An Ensemble Recognition Algorithm for Duplicated Product Post on Shopee

Group21-Members

Name	Qinren Zhou	Lu Zhang	Yongrui Chen	Zhaoyuan Qiu	Zhibao Li
NetID	Qz142	Lz468	Yc910	Zq37	Zl424
Project GitHub	https://github.com/ZhaoyuanQiu/ANLY590.git				

Project Goal & Objective Summary

Our goal is to achieve an algorithm which can automatically find out the original post for a given product post, if any previous post pointing to the same item is found and returns FALSE if the product is posted for the first time.

This model can be used to help prevent duplicate post on e-shopping websites and thus improve product posting efficiency. This algorithm can also be used to improve recommending system and reduce duplicated recommendations for users.

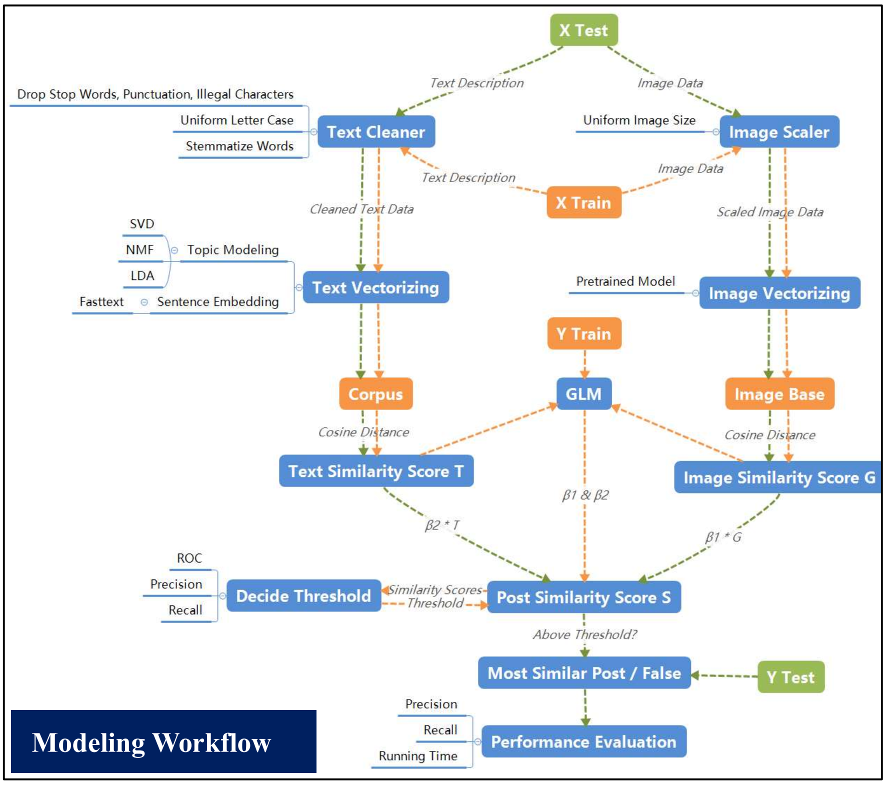
Proposed Data Source

The data used in this project was collected from Kaggle, which contains more than 30,000 product text descriptions and images from Shopee. Each post is identified with a unique ID. The goal is to find pairs of post ID which actually post for the same stuff.

Methods

To accurately predict the label for a given $post1$, we will define a similarity score $S(post1, post2)$ to measure the similarity between $post1$ and any other $post2$ in the dataset. Then we return the most similar post. Since the original data consists of image data and text data, the task can be divided into 2 parts. Calculating image similarity score $G(post1, post2)$ and calculating text similarity score $T(post1, post2)$. And then we combine the 2 metrics via ensemble method and get the final similarity score $S = \beta_1 G + \beta_2 T$.

To calculate score G, we can apply pre-trained model, which maps images to dense vectors. The similarity between two images can be obtained by calculating the cosine similarity between their embedded vectors. The calculate score T, first we can use Bag-of-Words or TF-IDF value to digitize the text data. Then we apply Latent Semantic Index (LSI) to compute the similarities between text data. We will compare different approaches including SVD, NMF and LDA.



The final stop is how to determine the weights β_1 and β_2. We can apply general linear regression with softmax to get predictions, then do gradient descent with $categorical\ entropy$ as loss function. The work scheme is as below,

Phase 1 Preparing Data	**Phase 2** Similarity Score Modeling	**Phase 3** Combining Metrics	**Phase 4** Test & Evaluation
Image Data Standardization	Transfer Pre-trained Model	GLM and Gradient Descent	Tuning Hyperparameters Writing Reports
Text Data Cleaning	Topic Models & LDA		

Expected Results

With the input of product post with text descriptions, our model is expected to identify if there are any repeated or similar posts. If there is a duplicated post, the model will return the IDs of the first n duplicate posts with the highest probability, otherwise return False.

Reference

[1] Common Signatures for Images | TensorFlow Hub. (n.d.). TensorFlow. Retrieved October 9, 2022, from https://www.tensorflow.org/hub/common_signatures/images

[2] Futrzynski, R. (n.d.). Image similarity with deep learning explained. Peltarion. Retrieved October 9, 2022, from https://peltarion.com/blog/data-science/image-similarity-explained

[3] Similar image: CNN + Cosine Similarity. (2022, February 17). Kaggle. Retrieved October 9, 2022, from https://www.kaggle.com/code/hamditarek/similar-image-cnn-cosine-similarity

[4] Wikipedia contributors. (2022, September 27). Topic model. Wikipedia. Retrieved October 10, 2022, from https://en.wikipedia.org/wiki/Topic_model

[5] Pagliardini, M., Gupta, P., & Jaggi, M. (2018). Unsupervised Learning of Sentence Embeddings Using Compositional n-Gram Features. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). https://doi.org/10.18653/v1/n18-1049