Inter IIT Tech Meet 12.0 - Adobe

Team 31

Abstract

This challenge involves two tasks: behavior simulation and content simulation, utilizing exploratory data analysis, feature engineering, and advanced machine learning models. The first task predicts user engagement, measured by likes, using **XGBoost**. The second task generates tweet text using media embedding and a **Mistral-7B** model for content simulation. Additionally, we have devised custom approaches to generate content. The goal is to assist marketers in optimizing user engagement and content creation on social media. [1]

Task 1

Behaviour Simulation based on tweet content.

Data Engineering

This section explains the data treatment process and how it was prepared for model training.

- Analysis

Upon examination, it was observed that the data exhibited a noticeable leftward skew, suggesting that the majority of 'likes' were concentrated at the lower end. Specifically, 94% of the total data had 'likes' below 20,000. This bias posed various challenges, including the initial difficulty of the model in understanding the higher range of 'likes' due to the scarcity of data points in that range.

- Username & Inferred Company

The technique of one-hot encoding was employed for both username and deduced business, resulting in a significant correlation with 'likes'. High-ranking usernames were assigned a unique one-hot encoding, while less popular usernames/companies shared the same encoding category. This approach enhanced the model's understanding of the data. A similar approach was adopted for deduced business as well.

[1] https://github.com/team-31-interiit/Adobe_31.git

-Other Quantitative Features

These features were obtained after plotting various correlations between the data . Refer Table 1.

FEATURES	DESCRIPTION
Languages	Language used in tweets
Thumbnail Status	Presence of a video thumbnail
URL Status	Presence of a URL in media
Video Bitrate	Bits transmitted per unit time
Video Duration	Duration of the video
Day	Day of the week
Is Famous	Likes surpassing a threshold
Number of Views	Views for a video
Number of Posts	Total posts by a user
Number of Posts (Past)	Posts by a user until time t
Number of Hashtags	Hashtags featured in a post
Fraction of Caps	Fraction of capital letters in a post
Number of Mentions	Mentions featured in a post
Number of Words	Words in a post
Aspect Ratio	Ratio of height to width
Emoji Count	Count of emojis in a post
Image Brightness	Brightness of the image
Saturation	Colorfulness of the image
Dominant Color	Maximum featured color
Image Size	Dimensions of the media

Figure 1: List of Features

Model Selection

Given a step-by-step brief overview of the approaches undertaken by the team.

- RoBERTa + ViT

To include the required media features, the application of a vision transformer appeared to be a practical approach. The architecture comprised a ***Vision Transformer*** that was used to make a feature space for media. This was then merged with the embedding space attained from RoBERTa using a custom *cross attention layer*. Following this, the data was sent to a feed-forward neural network to create predictions.

However we observed oscillation in the training curve due to outliers in the data. We used techniques like gradient clipping, gradient accumulation but the model was not learning the outliers in the data

- Binning + Classifier + Multiple Regressors

Due to the data being skewed and the limitation of not being allowed to gather more information, it was important to understand the data distribution in segments. The ideal solution we found was to divide the data into four *bins* based on the number of likes it received.

We then needed a classifier to predict which bin a data point belonged to before using the corresponding model to estimate the number of likes.

The classification process involved three stages: the first classified 'likes' less than 100 and more than 100, the second classified 'likes' fewer than 3.5K and more than 3.5K, and finally, the last stage separated 'likes' fewer than 10K and above.

For each of these categories, we used an independent model called ***BERTweet***, specifically trained on the range of data points that fell into the respective category, using the content and all the features noted in the table. For observations refer appendix.

- Final Model

In the construction of our model, an XGBoost classifier is employed for training, utilizing Exploratory Data Analysis (EDA) features as input to generate an outlier score. This outlier score is then concatenated with the EDA features and fed into an XGBoost Regressor. Simultaneously, features such as username, inferred company, and other textual characteristics undergo processing through a linear regressor layer. The outputs from both models are amalgamated and subsequently forwarded through an additional regressor layer, culminating in the formation of the ultimate meta model.

Results

Model Name	Range	RMSE	R2 Score
Bertweet	0-100	**11.8**	**0.84**
Bertweet not Pre-rocessed	100-3500	508	0.67
Bertweet Pre-processed	100-3500	408	0.71
Bertweet Final	0-∞	5009	0.02
XG-Boost Ensembled	0-∞	3880	0.15

Table 1: Model Evaluation Metrics

Task 2

Content generation conditioned on behaviour.

Media Analysis

We employed various models to process media, generating informative captions and embeddings. In the prompt engineering process, captions were initially added to the prompts for the Language Model (LLM). Later, the embeddings were used to calculate similarity scores, identifying tweets with similar content. These tweets were later appended to the prompts.

- Image Processing

We initially used EVA-CLIP to obtain image embeddings with dimensions (1,512). However, to maintain consistency across modalities, we transitioned to clip-ViT-B-32, which efficiently produced embeddings for images.
Additionally, we employed the ***vit-gpt2*** model to generate captions, providing supplementary information for the prompts.

- OCR

We harnessed Paddle OCR to extract text from images. For video processing, we utilized a two-step approach: Katna for keyframe extraction followed by Paddle OCR for text extraction from the keyframes.

- Video Processing

Our initial approach to video processing involved a custom model combining CLIP and LSTM. However, we found that Language Bind provided more consistent and efficient video embeddings. For the captioning process, we leveraged the power of the TimesFormer-GPT2 model. This model effectively transformed our inputs into tweet text, enhancing the overall content simulation.

- Audio Processing

We utilized the open-source model, NeMo, to convert audio into text. Subsequently, Language Bind was used to transform the text into audio embeddings. This comprehensive approach ensured uniformity in the dimensions of the embeddings across all modalities.

Model Selection

This section outlines our data processing pipeline. We divide the data into analogical and fine-tuning

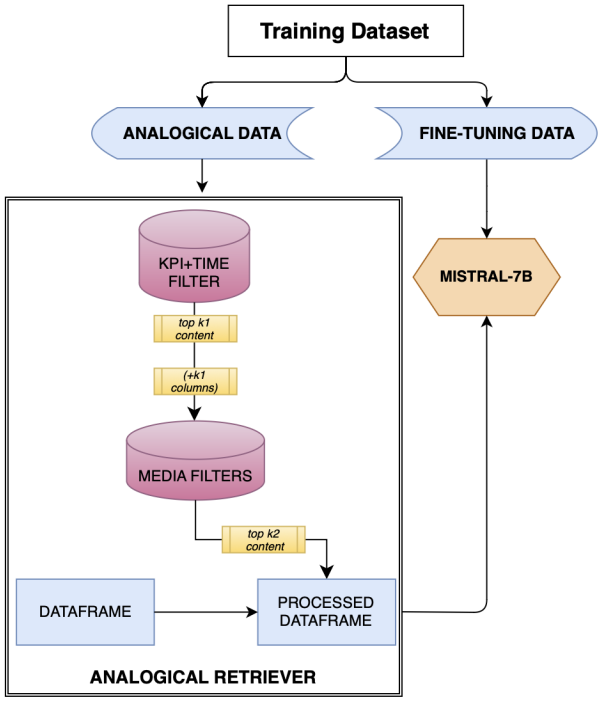

Figure 2: Training Pipeline

datasets with a split of 3:20 to minimize processing costs. We add k_2 additional data columns to each row of the analogical data by finding similar datapoints. We then finetune the LLM based on both data.

-Custom Approaches

We deployed two distinct models, FLAN T5 and Falcon7b, to achieve our goals. The FLAN T5 model utilized a variety of inputs, including image and video captions, OCR, and metadata like username, likes, company, date, time, and day. These inputs were then transformed into tweet outputs. Using the Falcon7b pipeline, we designed a custom model that incorporated modality tokens into the extracted features. To ensure compatibility with Falcon's requirements, we introduced linear layers and Conv1D to match the dimensions of the input features.

- Analogical Retriever

We create media content and embeddings for the analogical data using the ViT-B-32 model, as previously mentioned. Each row is filtered based on the inferred company, prioritizing companies. Tweets are sorted by date and likes (KPI+TIME FILTER), with the top k_1 tweets selected. These top k_1 tweets are further sorted based on similarity scores from the generated embeddings (MEDIA FILTERS). The final selection includes the top k_2 tweets, which are then added to each datapoint.

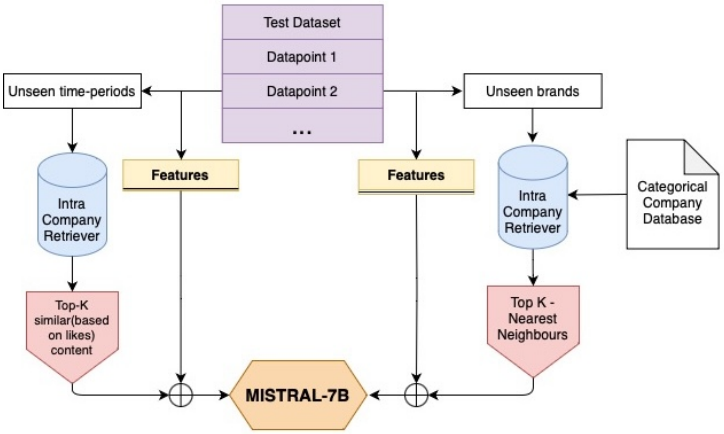

Figure 3: Inference Pipeline

Inference Pipeline

The test dataset can be of two types: unseen time periods and unseen brands. For a specific data point, we introduce additional columns of data that resemble the characteristics of that data point. This approach assists the model by providing more contextual information during inference.

- Unseen Time-periods

For each data point, we fetch tweets from the inferred company, selecting the top K tweets based on the number of likes. These top K tweets are then added to the data row. Additionally, all features of this data point are combined with the top K tweets and fed into the LLM.

- Unseen Brands

We iterate through all data rows and extract companies that are similar to the inferred company. Further more, we extract tweets from companies similar to the inferred company in each row. The tweets are sorted by likes, and the top K tweets are selected and appended to all datapoint features.

- Categorical Company Database

We use the list of companies from the training dataset and classify them based on the work they do. For instance, tech companies like Apple, Samsung, Microsoft and Adobe would belong to the same bin.

References

Marco Arazzi, Marco Cotogni, Antonino Nocera, and Luca Virgili. 2023. Predicting tweet engagement with graph neural networks. In *Proceedings of the 2023 ACM International Conference on Multimedia Retrieval*, ICMR '23. ACM.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning.

Michał Daniluk, Jacek Dąbrowski, Barbara Rychalska, and Konrad Gołuchowski. 2021. Synerise at recsys 2021: Twitter user engagement prediction with a fast neural model.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale.

Bin Guo, Hao Wang, Yasan Ding, Wei Wu, Shaoyang Hao, Yueqi Sun, and Zhiwen Yu. 2020. Conditional text generation for harmonious human-machine interaction.

Jiaming Han, Kaixiong Gong, Yiyuan Zhang, Jiaqi Wang, Kaipeng Zhang, Dahua Lin, Yu Qiao, Peng Gao, and Xiangyu Yue. 2023. Onellm: One framework to align all modalities with language.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Wang-Cheng Kang, Jianmo Ni, Nikhil Mehta, Maheswaran Sathiamoorthy, Lichan Hong, Ed Chi, and Derek Zhiyuan Cheng. 2023. Do llms understand user preferences? evaluating llms on user rating prediction.

Ankur Kumar. 2022. The illustrated image captioning using transformers. *ankur3107.github.io*.

Valerii Likhosherstov, Anurag Arnab, Krzysztof Choromanski, Mario Lucic, Yi Tay, Adrian Weller, and Mostafa Dehghani. 2021. Polyvit: Co-training vision transformers on images, videos and audio.

Hsien-Chin Lin, Christian Geishauser, Shutong Feng, Nurul Lubis, Carel van Niekerk, Michael Heck, and Milica Gašić. 2022. Gentus: Simulating user behaviour and language in task-oriented dialogues with generative transformers.

Yan-Bo Lin, Yi-Lin Sung, Jie Lei, Mohit Bansal, and Gedas Bertasius. 2023. Vision transformers are parameter-efficient audio-visual learners.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Mengmeng Ma, Jian Ren, Long Zhao, Davide Testuggine, and Xi Peng. 2022. Are multimodal transformers robust to missing modality?

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. Bertweet: A pre-trained language model for english tweets.

AJ Piergiovanni, Weicheng Kuo, and Anelia Angelova. 2022. Rethinking video vits: Sparse video tubes for joint image and video learning.

Rikaz Rameez, Hossein A. Rahmani, and Emine Yilmaz. 2022. Viralbert: A user focused bert-based approach to virality prediction.

Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gomez Colmenarejo, Alexander Novikov, Gabriel Barth-Maron, Mai Gimenez, Yury Sulsky, Jackie Kay, Jost Tobias Springenberg, Tom Eccles, Jake Bruce, Ali Razavi, Ashley Edwards, Nicolas Heess, Yutian Chen, Raia Hadsell, Oriol Vinyals, Mahyar Bordbar, and Nando de Freitas. 2022. A generalist agent.

Gaurav Sahu and Olga Vechtomova. 2021. Adaptive fusion techniques for multimodal data.

Muhammad Bilal Shaikh, Douglas Chai, Syed Mohammed Shamsul Islam, and Naveed Akhtar. 2023. Maivar-t: Multimodal audio-image and video action recognizer using transformers.

Nina Shvetsova, Brian Chen, Andrew Rouditchenko, Samuel Thomas, Brian Kingsbury, Rogerio Feris, David Harwath, James Glass, and Hilde Kuehne. 2022. Everything at once – multi-modal fusion transformer for video retrieval.

Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. 2019. Videobert: A joint model for video and language representation learning.

Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. 2023a. Eva-clip: Improved training techniques for clip at scale.

Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. 2023b. Generative pretraining in multimodality.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Haoqin Tu, Zhongliang Yang, Jinshuai Yang, Siyu Zhang, and Yongfeng Huang. 2022. Pcae: A framework of plug-in conditional auto-encoder for controllable text generation. *Knowledge-Based Systems*, 256:109766.

Subhashini Venugopalan, Marcus Rohrbach, Jeff Donahue, Raymond Mooney, Trevor Darrell, and Kate Saenko. 2015. Sequence to sequence – video to text.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. 2023. Cogvlm: Visual expert for pretrained language models.

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. 2023a. Next-gpt: Any-to-any multimodal llm.

Xuansheng Wu, Wenlin Yao, Jianshu Chen, Xiaoman Pan, Xiaoyang Wang, Ninghao Liu, and Dong Yu. 2023b. From language modeling to instruction following: Understanding the behavior shift in llms after instruction tuning.

Zihui Xue and Radu Marculescu. 2023. Dynamic multimodal fusion.

Shen Yan, Tao Zhu, Zirui Wang, Yuan Cao, Mi Zhang, Soham Ghosh, Yonghui Wu, and Jiahui Yu. 2023. Videococa: Video-text modeling with zero-shot transfer from contrastive captioners.

Michihiro Yasunaga, Xinyun Chen, Yujia Li, Panupong Pasupat, Jure Leskovec, Percy Liang, Ed H. Chi, and Denny Zhou. 2023. Large language models as analogical reasoners.

Lijun Yu, José Lezama, Nitesh B. Gundavarapu, Luca Versari, Kihyuk Sohn, David Minnen, Yong Cheng, Agrim Gupta, Xiuye Gu, Alexander G. Hauptmann, Boqing Gong, Ming-Hsuan Yang, Irfan Essa, David A. Ross, and Lu Jiang. 2023. Language model beats diffusion – tokenizer is key to visual generation.

Zhou Yu, Yuhao Cui, Jun Yu, Dacheng Tao, and Qi Tian. 2019. Multimodal unified attention networks for vision-and-language interactions.

Yue Zhao, Ishan Misra, Philipp Krähenbühl, and Rohit Girdhar. 2022. Learning video representations from large language models.

Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, HongFa Wang, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, Wancai Zhang, Zhifeng Li, Wei Liu, and Li Yuan. 2023. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment.

A Appendix: Images and Tables

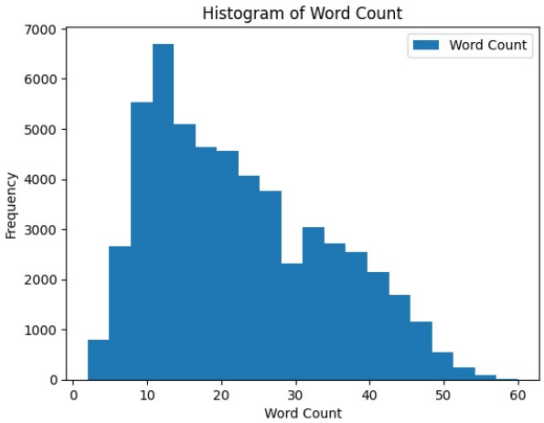

Figure 4: Frequency of words vs. Word count

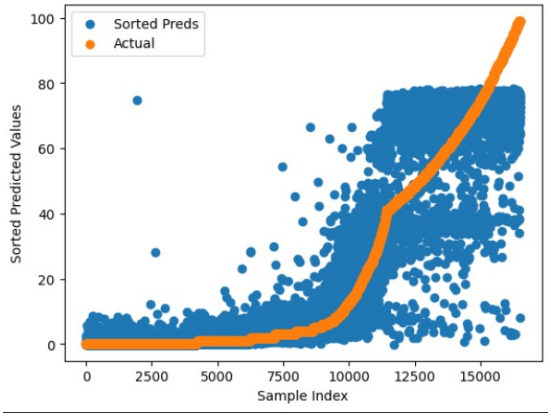

Figure 5: Predictions and Actual labels of likes in the range of 0-100

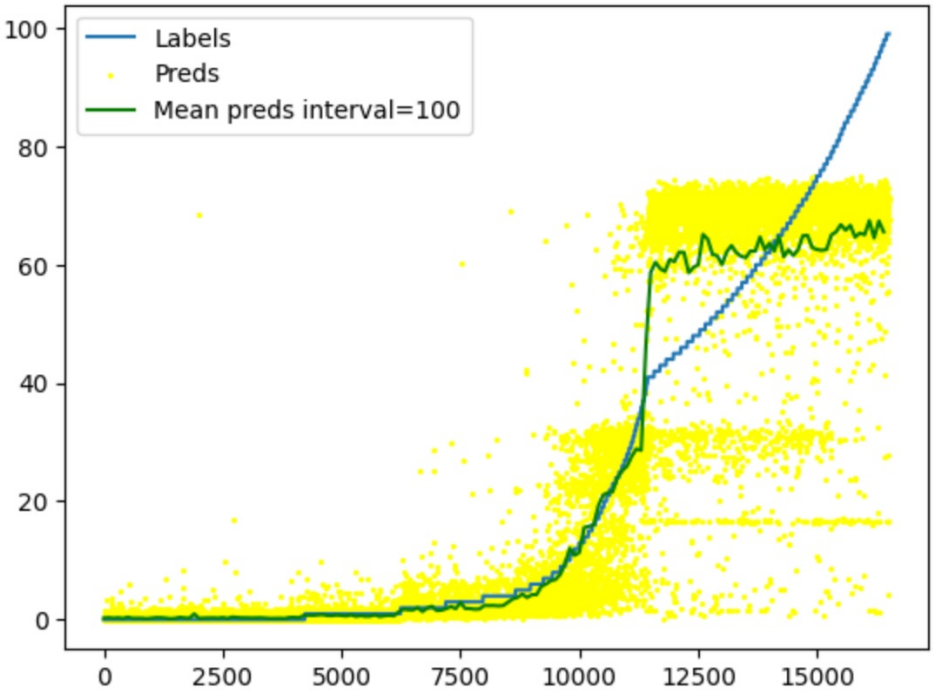

Figure 6: Predictions plotted alongside mean of predictions in bins of size 100 points across points whose likes lie in the range of 0-100

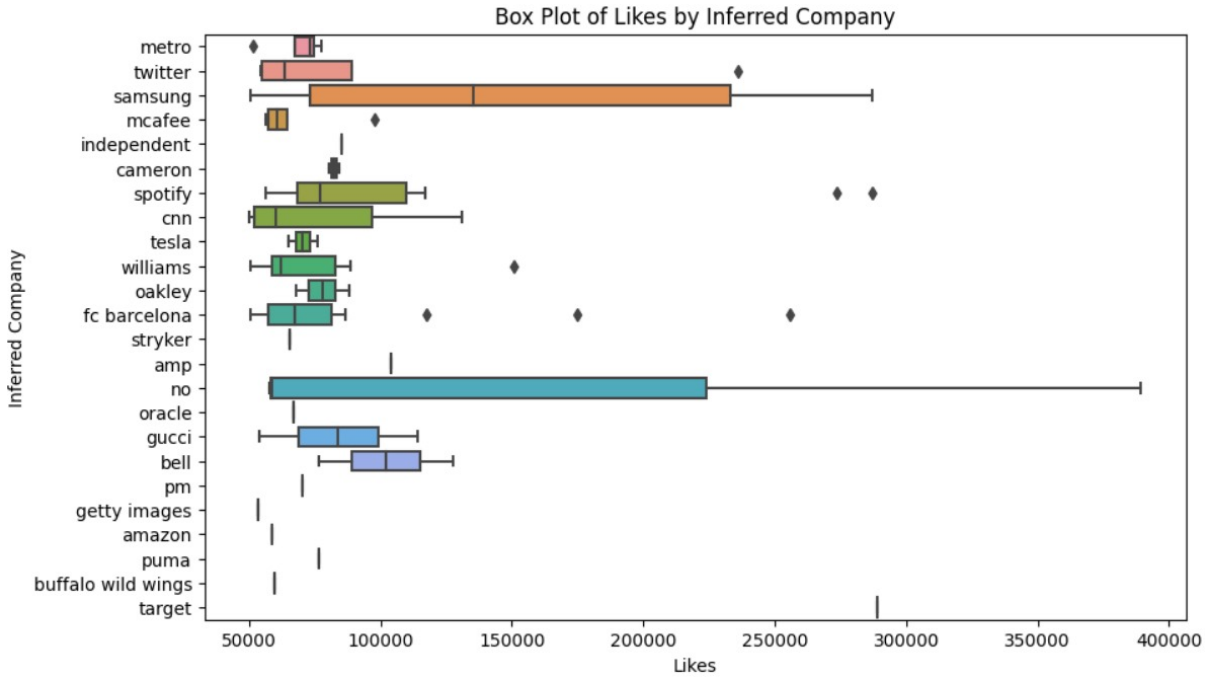

Figure 7: Box Plot for companies with viral posts

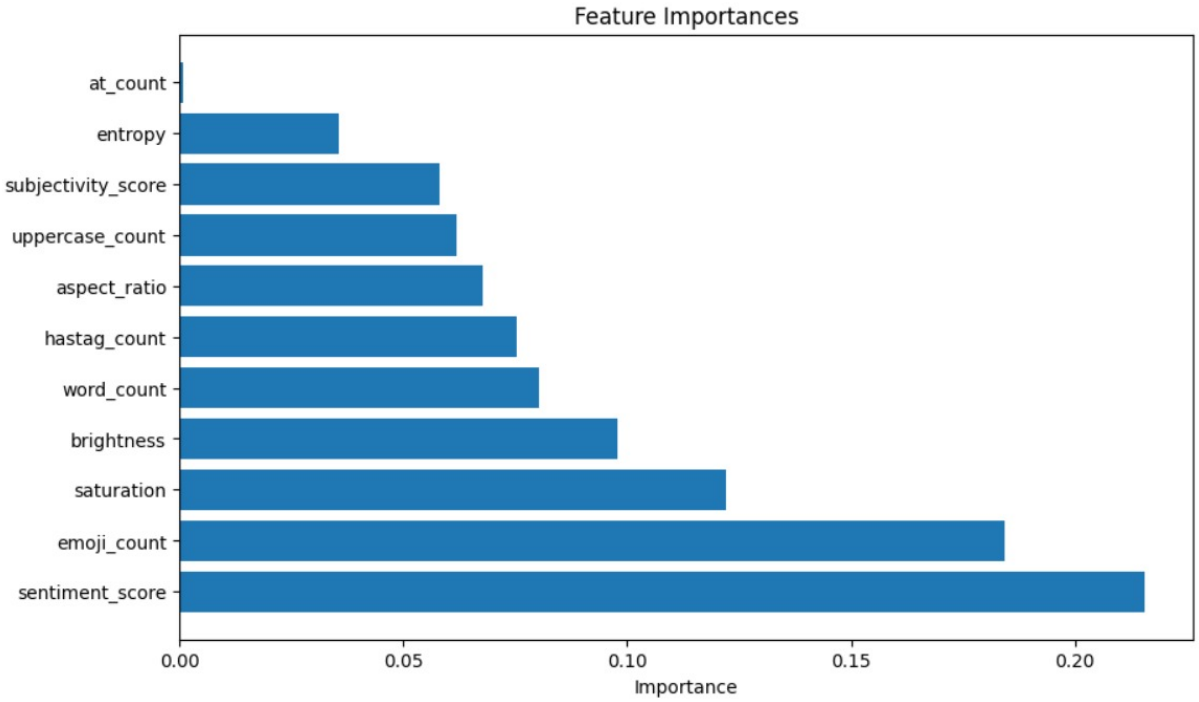

Figure 8: Analysis of feature's weights

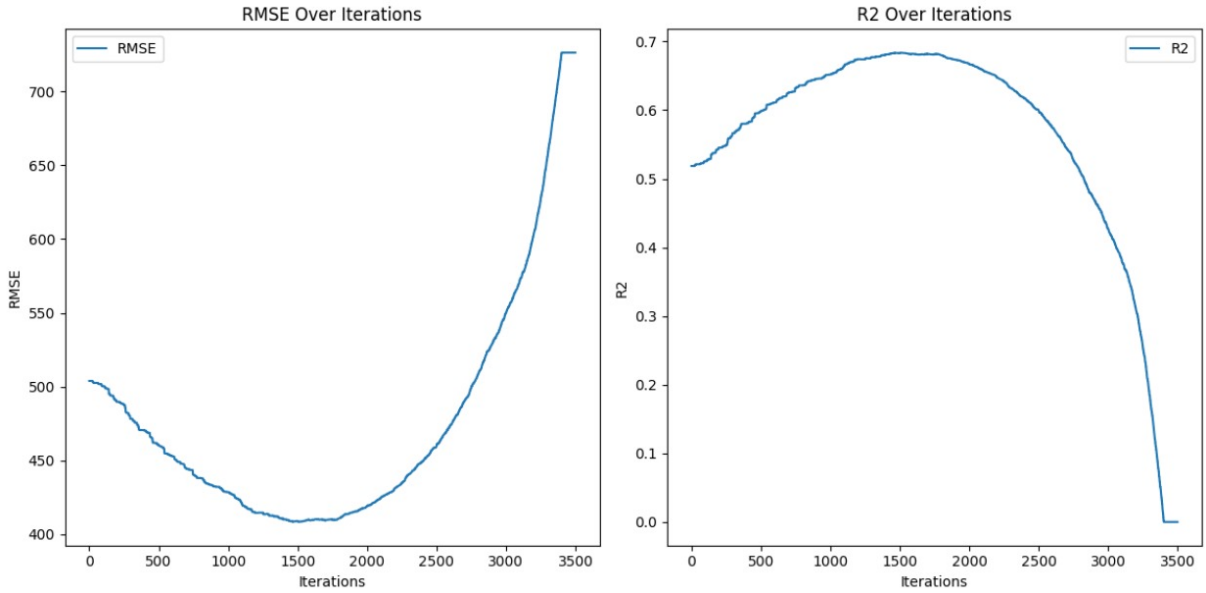

Figure 9: Graphical depiction of RMSE variation with R2 score, showcasing the trade-off when capping model predictions beyond a specified threshold

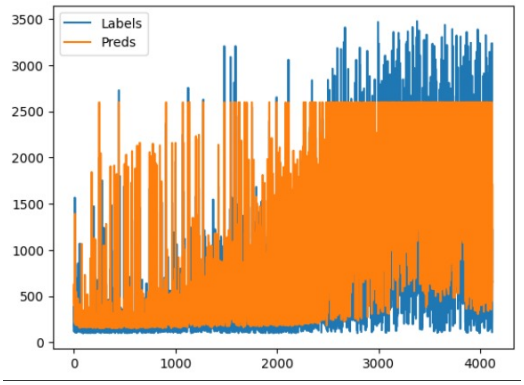

Figure 10: Predictions after optimum offset

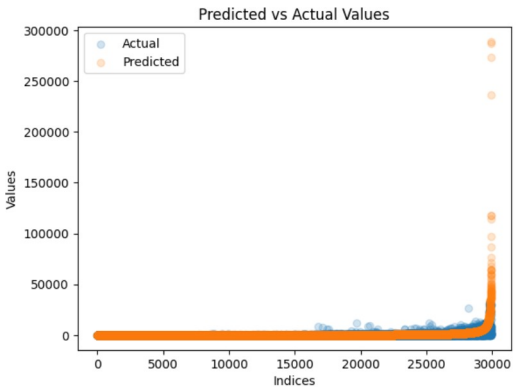

Figure 11: XGBoost model predictions vs labels

Changes	Dataset	Epoch 1	Epoch 3	Val	RMSE	R2
Base Model	0-100	0.96	0.49	0.61	36.26	-600
Including video views	0-100	1.01	0.33	0.48	-	-
Including username	0-100	0.908	0.199	0.217	16.18	0.73
Including duration	0-100	0.954	0.202	0.192	15.01	0.763
Including date	0-100	1.12	0.221	0.24	15.66	0.743
Including images and gifs	0-100	0.485	0.23	0.248	13.2	0.81
Removing the dates	0-100	0.46	0.224	0.252	13.7	0.787
Including prompts and dates	0-100	0.492	0.227	0.255	13.1	0.81
Removing gifs	0-100	0.46	-	0.259	13.33	0.82
Including captions	0-100	0.349	0.256	0.277	14.64	0.76
Tested only on views, img, text, user and date	0-100	0.446	0.243	0.264	11.88	0.84
Using roberta	0-100	0.407	-	0.254	11.56	0.85

Table 2: Results of feature engineering for the model in the range of 0-100 likes.

Changes to base model	Dataset	Epoch 1 loss	Epoch 2 loss	Val loss	RMSE	R2
Including views, duration and username	100-3.5K	2.55	0.34	0.345	503.88	0.512
After optimum offset*	100-3.5k	2.55	0.34	0.345	408.38	0.684
Including images and dates and removing duration of videos.	100-3.5K	1.2	0.426	0.425	527.78	0.435
Including duration	100-3.5K	1.16	0.43	0.56	576.16	0.33
Tested only on content, user, and views	100-3K5	1.2	0.43	0.46	543.8	0.4
Including images	100-3.5K	1.295	0.24	0.485	582	0.3
Including views, duration, username, and testing on Roberta	100-3.5K					

Table 3: Results of feature engineering for the model in range of 100-3500 likes

The base model we use is a pre-trained language model, called BERTweet, which utilizes only the textual features of the tweets to predict the number of likes they receive.

B Appendix: Task 2 - Prompt Used

Listing 1: Prompt

```
<s><s> [INST]
Using the given tweets as reference construct a tweet which has analogical
similarity to those, conditioned on the fact that the post is a
Moderate Engagement post.

Reference tweets:
Tweet: China built a hospital in just 10 days for coronavirus
patients <hyperlink>
Username: Independent
Company: independent
Likes: Moderate Engagement

Tweet: A massive construction effort is under way in Golokhvastovo, 50km
southwest of Moscow, to build a hospital for the treatment of COVID-19,
the disease caused by the coronavirus. The site is expected to have 500
patient beds and separate dormitories to house a staff of 1,000.
<hyperlink>
Username: RFERL
Company: free
Likes: High Engagement

Tweet: The tunnel boring machine "Prerna" has reached its destination at
Esplanade Metro station  today after a fascinating journey of
3.8 kms below the twin cities of Howrah and Kolkata including the
river Hooghly <hyperlink>
Username: metrorailwaykol
Company: metro
Likes: Low Engagement



The tweet is written by the user cnni belong to cnn.

New Tweet: An 85-year-old primary school in Shanghai has been
lifted off the ground     in its entirety     and relocated
using new technology dubbed the "walking machine."
<hyperlink> <hyperlink> </s></s><s>
```

Glossary

- AutoGluon Regressor

AutoGluon's Regressor offers a powerful automated approach to regression tasks, leveraging a suite of machine learning models and hyperparameter optimization techniques. Its strength lies in its ability to autonomously select the most suitable model architecture and hyperparameters for the given dataset, significantly reducing the need for manual tuning. By encapsulating state-of-the-art models and handling various aspects of the modeling pipeline, such as feature engineering, model selection, and ensemble methods, AutoGluon Regressor serves as an efficient and effective tool for practitioners seeking high-quality regression solutions without extensive manual intervention.

- LanguageBind

LanguageBind (Zhu et al., 2023) takes language as the binding element across different modalities because the language modality is well-explored and contains rich semantics. Specifically, the language encoder acquired by VL pretraining is used to train encoders for other modalities with contrastive learning. As a result, all modalities are mapped to a shared feature space, implementing multi-modal semantic alignment.

- RoBERTa

RoBERTa (Liu et al., 2019) stands for "A Robustly Optimized BERT Pretraining Approach." It is a variant of the BERT (Bidirectional Encoder Representations from Transformers) model. RoBERTa improves upon BERT's pretraining approach by using larger mini-batches, dynamic masking, and removing the next sentence prediction objective. It achieves state-of-the-art performance on various natural language understanding tasks, leveraging its bidirectional contextual embeddings to capture intricate patterns and semantics within text data.

- ViT

Vision Transformers (Lin et al., 2023) represent a breakthrough in computer vision by adopting the transformer architecture for image recognition tasks. Unlike traditional convolutional neural networks (CNNs), ViT divides images into fixed-size patches and processes them through a series of transformer layers, enabling global context aggregation and capturing long-range dependencies within the image. By leveraging self-attention mechanisms, ViT learns to associate information across different patches, achieving competitive performance on various image classification benchmarks. Its ability to handle varying image resolutions and scalability to larger datasets makes ViT a promising architecture for image understanding tasks.

- Bertweet

BERTweet (Nguyen et al., 2020) is a variant of the BERT (Bidirectional Encoder Representations from Transformers) model that is specifically pretrained on Twitter data. It incorporates the unique characteristics of tweets, such as hashtags, mentions, and informal language, to better understand and process text in a Twitter context. By training on a large corpus of tweets, BERTweet aims to capture the intricacies of language used on Twitter, enabling more effective natural language understanding and processing for tasks related to social media analysis, sentiment analysis, and other Twitter-specific applications.

- XGBoost

XGBoost, short for eXtreme Gradient Boosting, is a widely-used and powerful machine learning algorithm known for its exceptional performance in regression, classification, and ranking tasks. It operates by iteratively building a sequence of decision trees, each correcting the errors of its predecessor. XGBoost excels in handling large datasets, thanks to its optimized implementation that incorporates techniques like gradient boosting, regularization, and parallel processing.

- EVA-CLIP

EVA-CLIP (Sun et al., 2023a) introduces evolutionary algorithms to fine-tune CLIP's vision-language capabilities for downstream tasks. By leveraging evolutionary strategies, EVA-CLIP aims to enhance model performance through population-based optimization, enabling it to adapt and improve across various domains. This approach facilitates efficient fine-tuning and adaptation of CLIP's multimodal understanding, potentially yielding better performance in tasks requiring complex vision-language interactions.

- ViT-GPT2

ViT-GPT2 (Kumar, 2022), a hybrid model combining the Vision Transformer (ViT) architecture with the Generative Pre-trained Transformer 2 (GPT2),

represents a significant stride in multimodal learning. This fusion marries the prowess of ViT in visual understanding with the language generation finesse of GPT2. By leveraging ViT's attention mechanism for image input and GPT2's transformer-based language modeling capabilities, ViT-GPT2 excels in understanding visual context while generating coherent and contextually rich textual outputs. Its ability to jointly process images and text offers promising potential for various tasks demanding a nuanced comprehension of both visual and textual information.

- PaddleOCR

PaddleOCR, a part of the PaddlePaddle deep learning platform, stands as a robust and versatile tool for optical character recognition tasks. Leveraging advanced deep learning techniques, PaddleOCR offers a comprehensive suite capable of recognizing text in various languages and diverse formats within images and documents. Its flexibility, coupled with pre-trained models and the ability for fine-tuning, empowers users to efficiently extract text information, making it a valuable asset for projects requiring accurate and efficient text recognition capabilities.

- Katna

Katna automates the boring, error-prone task of video key/best frame extraction, video compression, and the manual time-consuming task of image cropping and resizing using machine learning.

- TimesFormer-GPT2

TimesFormer-GPT2 is an advanced language model that merges the Transformer architecture with the principles of the GPT-2 model. Developed as an extension of the GPT-2 framework, TimesFormer-GPT2 exhibits enhanced capabilities in understanding and generating text by leveraging self-attention mechanisms. Its architecture enables the model to grasp contextual dependencies efficiently, making it adept at various natural language processing tasks, including text generation, summarization, and language understanding.

- NeMo

NVIDIA is a conversational AI toolkit built for researchers working on automatic speech recognition (ASR), text-to-speech synthesis (TTS), large language models (LLMs), and natural language processing (NLP).

- Mistral-7B

The Mistral-7B-v0.1 Large Language Model (LLM) (Jiang et al., 2023) is a pretrained generative text model with 7 billion parameters. Mistral 7B is engineered for superior performance and efficiency. It outperforms the best open 13B model (Llama 2) across all evaluated benchmarks, as well as the best released 34B model (Llama 1) in reasoning, mathematics, and code generation. The model leverages grouped-query attention (GQA) for faster inference, coupled with sliding window attention (SWA) to effectively handle sequences of arbitrary length with reduced inference cost.