): 1783.

Akritidis, L., Fevgas, A., Bozanis, P. and Makris, C. (2020). A self-verifying clustering approach to unsupervised matching of product titles, *Artificial Intelligence Review* pp. 1–44.

Angelovska, M., Sheikholeslami, S., Dunn, B. and Payberah, A. H. (2021). Siamese neural networks for detecting complementary products, *Proceedings of the 16th Conference*

[6] Dataset: `http://shopee.com`

of the European Chapter of the Association for Computational Linguistics: Student Research Workshop, pp. 65–70.

Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. (1996). The kdd process for extracting useful knowledge from volumes of data, *Communications of the ACM* **39**(11): 27–34.

Fuchs, G., Acriche, Y., Hasson, I. and Petrov, P. (2020). Intent-driven similarity in e-commerce listings, *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pp. 2437–2444.

Han, G., Du, H., Liu, J., Sun, N. and Li, X. (2019). Fully conventional anchor-free siamese networks for object tracking, *IEEE Access* **7**: 123934–123943.

Hasson, I., Novgorodov, S., Fuchs, G. and Acriche, Y. (2021). Category recognition in e-commerce using sequence-to-sequence hierarchical classification, *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pp. 902–905.

Huang, H. J., Yang, J. and Zheng, B. (2019). Demand effects of product similarity network in e-commerce platform, *Electronic Commerce Research* pp. 1–31.

Khan, R. U., Zhang, X., Kumar, R. and Aboagye, E. O. (2018). Evaluating the performance of resnet model based on image recognition, *Proceedings of the 2018 International Conference on Computing and Artificial Intelligence*, pp. 86–90.

Li, J., Dou, Z., Zhu, Y., Zuo, X. and Wen, J.-R. (2020). Deep cross-platform product matching in e-commerce, *Information Retrieval Journal* **23**(2): 136–158.

Moraes, F., Yang, J., Zhang, R. and Murdock, V. (2020). The role of attributes in product quality comparisons, *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval*, pp. 253–262.

Niu, X., Li, C. and Yu, X. (2017). Predictive analytics of e-commerce search behavior for conversion.

Rajest, S. S., Sharma, D., Regin, R. and Singh, B. (2021). Extracting related images from e-commerce utilizing supervised learning, *Innovations in Information and Communication Technology Series* pp. 033–045.

Rajput, P. S. and Aneja, S. (2021). Indofashion: Apparel classification for indian ethnic clothes, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3935–3939.

Sarvi, F., Voskarides, N., Mooiman, L., Schelter, S. and de Rijke, M. (2020). A comparison of supervised learning to match methods for product search, *arXiv preprint arXiv:2007.10296* .

Sezavar, A., Farsi, H. and Mohamadzadeh, S. (2019). Content-based image retrieval by combining convolutional neural networks and sparse representation, *Multimedia Tools and Applications* **78**(15): 20895–20912.

Shen, Z. and Savvides, M. (2020). Meal v2: Boosting vanilla resnet-50 to 80%+ top-1 accuracy on imagenet without tricks, *arXiv preprint arXiv:2009.08453* .

Shrivastava, R. and Sisodia, D. S. (2019). Product recommendations using textual similarity based learning models, *2019 International Conference on Computer Communication and Informatics (ICCCI)*, IEEE, pp. 1–7.

Targ, S., Almeida, D. and Lyman, K. (2016). Resnet in resnet: Generalizing residual architectures, *arXiv preprint arXiv:1603.08029* .

Trappey, C. V., Trappey, A. J. and Lin, S. C.-C. (2020). Intelligent trademark similarity analysis of image, spelling, and phonetic features using machine learning methodologies, *Advanced Engineering Informatics* **45**: 101120.

Tsagkias, M., King, T. H., Kallumadi, S., Murdock, V. and de Rijke, M. (2021). Challenges and research opportunities in ecommerce search and recommendations, *ACM SIGIR Forum*, Vol. 54, ACM New York, NY, USA, pp. 1–23.

Zuo, Z., Wang, L., Momma, M., Wang, W., Ni, Y., Lin, J. and Sun, Y. (2020). A flexible large-scale similar product identification system in e-commerce.