# Multi-modal for News Popularity Prediction

311553054 王詠仟 311551098 吳泓緯 311551087 歐亭昀

## Abstract

This work investigates the prediction of news popularity by leveraging multi-modal techniques that combine textual content and accompanying images. This work explores the relationship between post content and its corresponding engagement metrics, focusing on the number of likes and comments. Three models are proposed: Baseline, Pre-CoFactv2, and CLIP+BERT, each utilizing different methods to learn joint representations. The models are evaluated on a comprehensive dataset of news articles, and their performance is measured using Mean Absolute Error (MAE). The results demonstrate the superiority of learning joint representations through contrastive pretraining on a large dataset, as showcased by the CLIP model. The research highlights the importance of incorporating both textual and visual information for accurate news popularity prediction in today's digital landscape.

## Introduction

In the era of social media, accurately predicting the reach and engagement of posts has become a captivating research endeavor. This work delves into the relationship between post content and its corresponding engagement metrics by employing multi-modal techniques that leverage textual content and accompanying images. Initially, the aim was to predict the number of likes on Instagram posts, but due to restrictions imposed by Instagram's crawling policy, the target shifted to predicting the number of likes on LINE TODAY news. To enhance prediction accuracy, a multi-modal framework is designed, incorporating various model architectures. The study validates the existence of a certain level of correlation between the content of an article and its corresponding engagement metrics.

The problem addressed in this work is to develop a model that predicts the number of likes and comments based on the combination of news images and news texts. The dataset consists of news articles from LINE TODAY, where each instance represents a news article with attributes such as the image, text, timestamp, and the target variable, which is the total number of likes and comments.

**Previous Studies**

The prediction of news popularity using multi-modal information has garnered significant attention. This task involves integrating textual, visual, and temporal features to assess the popularity of news articles. Drawing inspiration from previous works such as VisualBERT, ViLBERT, and UNITER, which contribute to multi-modal models in vision and language tasks, this research proposes a novel multi-modal model for news popularity prediction. By incorporating textual content, image features, and temporal information, the model captures the factors influencing news popularity, presenting an effective solution for this task.

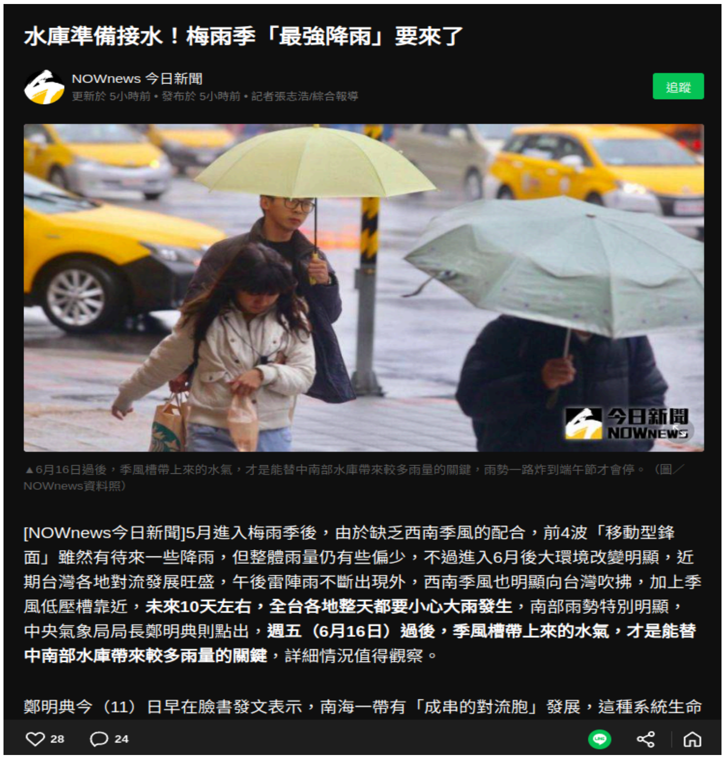
**Approaches**

This work introduces three models for news popularity prediction: Baseline, Pre-CoFactv2, and CLIP. The Baseline model concatenates textual and visual features extracted from Pretrained BERT and Pretrained ResNet-50, respectively, while the Pre-CoFactv2 model incorporates cross-modal and cross-type relations using parameter-efficient foundation models. The CLIP+BERT model combines the strengths of BERT, CLIP-image-encoder, and CLIP-text-encoder to obtain better semantic features for multi-modal problems. Experimental evaluation is conducted using a comprehensive dataset of crawled news articles from the LINE TODAY news website. The dataset undergoes rigorous preprocessing to ensure data quality and consistency. The performance of the models is assessed using Mean Absolute Error (MAE), and the results show that learning joint representations through contrastive pretraining on a large dataset, as demonstrated by CLIP, leads to superior performance.

In conclusion, this research contributes to the prediction of news popularity by exploring the relationship between news content and engagement metrics. By incorporating multi-modal information, including textual and visual cues, the proposed models provide more accurate predictions. The findings highlight the significance of considering both textual and visual content in today's digital landscape, bridging the gap left by traditional text-based approaches.

## Problem Definition

The problem addressed in this work is to develop a model that predicts the number of likes and comments based on the combination of news images and news texts. Denote the dataset as $\chi = (I, T, t, y)$ , every instance in $\chi$ is a news from LINE TODAY, $I$ is the image, $T$ is the text in the post, $t$ is the timestamp which indicates the publish time of the post, $y$ is the target in our task ,which is the total of likes and comments.



水庫準備接水！梅雨季「最強降雨」要來了

NOWnews 今日新聞
更新於 5小時前 • 發布於 5小時前 • 記者張志浩/綜合報導

▲6月16日過後，季風槽帶上來的水氣，才是能替中南部水庫帶來較多雨量的關鍵，雨勢一路炸到端午節才會停。（圖／NOWnews資料照）

[NOWnews今日新聞]5月進入梅雨季後，由於缺乏西南季風的配合，前4波「移動型鋒面」雖然有待來一些降雨，但整體雨量仍有些偏少，不過進入6月後大環境改變明顯，近期台灣各地對流發展旺盛，午後雷陣雨不斷出現外，西南季風也明顯向台灣吹拂，加上季風低壓槽靠近，**未來10天左右，全台各地整天都要小心大雨發生**，南部雨勢特別明顯，中央氣象局局長鄭明典則點出，**週五（6月16日）過後，季風槽帶上來的水氣，才是能替中南部水庫帶來較多雨量的關鍵**，詳細情況值得觀察。

鄭明典今（11）日早在臉書發文表示，南海一帶有「成串的對流胞」發展，這種系統生命

***Figure 1.*** Raw data of LINE TODAY news

## Related Work

The prediction of news popularity using multi-modal information has received significant attention. This task involves assessing news article popularity by integrating textual, visual, and temporal features. We review relevant literature that contributes to multi-modal models in vision and language tasks, inspiring our approach.

Lu et al. (2019) introduced VisualBERT, a transformer-based model that effectively integrates visual and textual modalities. It achieves state-of-the-art performance in vision and language tasks. ViLBERT (Lu et al., 2019) extends VisualBERT by learning task-agnostic visio linguistic representations. Through joint modeling of visual objects, regions, and text, it excels in vision-and-language benchmarks. UNITER (Chen et al., 2020) proposes a universal model for image-text representation learning, achieving outstanding results.

In our research, we draw inspiration from VisualBERT, ViLBERT, and UNITER to develop a novel multi-modal model for news popularity prediction. By incorporating textual content, image features, and temporal information, our model captures the factors influencing news popularity. Leveraging multimodal learning, our approach presents an effective solution for this task.

In summary, VisualBERT, ViLBERT, and UNITER contribute to multi-modal models but not specifically for news popularity prediction. Our research fills this gap by proposing a dedicated multi-modal model that predicts news popularity, building upon these foundational works.

## Method

In our study, we propose three models for news popularity prediction: Baseline, Pre-CoFactv2, and CLIP+BERT. These models progress from simple to more complex architectures.

### Baseline

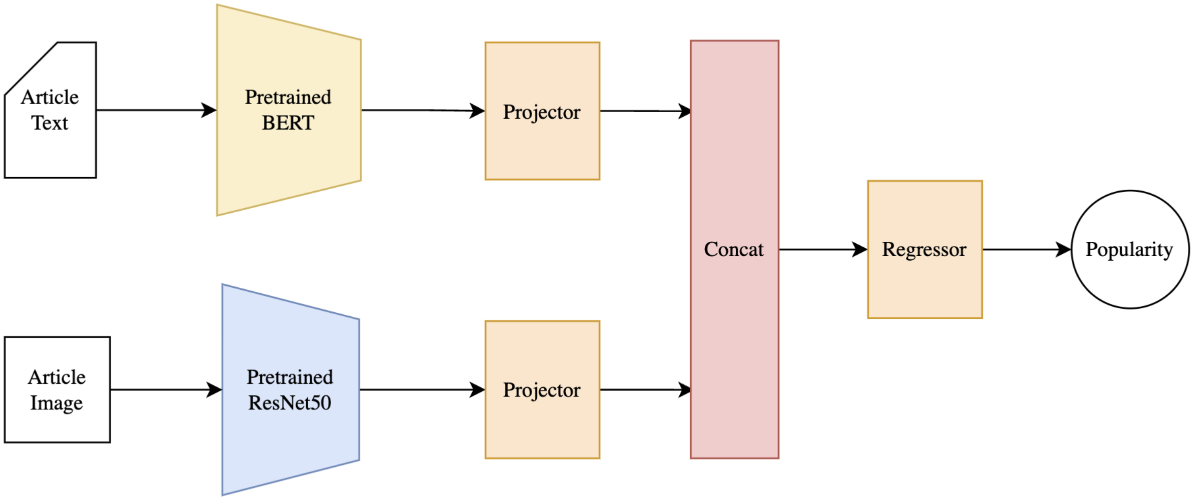


***Figure 2*** Baseline Model Architecture. It combines Pretrained BERT and Pretrained ResNet-50 to extract 512-dimensional embeddings from news text and images, respectively. These features are concatenated to form a 1024-dimensional vector. The MLP captures relationships between the combined features and news popularity levels.

In the Baseline model, we utilize Pretrained BERT to obtain a 512-dimensional embedding from the news text, and Pretrained ResNet-50 to extract a 512-dimensional embedding from the corresponding news images. The textual and visual features are then concatenated to form a single 1024-dimensional feature vector. The overall architecture of the model is depicted in ***Figure 2***.

During the training of the Baseline model, we fix the parameters of Pretrained BERT and Pretrained ResNet-50 and only train the MLP. We employ the AdamW optimizer with a learning rate of 5e-4 and weight decay of 1e-5.

The Baseline model takes advantage of the expressive power of Pretrained BERT and Pretrained ResNet-50 to capture meaningful representations from news text and images, respectively. By concatenating the 512-dimensional embeddings, we create a joint representation that combines textual and visual cues for news popularity prediction. The MLP is trained to learn the relationships between the combined features and the popularity levels of the news articles.

### Pre-CoFactv2

***Figure 3*** illustrates an overview of the proposed Pre-CoFactv2 framework. The additional features are generated by the feature extractor from the given claim text, claim image, document text, and document image. Then, we adopt two parameter-efficient foundation models for learning in-domain knowledge from pre-trained embeddings with adapters and a multi-modal multi-type fusion module for modeling not only cross-modality (i.e., text and image) relations but also cross-type (i.e., claim and document) relations. Outputs of these embeddings are fused by the verification classifier with cross-entropy loss as well as supervised contrastive loss to separate embeddings and find clearer boundaries.

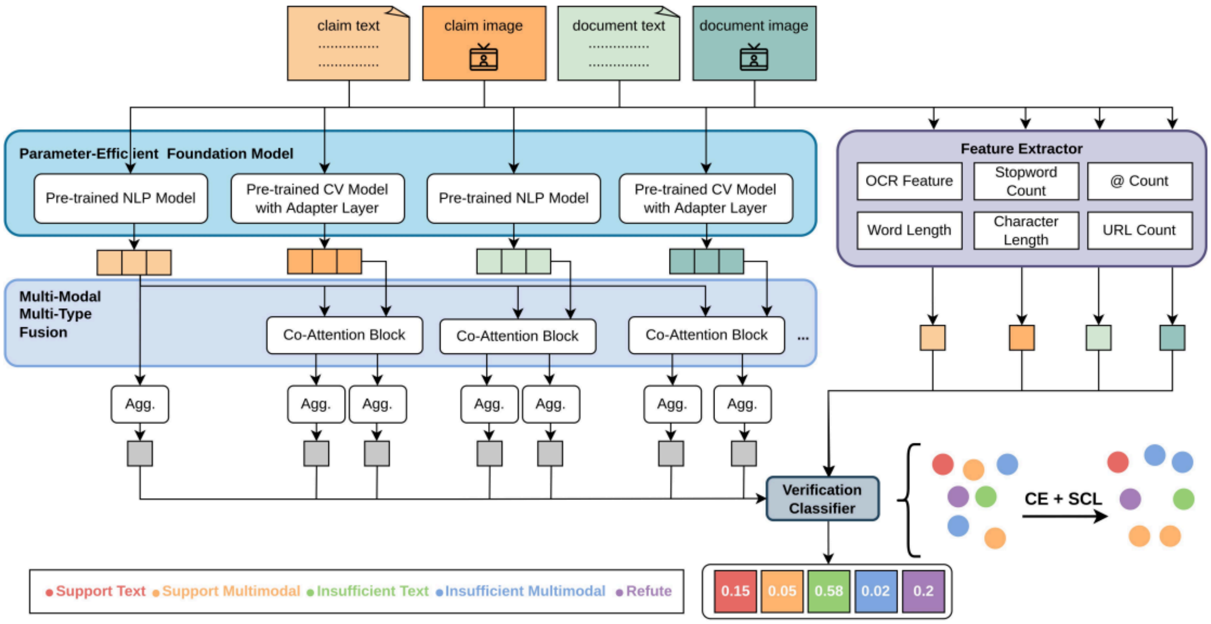


***Figure 3.*** Illustration of the Pre-CoFactv2 framework. The parameter-efficient foundation model aims to transform the input text and images into embedding by the pre-trained language model. Then, the multi-modal multi-type fusion fuses the information from the same modality (images/text from the claim and document), different modalities (images and text from the claim/document), and different types (image from the claim and text from the document, and text from the claim and image from the document) to obtain contexts. Besides, the feature extractor is designed to convert input text and images into several features. In the end, the verification classifier contains cross-entropy loss and contrastive loss to predict the possible class based on the embeddings from previous outputs.

## CLIP+BERT

In CLIP+Bert model, we decided to use three types of encoders to obtain better semantic features for the multi-modal problems: (1) **Bert model** which is pretrained on the Chinese Wikipedia text and fine-tuned on our LINE TODAY News dataset. (2) **CLIP-image-encoder** (3) **CLIP-text-encoder.** CLIP is composed of an image encoder and a text encoder, both jointly pre-trained to project the image and the caption onto the same embedding space in a contrastive manner. In this way, the extracted image embeddings and the caption embeddings are aligned, and the images will be near the captions with similar semantic features. We adopt ViT as the CLIP-image-encoder and Bert as the CLIP-text-encoder in our CLIP+Bert model.

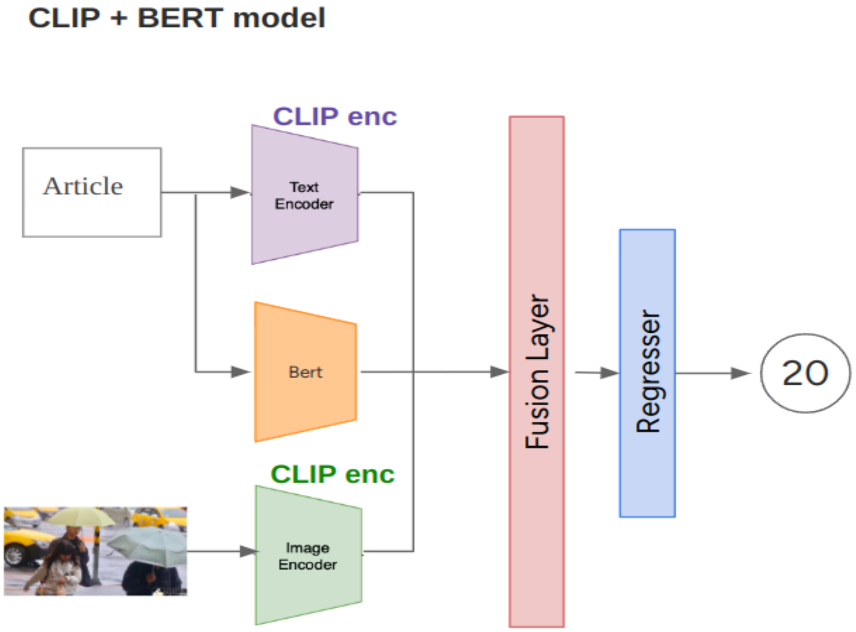


*Figure 4.* Illustration of the CLIP+BERT model. We concatenate the output of CLIP-text-encoder, CLIP-image-encoder and Bert to predict the likes and comments of the news.

Therefore, we get two text embeddings separately from the CLIP-text-encoder and bert model and image embedding from CLIP-image-encoder. We regard the connection of three embeddings as the input of the fusion layer , then the fusing embedding is fed to the regressor with the multi-modal embedding to predict the corresponding numbers of popularity.

**Loss function**

Mean Absolute Error (MAE) is a measure of the average size of the mistakes in a collection of predictions, without taking their direction into account. It is measured as the average absolute difference between the predicted values and the actual values and is used to assess the effectiveness of a regression model. Thus, we use MAE as the loss function for all the models we proposed. The following shows the formula we use:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|f_i - y_i|$$

# Experiment

## Dataset

| Attribute | Value |
|---|---|
| Title | ‘桃竹深夜淹水警報！暴雨像整桶灌\u3000網友:騎車好像在划船’ |
| Content | '中央氣象局今晚（22）日9時30分已經針對苗栗以北7縣市發布「大雨特報」，並於10時31分在竹桃發布「大雷雨即時訊息」。水利署10時40分許指出，中壢站時雨量達40.5毫米、...’ |
| Publish Time | 300 |
| Image URL | 'https://obs.line-scdn.net/0ht3EVz10qKxx4LzkQ-YFUS0B5J21LSTEVWk1nLV58dSlTA2sfQEB4f1l6dTBdTD5DWBpgLQ18dHtQGWwZFA/w644' |
| Like Count | 244 |
| Comment Count | 250 |

***Table 1.*** Columns of dataset and a example of one instance

We constructed a comprehensive dataset by crawling news articles from the LINE TODAY news website using the Selenium automation tool. The dataset comprises 8,908 news articles, covering diverse topics such as "Entertainment," "Domestic," "International," "Sports," and "Finance." This careful selection ensures the dataset's diversity and represents various domains for accurate news popularity prediction. Each article in the dataset includes attributes such as the title, content, time elapsed since publication, image link, number of likes, and number of comments. Rigorous preprocessing techniques were applied to clean the text and standardize the image sizes, ensuring data quality and consistency.

The dataset's curation involved systematic crawling, targeted topic selection, and preprocessing steps to optimize its suitability for our research objectives. This curated dataset provides a valuable resource for training and evaluating our proposed models, enabling a comprehensive exploration of news popularity prediction across multiple domains.

### Testing Performance

CLIP, PreCoFact, and Baseline are all multi-modal models designed to learn joint representations, albeit through different methods. CLIP employs contrastive pretraining on a large-scale multi-modal dataset, PreCoFact utilizes cross-attention, and Baseline simply concatenates features. By performance of CLIP ,we can find that learning joint representations through contrastive pretraining on a large dataset leads to superior results.

| Model Name | MAE | Hit 5 |
|---|---|---|
| Baseline | 13.45 | 77.02 |
| CLIP | 13.10 | 79.65 |
| Pre-CoFactv2 | 13.80 | 76.11 |

***Table 2.*** Performance of our two methods and baseline model.

### Ablation

The baseline model is composed of image encoder and text encoder, while BERT and Sentence (T5) only consider the text information. We can get two conclusions from ***Table 2,***

- The Image information didn’t really much.
- The performance slightly improves when using larger pretrained language models such as T5 compared to BERT.

| Model Name | MAE | Hit 5 |
|---|---|---|
| BERT | 13.52 | 76.89 |
| Sentence-T5 | 13.30 | 79.36 |
| Baseline | 13.45 | 77.02 |

***Table 3.*** Performance with only text information compared to our baseline (with both image and text information.)

## Conclusion

In conclusion, we tackled the task of predicting news popularity by implementing three multi-modal combining text and image models. Through our research, we aim to explore the relationship between news content and the number of likes and comments  by utilizing visual and textual information.  In our experiment, we recognized the limitations of traditional approaches that focused solely on text-based analysis, overlooking the significant impact of visual content in today's digital landscape. Our method addresses the prediction gap by combining news video and text to create a more comprehensive and accurate prediction model, and slightly improves on the evaluation of Hit 5.

### Team Member Contributions

Data Crawling, Related work, baseline Model: 吳泓緯

CLIP+BERT model, video: 歐亭昀

Pre-CoFactv2, video: 王詠仟

# References

Tang, Yu-Chien, et al. "NYCU-TWO at Memotion 3: Good Foundation, Good Teacher, then you have Good Meme Analysis." *arXiv preprint arXiv:2302.06078* (2023).

Radford, Alec, et al. "Learning transferable visual models from natural language supervision." International conference on machine learning. PMLR, 2021.

Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).

Dosovitskiy, Alexey, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." *arXiv preprint arXiv:2010.11929* (2020).

Ni, Jianmo, et al. "Sentence-t5: Scalable sentence encoders from pre-trained text-to-text models." *arXiv preprint arXiv:2108.08877* (2021).

Li, Liunian Harold, et al. "Visualbert: A simple and performant baseline for vision and language." arXiv preprint arXiv:1908.03557 (2019).

Lu, Jiasen, et al. "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks." Advances in neural information processing systems 32 (2019).

Chen, Yen-Chun, et al. "Uniter: Universal image-text representation learning." Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX. Cham: Springer International Publishing, 2020.