Adobe Behaviour Simulation Challenge: InterIIT Tech Meet 12.0

Team 99

Abstract

This study outlines our suggested solution for the Adobe Behaviour Simulation challenge. In the initial section, we present a solution for predicting likes. Our baseline model consists of a neural network with embedded inputs. Next, we implement clustering by employing K-Means Clustering to create clusters based on firm names. Subsequently, we train a small-scale classification model for each cluster to facilitate downstream tasks. In the second phase, we once again utilise K-Means Clustering. Subsequently, we evaluate the performance of GIT (Wang et al., 2022), VIT + GPT2, and BLIP (Li et al., 2022) on each individual cluster. Our findings indicate that BLIP offers superior generation quality with less inference time. Ultimately, we introduce our training methodology tailored to the limited timeframe of the competition, as well as a means to enhance this technique for improved outcomes in future endeavours.

1 Introduction

Producing content that is both timely and targeted is crucial in order to achieve the intended marketing outcomes for any firm. This is achieved through increased user engagement, which subsequently leads to more sales and a wider consumer base. Achieving high scores on key performance indicators (KPIs) related to user interaction enhances the legitimacy of a brand, expands its reach, and increases product recall. This also aids advertisers in enhancing algorithmic favorability, which is the process by which a social media firm determines which posts to promote based on the level of engagement they generate. Our suggested method aims to showcase the ability of Deep Neural Networks to mimic human behaviour by accurately predicting the likes and content of a tweet based on the provided metadata. Due to the intricate nature of human activity, it is extremely challenging to accurately represent it using a single mathematical tool. Therefore, we suggest implementing the approach of task downstreaming. Consequently, we develop distinct models to handle various categories of tweets, thereby diminishing the spectrum of human emotions and responses that a single model must address.

2 Methodology

2.1 Task 1: Likes Prediction

For this purpose, our starting point is a simple feed forward neural network that receives vectorized embeddings of the textual columns from all the training data. As anticipated, the extensive training data coverage across the domain and the limited number of layers in our model result in a significantly subpar performance and prolonged training time for this technique.

Therefore, in an attempt to downstream our data, we deduced from patterns observed in social media that there is a strong correlation between user engagement with a social media post and its relevance to the corresponding business. This strategy excludes numerous individual tweets, but we can consider them collectively.

For the purpose of putting this into reality, we have made the decision to cluster our data according to the inferred company. For the purpose of determining the optimal number of clusters, we employ the elbow technique, which ends up being somewhere about 55. A graph to support this can be seen in figure 1.

Now, for each of these tasks, we experimented with two methodologies:

- Preprocessing image into embeddings using MobileNetV2 (Sandler et al., 2019)
- Generating text from images using BLIP (Li et al., 2022) and using this text as an input to the regression model in vectorized form.

For each of the clusters, this is carried out in a separate manner. Consequently, we construct an independent regression model for each of the clusters. In the process of inference, we begin by determining the cluster that is closest to each point by utilising the Euclidean distance, and then we use that information as input to the model that is associated with that cluster. An example of a feed forward neural network that has been trained for regression is the model itself. It is possible to observe the full pipeline in the diagram of the pipeline that is depicted in figure 2.

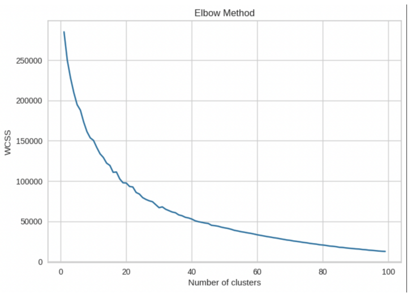

Figure 1: Elbow method for determininig the ideal number of clusters

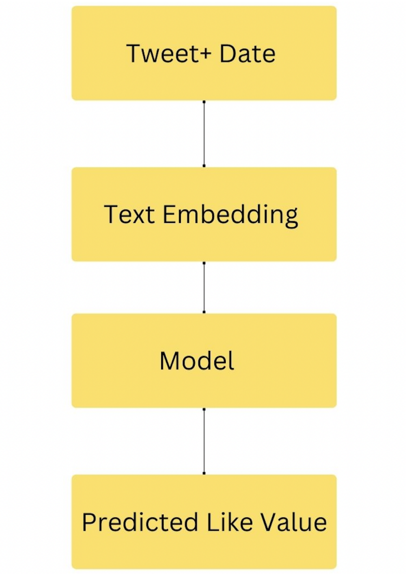

Figure 2: Pipeline for inference for Task-1

2.2 Task 2: Content Generation

As a result of the Exploratory Analysis of the data, we discovered that clustering can be implemented, and as a result, we discovered that the optimal number of clusters is close to 55.

Then, initially, we intended to create distinct models for each cluster, and then map the companies that were seen to their respective clusters, while mapping the companies that were not seen to the clusters that were closest to them based on the product domains that they offered. On the other hand, we were unable to apply this strategy because of size and computational constraints on our end.

After that, we changed our focus to the process of fine-tuning existing open-source LLMs, and since GPT2 offered the optimal balance of inference time and accuracy, we decided to go with it. We used the image captions, the date, the number of likes, and the content as features for the purpose of fine-tuning. We examined GIT (Wang et al., 2022), VIT (Vision Transformers)+Gpt2, and BLIP (Li et al., 2022) for the purpose of image captioning, and we discovered that BLIP-base had the optimal balance of modest inference time while also supplying the necessary features in the captions of the images and achieving the best results.

Given the required time and computational resources, we believe clustering of data for downstreaming the task of tweet generation can still provide us with better results that a generalized approach.

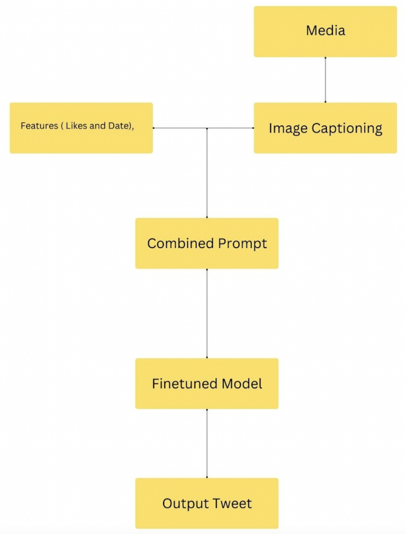

Figure 3: The pipeline utilized for tweet generation

As it can be seen in Figure 3, the Media from the given data is passed into an image captioning model, the output is then concatenated with the other textual features from the data set (likes, date etc.). This is combined into a prompt for tweet generation which is sent to the finetuned model as described above.

The finetuned model then outputs the predicted tweet. As of the current implementation, the model is cluster agnostic. We believe a cluster specific approach can provide better results but are impractical to implement given the computational resources required and the timeframe of the competition.

3 Training Scheme Employed

3.1 Current Approach

We were unable to provide independent training on each cluster due to the limited amount of time we had available. Rather than using it for inference, we used a model that had been pretrained on 25 data points from each cluster to ensure that it was accurate. Vectorization of textual data is accomplished with the help of a tf-idf vectorizer, and then the media is downloaded and embeddings are generated with the help of the MobileVNet2 model. After being trained on around 3000 points, the base model is then fine-tuned by using twenty-five data points chosen from each of the clusters.

3.2 Better Approach

Utilising a strategy that involves training different models for each cluster is a viable option once sufficient computational resources are available. It is possible to store these models. The Euclidean Distance can be used to assign a datapoint to a cluster during the inference process. This datapoint can then be used as input to the model that is most appropriate for the situation.

Similar approach can also be used for tweet generation where we downstream the area of content generation to a single cluster as opposed to the entire dataset and then generate the tweet following the earlier stated pipeline.

We were unable to run this approach due to a paucity of time and resources. This app

4 Future Work

This work has the potential to be expanded in the future in a variety of different ways. A superior clustering strategy would be the first option. We are able to include Psychographic Segmentation into our strategy in order to accomplish this goal. This strategy can be implemented for a particular time period, but it necessitates additional data from outside sources.

One further significant enhancement that might be made is to the design of the regressor, which is now a straightforward feed forward neural network configuration.

In the current implementation of the tweet generating process, BLIP is being utilised; however, it is possible to upgrade to more advanced models that have higher processing requirements.

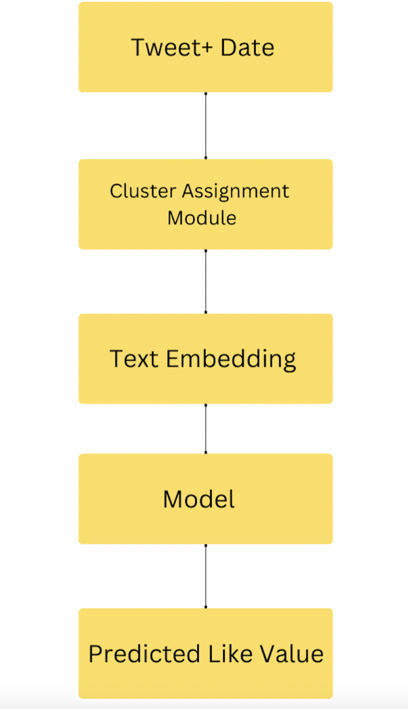

Figure 4: A diagrammatic representation of the Better Approach. Similar approach can be followed for tweet generation

5 Conclusion

Thus, we provide our proposed solution for both the subparts of the Problem Statement. We also propose methods to make our solution better with more computational resources.

The code of the solution can be found in the following git repository.

Link: https://github.com/immortql-coder/Adobe-InterIIT-Team-99

Note that the repo is private till the date of the presentation and will be made public after it.

References

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation.

Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. 2019. Mobilenetv2: Inverted residuals and linear bottlenecks.

Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. 2022. Git: A generative image-to-text transformer for vision and language.