Non tabular feature-based Prediction of Product Acceptance Rates

SDSC 8007- Deep Learning

Group 5

LIU MING	OU YANG	Siu Him Kan, Steve
mliu239-c@my.cityu.edu.hk	yangou3@cityu.edu.hk	shkan9@cityu.edu.hk
ID 56996684	ID 58049584	ID: 58193441

1. Motivation

With rapid advancements and breakthroughs in the subfield of machine learning - Natural Language Processing (NLP), the importance of a feedback system that evaluates the success factors of a product have increased as content generation becomes more accessible.

According to Amazon's official forum, a generative AI solution was launched in 2023 to help sellers write product descriptions. The objective is to create a seamless product listing procedure, allow content enrichment and help customers to make confident purchase decisions (Westmoreland 2023) [1]. Similar content generative solutions such as GPT, Imagen and Mid journey will be changing the routine operations at many traditional ecommerce platforms and marketing firms. A recent study says that 87% of buying online decisions are reliant on product descriptions (Alibaba 2023) [2]. It is also widely recognized that user ratings and appealing images significantly enhance online shopping experience. 87.6% online shoppers were reported to consider product images as the main component of their shopping experience in 2023 (Dropics 2023) [3].

There are many novel approaches to evaluate user acceptance rates given product images, descriptions and user feedback. In the field of machine learning, algorithms such as Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM) are commonly used for data that has sequential dependencies. Unlike traditional statistical algorithms such as the Auto Regressive Integrated Moving Average (ARIMA), these deep learning algorithms have more trainable parameters and considerations of past information. Transformers models can also be considered as they provide more predictive power and introduce parallelism that RNN and LSTMs lack. This was made possible with the introduction of an attention mechanism and an encoder / decoder framework (Vaswani et al 2017) [4]. In most business scenarios, these algorithms are trained on user feedback as a classification problem to calculate an appropriate rating for a particular product. On the other hand, images are commonly evaluated with Convolutional Neural Networks (CNN). These networks combine convolutional layers with sizable kernels to extract latent features and fully connected layers acting as a classifier. There is a wide range of pre-trained CNNs for selection. Most notably, Efficient_net, VGG16, Mobilenet and RestNet. These models are widely used for classification tasks such as identifying the product entity, gender and product classifications.

The primary contributions of this study can be outlined as follows:

- Data analysis and preprocessing that contains word vectorization, class balancing and image feature selection
- Experiments with a novel modelling approach that utilities both product reviews and descriptions
- Proposed the idea of combining image latent features with natural languages
- Validate the proposed model's efficiency with baseline models
- Model implementations / deployments alongside with future analysis

2. Background / Relevant Work

Sentiment analysis and predicting product acceptance rates have been well researched in the field of natural language processing. In the paper (Habib et al 2023) [5], a Convolutional Neural Network (CNN) was used alongside Long Short-Term Memory (LSTM) to make inferences on user product ratings. The proposed model shows an accuracy of 95% when compared to other algorithms such as Random Forest, Logistic Regressions and XGBoost. The author of the paper also had an emphasis on data processing standards, which includes lower casing and stop words removal as they were considered not significant and not related to emotions.

In the paper (Chin et al 2018) [6], the idea of aspect extraction was proposed under the consideration that not all words carry the same amount of importance in a sentence. The proposed model uses both user and item documents as inputs. Aspect level layers were used after the embedding layer to extract latent features and later concatenated together to make rating predictions. The proposed model was applied on multiple dataset, and the lowest Mean Squared Error of 0.688 was achieved for the Digital Music Category within the Amazon Product Dataset.

This idea of feature extraction was expanded even further in the paper (Xu et al 2022) [7]. The authors proposed a parallel deep neural network that trains on both review sentiments, product descriptions and images to predict product ratings. Aspect sentiment and word aware attention layers were also used in their paper to extract latent features and aspects within the word corpus. Attention layers are the founding block of this model as the authors indicated that comments made by users on an item can be represented with words that carry vastly different sentiments. This model also incorporated a visual embedding layer alongside with an attention layer before concatenating with natural languages. This proposed model achieved a better mean squared error than baseline models such as the ANR model recommended by (Chin et al 2018) [6] in most of the Amazon product categories.

While it may be considered unconventional, there are existing papers that detail different approaches to anticipate ratings with the usage of non-tabular data. In the paper (Truong et al 2017) [8], the author used AlexNet as their foundational model with customised CNNs and Fully Connected layers to encode item orientations. Review images from Yelp covering businesses in different US cities were used in their experiment. Various implementations of this concept were examined to assess result variations. Among all the models tested, introducing item orientations to the Fully Connected layer yielded the highest accuracy of 63% for review images from Chicago. Baseline models such as the Naive Bayes and modified AlexNet achieved an accuracy of sub 56%.

In the paper (Chaudhuri et al 2018) [9], optimal image orders were examined with the use of well known algorithms such as VGG19 and Resnet50. The main contribution of the paper resides in dealing with data sparsity and non-compliant or inappropriate images. Experiments within the study train on images of durable goods and the average accuracy of each classifier for 4 different product segments (furniture, apparels, electronics and everyday living) is above 75%.

3. Dataset, Exploratory Data Analysis and Data Preprocessing

In this paper, we will be using Amazon's 2018 Review Data [10]. The prepared dataset has a total of 24 categories, which includes but not limited to Books, Elecontrics, Digital Music and Beauty. For computational efficiency, we have decided to use a subset of this particular dataset. Specifically, this subset, also known as 5-cores, is a truncated version of the original dataset where users and items have a total of 5 reviews each. In the following analysis, we have selected Video Games as our category of interest, and the subset has been merged together with the metadata to retrieve other information including product descriptions, price and image. Table 1 displays the aggregation count per product rating.

Table 1: Distribution of Product Rating

Product Rating	Reviews Count
5.0	294155
4.0	95437
3.0	50006
2.0	30084
1.0	24544

From Table 1, we can see that the issue of class imbalance is very prominent in our subset. Ratings of 4.0 and above is at least 75.46% and the rating of 5 takes up at least 50% of our subset. To allow our model to generalise, we will be combining oversampling and undersampling so that the class ratio is approximately 1:1.

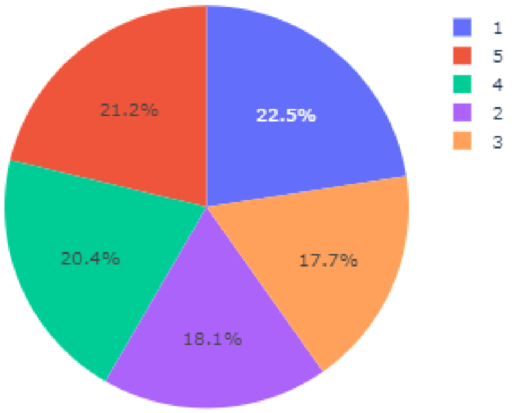


Fig 1: Pie chart of Rating Distribution after sampling

As for non-tabular features such as Natural Languages, we have used both Bag-of-words and Term Frequency - Inverse Document Frequency (TF-IDF) approaches to analyse the common words used by Amazon users for the Video Games category.

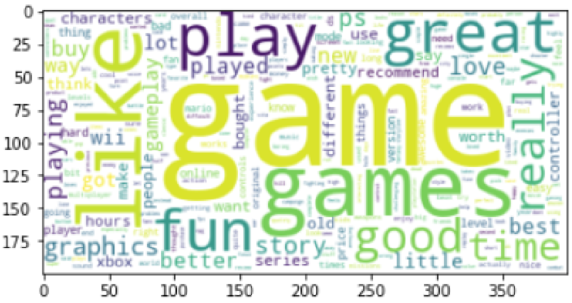


Fig 2: Word cloud using Bag of Words

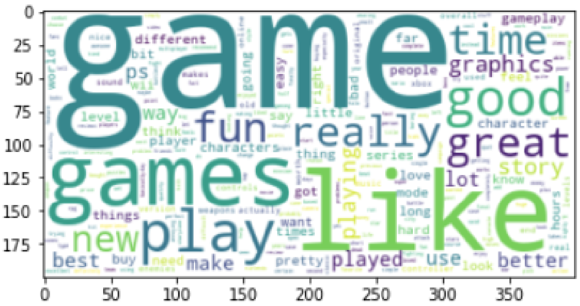


Fig 3: Word cloud using TFIDF

Common words that have positive cognitions such as good, fun, like, recommend, great and best appeared the most as shown from the word clouds above. This phenomenon is expected as we have shown that a large number of positive reviews plagues this underlying subset. We can approach this analysis again after sampling and focus on negative reviews only.

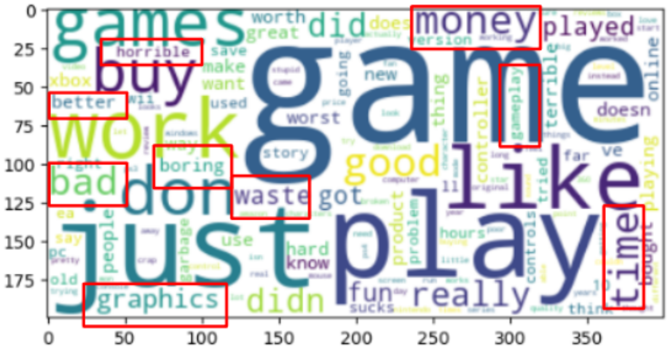


Fig 4: Word cloud with rating =1 (subset with 10% of reviews)

Review text and product descriptions will be vectorized using a pretrained embedding model, GoogleNews-vectors-negative300, in most of our models. Our proposed model will use a pretrained BERT model embedding instead. Punctuation and stopwords removal can be considered, but they did not show a significant performance boost in our experiments. All vectors are padded to an equal length and unseen words will be initialised with zeros. Traditional algorithms such as TFIDF and Bag-of-words will not be used during model training as they do not account for word meanings and underlying semantic relevance at the word level.

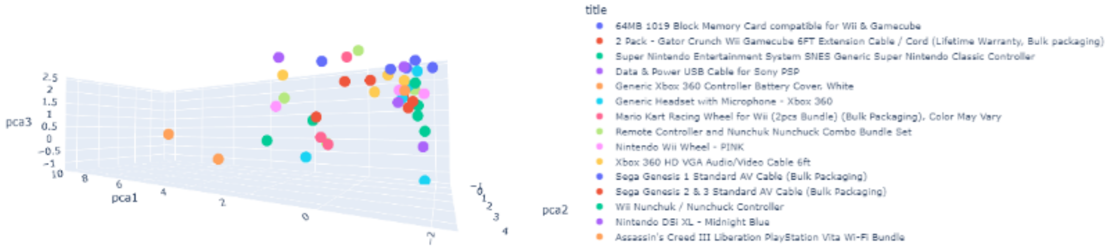


Fig 5: Product Similarity in 3-dimensional space using TFIDF (20 products)

The dataset consists of three primary image categories: Video Games Cover, Equipments and Screenshots. These images are different in nature, but the focus of this paper lies in examining the artistic elements such as tone, vectors, colour choice and the overall art style of their product listing. Comparisons made against two different types of image will be more focused on their image colour representation and rather than their product characteristics. For this exact reason, we have simply chosen one image to present each product in our experiments. Future studies should consider implementing a more comprehensive image selection process, as briefly discussed in the previous section, to account for the correlation between image selection and customer engagements.

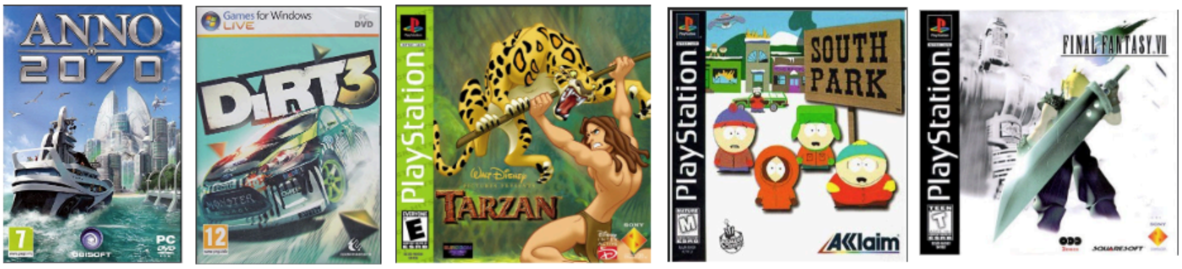


Fig 6: Product Images (Video Games Cover)

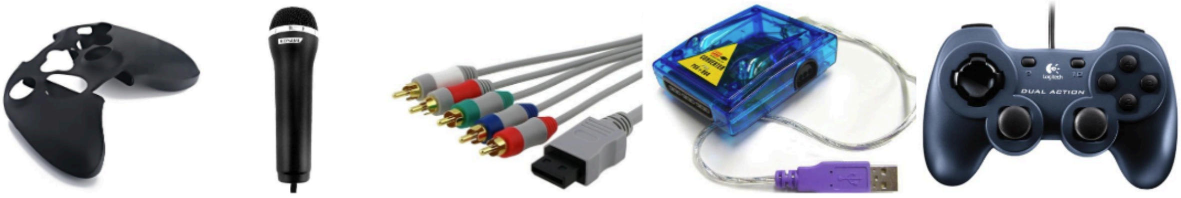


Fig 7: Product Images (Video Games Equipment)

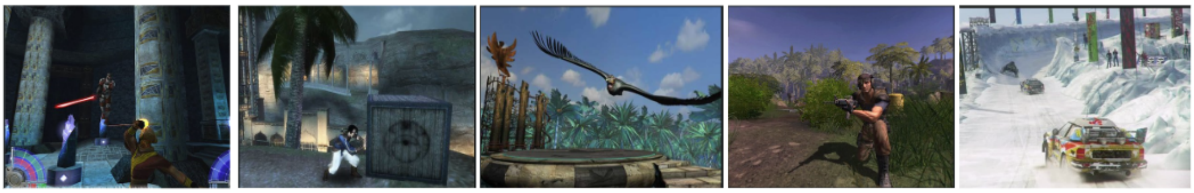


Fig 8: Product Images (Video Games Screenshots)

4. Proposed Model

Bert DistilBertModel was applied in the embedding of the text, including both description and reviews. DistilBERT is a small, fast, cheap and light Transformer model trained by distilling BERT base. It has 40% less parameters than bert-base-uncased, runs 60% faster while preserving over 95% of BERT’s performances as measured on the GLUE language understanding benchmark (DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter).As the model required length no longer than 512, so we just use 50 length of description and 462 length of review because the review was assumed to be more relevant to the rating score.

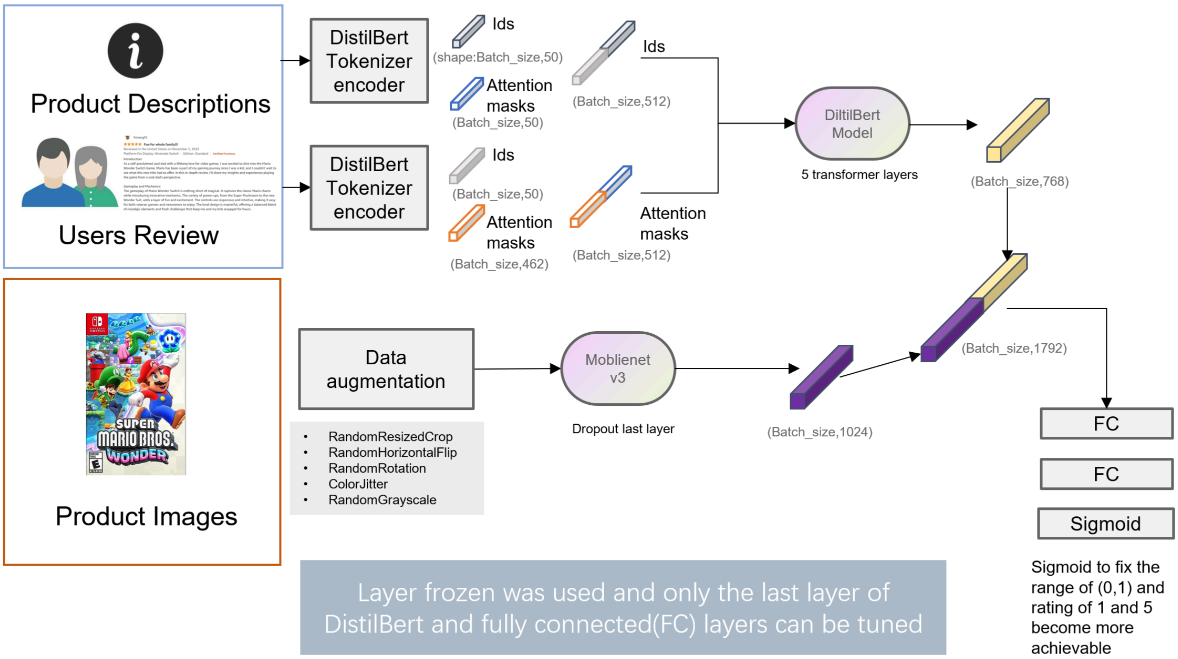


Fig. 9: The model structure of the proposed model

As is shown in Fig.9, from the text encoder side, the product description and user review all feed into DistilBerTokenizer encoder, each producing separated ids and attention masks. Then, their ids and attention masks were merged together separately with the same dimension. The DistilBert model then processed them into a 1D vector.

From the image encode side, Mobilenet v3 was used to extract the features of images, the images in the training set will first be augmented before feed into the Mobilenet. The last layer of Mobilenet was dropped as it is just used as an image feature extractor. After the feature extraction, the output was merged with the output of the text feature extractor to form a unified vector.

The joined vector will be passed into two fully connected layers and a sigmoid layer. The sigmoid layer will process the output ranging from 0 to 1. 5 fold of the output will be used as the final output. This process will give a relatively large feature range for rating 1 and 5.

For the training process, for speed consideration and extractor consideration, layer frozen was used and only the last layer(fifth) and fully connected layers can be tuned. On the contrary, the Mobilenet and the other 4 layers of transformer in the DistilBert model were set to constant weight. Other settings like optimiser and learning rate schedule were the same as the base model. The model has proven to show better results than the baseline model.

The training, validation and test set follow an 8:1:1 ratio split. The final result used the model with the lowest loss in the validation set to be implemented on the test set.

5. Baseline Models

All of our models will be under the regression spectrum as we have determined that user ratings are more realistic and insightful if they are presented on a continuous scale. Our loss function will be Mean Squared Errors (MSE), and the accuracy metric can be calculated if we simply round up the raw predicted user ratings.

To accurately assess the effectiveness of our proposed model, we have trained two additional baseline models (LSTMs + 1d CNNs & LSTMs +1d CNNs+ MobileNet). To establish a benchmark for comparison, we will be including the model performance of ANR and MSVA models proposed by (Xu et al 2022) [7] and (Chin et al 2018) [6]. While these models were also trained with Amazon Product Review Dataset, our evaluation process will not be a direct comparison of performance as we have simply trained on a subset of the original data in our experiments. A direct comparison of Mean Squared Errors (MSE) is not a good indicator of the better model in this situation.

Table 2. Performance Metric (MSE) for different models

Dataset	a) ANR	b) MSVA	c) Baseline 1	d) Baseline 2	(e) Proposed Model
Video Games	1.257	1.182	0.67	0.62	**0.546**

Instead, we will be evaluating the average percentage improvement of the MSVA model over ANR against the improvement we get from our proposed model over baseline 1 & 2. This decision stems from the similarity in the interest of our paper and the proposal of the MSVA model over the ANR model by incorporating both images and text, which the ANR model lacks consideration for.

Table 3. Model Comparisons

Metric	(a) vs (b)	(c) vs (d)	(c) vs (e)	(d) vs (e)
Raw reduction in MSE	-0.057	-0.108	-0.124	-0.074
Improvements (%)	4.53%	16.1%	18.5%	11.9%

a) Aspect-based Neural Recommender (ANR) uses both review text and product descriptions to approximate user ratings. It utilises aspect aware mechanism combined with aspect attention layers to extract latent features from users. The word embedding of choice is the pre-trained word2vec model, *GoogleNews-vectors-negative300*. (Chin et al 2018) [6]

b) MSVA combines user reviews, product images and user ratings to analyse sentiment. The model utilises both natural languages and images by using two parallel embedding layers and attention layers. To capture the sentimental properties related to different aspects in the data, a (aspect / visual) transformation layer is added to further improve the model performance. (Chin et al 2018) [6]

c) Baseline model #1 combines LSTMs and convolutional layers. Our implementation follows the general idea proposed in (Habib et al 2023) [5] where a series of convolution layers with filters are combined with the results from the LSTM layer. We have stacked 3 bidirectional LSTMs in our experiment. 3 different kernels sizes (2,4,6) were chosen in our convolutional layer to extract latent features from both product description and reviews. Parameter wise, dropout rate p(s) is set to 0.5 and LeakyReLU is used to introduce nonlinearity. This model was trained exclusively on Review text and product descriptions only.

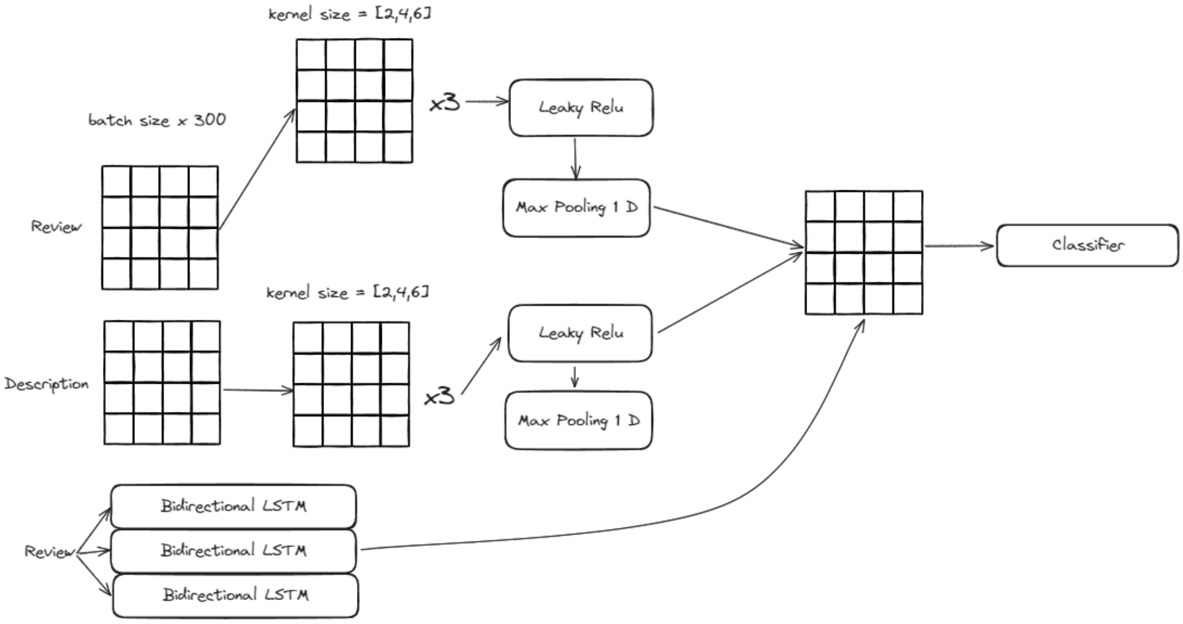


Fig 10: Model Structure for model 1

Different setups using Baseline #1 were also considered before constructing our proposed model. Results suggest that our selected pre-trained word embedding was able to reduce approximately 50% of the Mean Squared Errors.

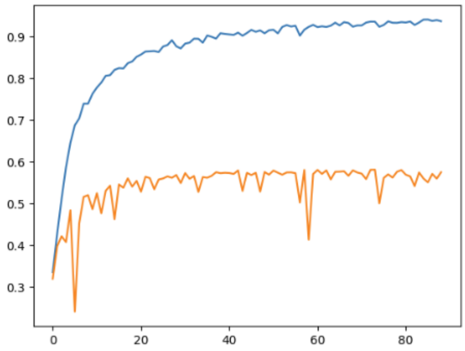
Table 4: Experimental Results using Baseline with / without pretrained embeddings

Setup	MSE
Without Pre-trained Embeddings	1.24
With Pre-trained Embeddings	0.67

d) Baseline model #2 combines baseline model #1 with a pretrained MobileNet model to extract latent features from images. This model was trained on both images and natural languages.

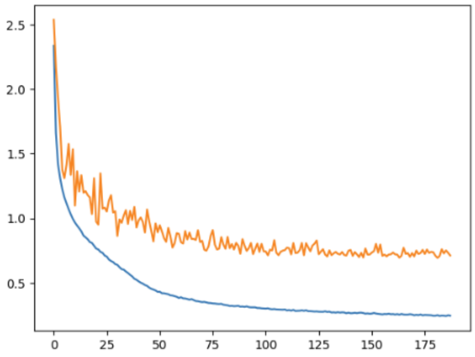
6. Model Comparisons

From Table 3, the performance boost using MSVA over ANR is approximately 4.53% for the Video Games Category. Our implementation of Baseline #2 that utilises both natural language and images achieved a lower MSE than Baseline #1 that was trained exclusively on text. The performance boost is approximately 16.1%. If we round up the predictions from both baselines, however, the accuracy for both models is approximately 58%. The performance boost of Baseline #2 and MSVA is approximately 11.57%. The test accuracy of our proposed model#3 is **66%**, which is much higher than the result of the baseline.



Blue: Training set | Orange: Validation set

Fig 10: Accuracy Plot for Baseline #1



Blue: Training set | Orange: Validation set

Fig 11: Loss Plot for Baseline #2

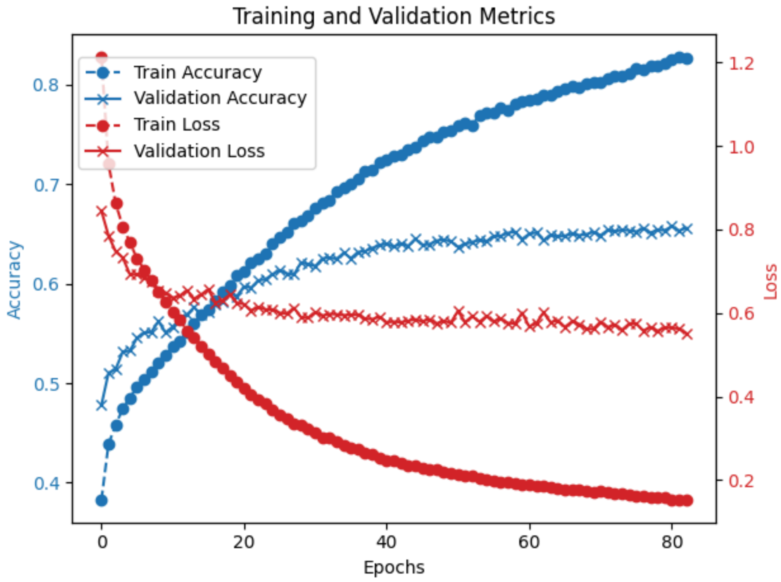


Fig 12: Accuracy and Loss plot for Proposed model #3

7. Conclusion

This paper presented a novel approach to approximate product acceptance rates by extracting latent features from both natural languages and images. We demonstrated that the inclusion of product images allows better approximation of user acceptance rates. The usage of Convolutional Layers on text data is also presented in this paper. Our proposed model with dedicated encoding layers exhibits better accuracy than traditional sequential models like LSTM and RNN. Bert model has proven to be better text feature extractor than traditional word embedding as it can utilise the giant transformer models and by training with large amounts of unannotated data, the Bert model can capture more semantic information. Frozen layers and fine-tuning are helpful to get expected results within a relatively short time.

The increasing accessibility of content generation solutions has led to a significant acceleration in the production of marketing campaigns and promotional materials in general. While it can benefit many traditional marketing firms and ecommerce platforms, we reckon that an evaluation module that incorporates the idea proposed in this paper is needed to address the question of how effective AI generated content will resonate with a certain target audience. In a production environment, it is highly recommended to re-trained or update the evaluation module on a scheduled basis to capture the ever-changing public opinion.

In future studies, a sophisticated image selection procedure should be considered to enhance the quality of the training process. Word embedding weights should be fine-tuned and other product related features such as price, helpfulness and popularity should be considered in the future as well.

References

1. Mary Beth Westmoreland. Amazon launches generative AI to help sellers write descriptions. September 2023. URL https://www.aboutamazon.com/news/small-business/amazon-sellers-generative-ai-tool

2. Alibaba. How to use SEO to write eye-catching product descriptions. April 2022. URL https://seller.alibaba.com/businessblogs/px001uc5n-how-to-use-seo-to-write-eye-catching-product-descriptions

3. Dropicts Pte Ltd. The Importance of Product Images in E-Commerce. July 2023. URL https://www.linkedin.com/pulse/importance-product-images-e-commerce-dropicts-pte-ltd

4. Ashish Vaswani, Noam Shazzer, Niki Parmar, Jakon Uszkoreit, Lilon Jones, Aidan N.Gomez, Lykasz Kaiser, Illia Popsukhin. Attention Is All You Need. *arXiv preprint arXiv:1706.03762v7 [cs.CL] 2 Aug 2023*

5. Md. Ahsan Habib, Asma Akter. Forecast the rating of online products using novel deep learning technique. *All Engineering Journal Volume 9, Issue 2, 2023 Page No. 28-37*

6. Jin Yao Chin, Kaiqi Zhao, Shafiq Joty, Gao Cong. ANR: Aspect-based Neural Recommender. *In: Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM 2018, Torino, Italy, 22–26 October 2018, pp. 147–156. ACM (2018)*

7. Zhu Zhan, Bugao Xu. Analyzing review sentiments and product images by parallel deep nets for personalized recommendation. *Information Processing & Management, Volume 60, Issue 1,2023, 103166, ISSN 0306-4573*

8. Quoc Tuan Truong, Hady Wirawan Lauw. Visual sentiment analysis for review images with item-oriented and user-oriented CNN *(2017). MM '17: Proceedings of the ACM Multimedia Conference, Mountain View, CA, October 23-27. 1274-1282. Research Collection School Of Information Systems*

9. Abon Chaudhuri, Paolo Messina, Samrat Kokkula, Aditya Subramanian, Abhinandan Krishnan, Shreyansh Gandhi, Alessandro Magnani, Venkatesh Kandaswamy. A Smart System for Selection of Optimal Product Images in E-Commerce. *arXiv preprint arXiv:1811.07996v1 [cs.CV] 12 Nov 2018*

10. Amazon Review Dataset (2018). Retrieved from https://cseweb.ucsd.edu/~jmcauley/datasets/amazon_v2/