

A Multimodal Analysis of Influencer Content on Twitter

Danae Sánchez Villegas^α Catalina Goanta^β Nikolaos Aletras^α

^α Computer Science Department, University of Sheffield, UK

^β Utrecht University

{dsanchezvillegas1, n.aletras}@sheffield.ac.uk

e.c.goanta@uu.nl

Abstract

Influencer marketing involves a wide range of strategies in which brands collaborate with popular content creators (i.e., influencers) to leverage their reach, trust, and impact on their audience to promote and endorse products or services. Because followers of influencers are more likely to buy a product after receiving an authentic product endorsement rather than an explicit direct product promotion, the line between personal opinions and commercial content promotion is frequently blurred. This makes automatic detection of regulatory compliance breaches related to influencer advertising (e.g., misleading advertising or hidden sponsorships) particularly difficult. In this work, we (1) introduce a new Twitter (now X) dataset consisting of 15,998 influencer posts mapped into commercial and non-commercial categories for assisting in the automatic detection of commercial influencer content; (2) experiment with an extensive set of predictive models that combine text and visual information showing that our proposed cross-attention approach outperforms state-of-the-art multimodal models; and (3) conduct a thorough analysis of strengths and limitations of our models. We show that multimodal modeling is useful for identifying commercial posts, reducing the amount of false positives, and capturing relevant context that aids in the discovery of undisclosed commercial posts.¹

1 Introduction

Social media influencers are content creators who have established credibility in a specific domain (e.g., fitness, technology), are sometimes followed by a large number of accounts and can impact the buying decisions of their followers (Keller and Berry, 2003; Brown and Hayes, 2008; Nandagiri and Philip, 2018; Lee et al., 2022). Influencer marketing (i.e., promoted content via influencer posts



Commercial: For a truly beautiful and delicate summer fragrance you have to try @USER's newest scent.



Non-commercial: So that's tonight's dinner, tomorrow's lunch, dinner & inbetweens sorted.

Figure 1: *Commercial* and *non-commercial* tweets in our dataset.

in social media) has gained popularity as an alternative to traditional advertising (e.g., magazines, television, billboards) and mainstream digital marketing such as pop-up and platform ads (Leerssen et al., 2019; Nandagiri and Philip, 2018; Lou et al., 2019; Jarrar et al., 2020; Fang and Wang, 2022) for reaching a larger and more targeted audience (Gross and Wangenheim, 2018).

Influencer marketing is dominated by *native advertising* where there is no obvious distinction between *commercial* (i.e., content that is monetized) and *non-commercial* content such as personal thoughts, sentiment and experiences (Chia, 2012). Even though the disclosure of *commercial* content (via keywords such as #ad, #sponsored) by influencers has become a requirement in some countries due to consumer protection obligations,² identifying *commercial* content in influencer posts is challenging in practice because (1) disclosure guidelines are not always followed, e.g., not including or hiding standard disclosure terms³ (Wojdyski, 2016; Boerman and van Reijmersdal, 2016; Mathur et al., 2018; Alassani and Göretz, 2019; De Gregorio and Goanta, 2020); and (2) brand cues (i.e., elements that may affect buying behavior) may appear in different modalities such as text, images

²<https://icas.global/advertising-self-regulation/influencer-guidelines/>

³Only about 10% of affiliate marketing content on Pinterest and YouTube contains any disclosures (Mathur et al., 2018).

¹Code and data are available at <https://github.com/danaesavi/micd-influencer-content-twitter>

or both (Sánchez Villegas et al., 2021). Figure 1 shows an example of a *commercial* and a *non-commercial* post. Both examples appear to include products, however only the top example is *commercial*. This makes it difficult for the users to distinguish between paid promotion and personal opinions.

Therefore, automatically detecting whether an influencer’s post involves paid promotion of products or services is of utmost importance for addressing issues related to transparency and regulatory compliance, such as misleading advertising or undisclosed sponsorships in large scale (Mathur et al., 2018; Evans et al., 2017; Wojdowski et al., 2018; Ducato, 2020; Ershov and Mitchell, 2020). Previous work on identifying influencer commercial content has focused on analyzing user features (e.g., popularity and engagement) and network characteristics of influencers (Zarei et al., 2020; Kim et al., 2021b), while the use of language and its relationship to images has not been explicitly explored.

In this work, we present a new expert annotated Twitter (now X) dataset and an extensive empirical study on influencer multimodal content focused on analyzing the contribution of text and image modalities to *commercial* and *non-commercial* posts. Our main contributions are as follows:

- We present a large publicly available dataset of 14,384 text-image pairs and 1,614 text-only influencer tweets written in English. Tweets are mapped into *commercial* and *non-commercial* categories;
- We benchmark an extensive set of state-of-the-art language, vision and multimodal models for automatically identifying *commercial* content, including prompting large language models (LLMs);
- We propose a simple yet effective cross-attention multimodal approach that outperforms all text, vision and multimodal models;
- We conduct a qualitative analysis to shed light on the limitations of automatically detecting *commercial* content, and provide insights into when each modality is beneficial.

2 Related Work

2.1 Computational Studies on Influencers

Previous work has analyzed the characteristics of influencers on social media platforms such as Twit-

ter (Huang et al., 2014; Lagrée et al., 2018; Han et al., 2021), Instagram (Kim et al., 2017, 2021a; Fernandes et al., 2022) and Pinterest (Gilbert et al., 2013; Mathur et al., 2018). Kim et al. (2017) investigate the social relationships and interactions among influencers while Kim et al. (2021a) explore the audience loyalty and content authenticity. On Twitter, Lagrée et al. (2018) leverage social network analysis to discover influencers that achieve high reach on advertising campaigns and Han et al. (2021) study the relationships among fashion influencers to understand who they follow, mention, and retweet. Using posts from Pinterest and YouTube, Mathur et al. (2018) examine whether influencers comply with advertising disclosure regulations and show that while influencer commercial content has increased over the years, its disclosure remains limited.

2.2 Data Resources for Influencer Content Analysis

Datasets for analyzing influencer content have been developed to analyze the influencers’ impact on spreading information (Han et al., 2021), categorizing influencers into different domains, e.g., fashion, beauty (Kim et al., 2020), and analyzing the characteristics of branded content (Yang et al., 2019). Yang et al. (2019) introduce a dataset to study how influencers mention brands in their posts. They collect 800K Instagram posts from 18K influencers that explicitly mention (@mention) a brand, and characterize them as sponsored or non-sponsored using three sponsorship indicators: *#ad*, *#sponsored*, *#paidAD*.

Datasets for analyzing commercial content shared by influencers have been developed by Zarei et al. (2020) and Kim et al. (2021b). Zarei et al. (2020) present a dataset consisting of 35K Instagram posts and 99K stories (i.e., posts that disappear after 24 hours) from 12K influencers and use an LSTM model (Hochreiter and Schmidhuber, 1997) to identify whether a post is sponsored or not. Kim et al. (2021b) develop a dataset of 38K influencer posts that explicitly mention (@mention) a brand. Similar to Yang et al. (2019), they label these posts as sponsored if they contain at least one of three sponsorship indicators: *#ad*, *#sponsored*, *#paidAD*. They propose an attention-based neural network model to classify posts as sponsored or non-sponsored.

Dataset	Publicly Available	Posts w/o brand mentions	Human Annotation	Keyword Matching	No. of Commercial Keywords	Platform	Modality	Time Range	Domains
Han et al. (2021)	✗	✗	✗	✗	0	Twitter	Text	not specified	fashion
Zarei et al. (2020)	✗	✓	✗	✓	7	Instagram	Text	Jul 2019 - Aug 2019	not specified
Yang et al. (2019)	✗	✗	✗	✓	3	Instagram	Text & Image	not specified	not specified
Kim et al. (2021b)	✓	✓	✗	✓	3	Instagram	Text & Image	not specified	not specified
Kim et al. (2020)	✓	✗	✗	✓	1	Instagram	Text & Image	Oct 2018 - Jan 2019	beauty, family, food, fashion, pet, fitness, interior, travel,
MICD (Ours)	✓	✓	✓	✓	26	Twitter	Text & Image	Jan 2015 - Aug 2021	beauty, travel, food fitness, technology, lifestyle

Table 1: A comparison of existing datasets for influencer content analysis

Limitations of existing resources Table 1 compares existing datasets for analyzing influencer content. We observe that current datasets have only used a limited set of keywords (e.g., *#ad*) for identifying posts with commercial content (seven or less). While some datasets include only text content (Zarei et al., 2020), others focus only on posts that explicitly mention (@mention) a brand (Yang et al., 2019; Kim et al., 2021b). In contrast to prior datasets for analyzing influencer commercial content that use Instagram, we use Twitter because it is a text-first platform and has rapidly increased in popularity as a tool for influencer marketing. For instance, 49% of Twitter users say that they have made a purchase as a direct result of a Tweet from an influencer.⁴

3 Multimodal Influencer Content Dataset (MICD)

We present a new multimodal influencer content dataset (MICD) consisting of Twitter posts mapped into *commercial* and *non-commercial* classes.

3.1 Retrieving Candidate Influencers

To map tweets into these two classes, we first need to identify candidate influencers on Twitter. We look for candidate accounts in six different domains (i.e., *Beauty*, *Travel*, *Fitness*, *Food*, *Tech* and *Lifestyle*) to ensure thematic diversity. The domains related to ‘Beauty’, ‘Fitness’, ‘Travel’ and ‘Lifestyle’ are among the most popular in Twitter,⁵ while *Food* and *Tech* have recently gained attention (Allassani and Göretz, 2019; Weber et al., 2021). To retrieve influencers, we query for accounts that contain domain-specific keywords in

⁴https://blog.twitter.com/en_us/a/2016/new-research-the-value-of-influencers-on-twitter

⁵<https://influencermarketinghub.com/influencer-marketing-benchmark-report-2021/>

their bios (e.g., *beauty vlogger*, *travel influencer*, *lifestyle blogger*, *food writer*) as influencers tend to provide such information in profile descriptions (Kim et al., 2020).⁶ We collect all available image-text tweets written in English from each account using the Academic Twitter API.⁷ Duplicate tweets with identical text are removed.

3.2 Keyword-based Weak Labeling

We initially use a keyword-based strategy to automatically map posts into the *commercial* and *non-commercial* categories (i.e., weak labeling). This is suitable in a real-world scenario of an automatic regulatory compliance system with limited resources for manually labeling all available posts (Zarei et al., 2020; Kim et al., 2021b).

Commercial Commercial tweets include content that promotes or endorses a brand or its products or services, a free product or service or any other incentive. Thus, we extract keywords strongly associated with influencer marketing following the official guidelines provided by the Federal Trade Commission (FTC, 2019) in the US, and the Advertisements Standards Authority and Competition and Markets Authority in the UK (CMA, 2020). These guidelines contain lists of keywords to appropriately disclose commercial content. In this work, we considerably extend the keyword lists (extended and verified by members of a national consumer authority) to not only include recommended sponsorship disclosure terms (e.g., *#ad*, *#sponsored*), but also terms that are relevant to different business models (i.e., market practices based on the obligations of the parties) such as gifting (e.g., *#gift*, *#giveaway*), endorsements (e.g., *#ambassador*) and

⁶Influencer accounts were manually validated to ensure bots are not included.

⁷<https://developer.twitter.com/en/products/tw-itter-api>

Domain	Accounts
Beauty	22
Travel	22
Fitness	15
Food	22
Tech	20
Lifestyle	31
Total	132

Table 2: Number of influencer accounts by domain

affiliate marketing (e.g., *#aff*, *discount code*). A complete list of keywords can be found in Appx. A. We label as *commercial* all tweets containing at least one of the influencer marketing keywords excluding tweets where the keyword is negated (e.g., *not ad*, *not an ad*). To avoid data leakage in the experiments, we remove all of the keywords used for data labeling (see Sec. 5.1) from the posts after labeling them. As a result, our models can identify *commercial* content without the use of such terms (see Sec. 4).

Non-commercial Non-commercial posts refer to organic content such as personal ideas, comments and life updates that do not aim for monetization. Thus, all tweets that do not include any of the keywords presented above are considered *non-commercial*. To balance the dataset, we sample *non-commercial* posts weighted according to the number of *commercial* tweets for each account.

3.3 Data Splits

Text-Image Sets We split the tweets into train, dev and test sets at the account level (i.e., tweets included in each split belong to different accounts) to ensure that models can generalize to unseen influencer accounts and prevent information leakage in our experiments.

Text-only Test Set We further collect text-only posts from influencer accounts in the test set. We sample text-only tweets according to the number of tweets for each influencer account in the test set, resulting in a total of 1,614 text-only tweets. This is done to account for cases where only text content is provided.⁸

3.4 Human Data Annotation

To ensure a high quality data set for evaluation, we use human annotators for labeling all tweets in both

⁸Note that while text-only tweets are prevalent on Twitter, image-only tweets are uncommon.

Split	Non-commercial	Commercial	Total
Train	5,781	5,596	11,377 (79.1%)
Dev	789	783	1,572 (10.9%)
Test	689	746	1,435 (10%)
Total	7,259	7,125	14,384
Text-only Test	1,377	237	1,614
All	8,636	7,352	15,998

Table 3: Dataset statistics showing the number of tweets for each split.

test sets (text-image and text-only test sets).⁹ Four volunteer annotators from our institution, each with a substantial legal background and knowledge of advertising disclosure regulations labeled the test dataset. A workshop was held to introduce the task to the annotators, explain the annotation guidelines and run a calibration round on a random set of 20 examples. All tweets in the test sets were labeled by two different annotators as *commercial*, *non-commercial*, or *unclear* (i.e., it is not clear whether the post contains *commercial* content or not). In cases of disagreement, a third independent annotator assigned the final label (*commercial* or *non-commercial*) after adjudication. Posts labeled as *unclear* (15) are removed, as well as posts written in other language than English (2).

The inter-annotator agreement between two annotations across all tweets is 0.78 Cohen’s-Kappa (Cohen, 1960) that corresponds to the upper part of the *substantial* agreement band (Artstein and Poesio, 2008). Furthermore, the agreement between the automatic weak labels and the resulting human annotations is 0.67 Cohen’s-Kappa which corresponds to *substantial* agreement and denotes weak labels of good quality for model training.

Our final dataset contains 14,384 text-image pairs (7,259 *non-commercial* and 7,125 *commercial*). Additionally, the text-only test set consists of 1,614 tweets (1,377 *non-commercial* and 237 *commercial*). Table 3 shows the distribution of *commercial* and *non-commercial* tweets by split.

3.5 Exploratory Analysis

Exploratory analysis of our dataset revealed that influencer accounts in our dataset have between 8K and 500K followers covering micro and macro influencers which are considered to create highly persuasive content (Kay et al., 2020). Table 2 shows the number of influencer accounts per domain. In average, each domain contains 22 accounts, and

⁹We received approval from the Ethics Committee of our institution. Annotation guidelines can be found in Appx. B.

all accounts have a minimum of 10 *commercial* tweets. Finally, we observe a different label distribution in text-image and text-only test splits. Text-only test split is unbalanced with most posts manually annotated as *non-commercial* (85.32% *non-commercial*, 14.68% *commercial*). On the other hand, text-image test set label distribution is balanced (48.01% *non-commercial*, 51.99% *commercial*). This highlights the use of visuals in influencer marketing for effectively advertising products, which is consistent with findings in conventional online advertising research (Mazloom et al., 2016). It also emphasizes the multimodal nature of the task.

3.6 Comparison with Related Datasets

Table 1 compares our dataset, MICD, to related datasets for influencer content analysis (see Sec. 2). Our dataset contains posts with and without explicit (i.e., @USER) brand mentions from influencers of different domains. We follow a similar approach for weak labeling *commercial* posts as previous work (Zarei et al., 2020; Kim et al., 2021b), but we considerably extend the list of keywords following relevant guidelines and experts feedback (see Sec. 3.2). Moreover, we include test sets with a total of 3,049 tweets annotated by experts in the legal domain. We anticipate that this dataset will be beneficial not only for this study, but also for future influencer content analysis research.

4 Influencer Content Classification Models

Given a social media post P (e.g., a tweet) consisting of a text and image pair (L, I) , the task is to classify a post P into the correct category (*commercial* or *non-commercial*).

4.1 Unimodal Models

Prompting We first experiment with prompting **Flan-T5** (Chung et al., 2022) and **GPT-3** (Brown et al., 2020). We use the following prompt: “Label the next text as ‘commercial’ or ‘not commercial’. Text: <TWEET>”. We map responses to the corresponding *commercial* or *non-commercial* class and report results for each model (zero-shot). We further experiment with few-shot prompting by appending four randomly selected training examples¹⁰ (two examples from each class) before each prompt (few-shot). We run this three times

¹⁰Appx. D includes the template we use for these prompts.

with a different set of examples and report average performance.

Image-only Models We fine-tune two pre-trained models that achieve state-of-the-art results in various computer vision classification tasks by adding an output classification layer: (1) **ResNet152** (He et al., 2016) and (2) **ViT** (Dosovitskiy et al., 2020). ResNet uses convolution to aggregate information across locations, while ViT uses self-attention for this purpose. Both models are pre-trained on the ImageNet dataset (Russakovsky et al., 2015).

Text-only Recurrent Model Zarei et al. (2020) propose a contextual Long-Short Term Memory (LSTM) neural network architecture for identifying posts in Instagram. Thus, we also experiment with a similar bidirectional LSTM network with a self-attention mechanism (Hochreiter and Schmidhuber, 1997) to obtain the tweet representation that is subsequently passed to the output layer with a softmax activation function (**BiLSTM-Att**).

Text-only Transformers We fine-tune two pre-trained transformer-based (Vaswani et al., 2017) models for commercial posts prediction: **BERT** (Devlin et al., 2019) and **BERTweet** (Nguyen et al., 2020) by adding a classification layer on top of the [CLS] token. **BERTweet** is a BERT based model pre-trained on a large-scale corpus of English Tweets.

4.2 Multimodal Models

Text & Image Transformers We fine-tune three multimodal transformer-based models: **MMBT** (Kiela et al., 2019), **ViLT** (Kim et al., 2021c) and **LXMERT** (Tan and Bansal, 2019). MMBT uses ResNet and BERT as image and text encoders respectively, ViLT uses a convolution-free encoder similar to ViT, and LXMERT takes *object-level* features as input (see Sec. 5.1). ViLT and LXMERT are multimodally pre-trained on visual-language tasks such as image-text matching and visual question answering.

Aspect-Attention Kim et al. (2021b) proposed an aspect-attention fusion model to rank Instagram posts based on their likelihood of including undeclared paid partnerships. Thus, we repurpose their model to identify commercial posts on Twitter. Aspect-attention fusion consists of generating a score for each modality by applying the attention mechanism across the image and text vectors.

Then, the multimodal post representation is produced by computing a linear combination of the score and the unimodal representations. The model is fine-tuned by adding a fully-connected layer with a softmax activation function (**Aspect-Att**).

ViT-BERTweet-Att We propose to combine unimodal pretrained representations via cross-attention fusion strategy so that text features can guide the model to pay attention to the relevant image regions. We use BERTweet to obtain contextual representations of the text content $L \in R^{d_L \times m_L}$, where L is the output of the last layer of BERTweet, d_L is the hidden size of BERTweet and m_L is the text sequence length. For encoding the images, we use the Vision Transformer pre-trained on ImageNet (Russakovsky et al., 2015). We obtain the visual representations of the image content $I \in R^{d_I \times m_I}$, where I is the output of the last layer of ViT, d_I is the hidden size of ViT and m_I is the image sequence length. We propose to capture the inter-modality interactions using a cross-attention layer. Specifically, given L and I , we compute the scaled dot attention with L as queries, and I as keys and values as follows: $\text{Cross-Att}(L, I) = \text{softmax}\left(\frac{[W^Q L][W^K I]^T}{\sqrt{d_k}}\right)[W^V I]$, where $\{W^Q, W^K, W^V\}$ are learnable parameters, $d_k = d^L = d^I$, and $\text{Cross-Att}(L, I) \in R^{m_L \times d_k}$.

The multimodal representation vector h is obtained by concatenating the ‘classification’ $[\text{CLS}]_L$ token from L (output from the last layer of BERTweet), and the $[\text{CLS}]_{Att}$ token from the output of the cross-attention layer ($\text{Cross-Att}(L, I)$). In this way, we leverage the text content of the influencer posts, and the relevant information from the image content. We fine-tune the model on the commercial content classification task by adding a fully-connected layer with a softmax activation function.¹¹

5 Experimental Setup

5.1 Data Processing

Text For each tweet, we lowercase and tokenize text using DLTK (Schwartz et al., 2017). We also replace URLs and user @-mentions with placeholder tokens following the BERTweet pipeline (Nguyen et al., 2020). Emojis are replaced with their corresponding text string, e.g thumbs_up. Keywords used in the weak labeling process (Sec. 3.2) are removed from all *commercial* tweets.

¹¹Figure 4 shows a diagram of the model.

Image Images are resized to (224×224) pixels representing a value for the red, green and blue color in $[0, 255]$. The pixel values are normalized to $[0 - 1]$. For LXMERT, we extract *object-level* features using Faster-RCNN (Ren et al., 2016) as in Anderson et al. (2018) and keep 36 objects for each image as in Tan and Bansal (2019).

5.2 Most Freq. Baseline and Evaluation

Most Freq. Baseline We assign the most frequent label in the training set to all instances in the test set.

Evaluation We evaluate all models using weighted-averaged¹² F1, precision, and recall to manage imbalanced classes. Results are obtained over three runs using different random seeds reporting average and standard deviation.

5.3 Implementation Details

We select the hyperparameters for all models using early stopping by monitoring the validation loss. We use the Adam optimizer (Kingma and Ba, 2014). We estimate the class weights using the ‘balanced’ heuristic (King and Zeng, 2001). All experiments (unless indicated) are performed using an Nvidia V100 GPU with a batch size of 16.

Prompting We use one GPU T4 to obtain the inference results from Flan-T5 (Chung et al., 2022) model. We use the large version from HuggingFace library (780M parameters) (Wolf et al., 2019). For GPT-3 (Brown et al., 2020), we use the *text-davinci-003* model via the OpenAI¹³ Library. Prompt templates are included in Appx. D.

Image-only For ResNet152 (He et al., 2016), we fine-tune for 1 epoch with learning rate $\eta = 1e^{-5}$ and dropout $\delta = 0.05$ before passing the image representation through the classification layer. We fine-tune ViT (Dosovitskiy et al., 2020) for 3 epochs with learning rate $\eta = 1e^{-5}$ and dropout $\delta = 0.05$. $\eta \in \{1e^{-3}, 1e^{-4}, 1e^{-5}\}$ and δ in $[0, 0.5]$, random search.

Text-only Recurrent Model For BiLSTM-Att we use 200-dimensional GloVe embeddings (Pennington et al., 2014) pre-trained on Twitter data. The maximum sequence length is set to 50. The LSTM size is $h = 32$ where $h \in \{32, 64, 100\}$ with dropout $\delta = 0.3$ where $\delta \in [0, 0.5]$, random search.

¹²Macro-averaged results are included in Appx. C.

¹³<https://platform.openai.com/docs/>

We use Adam (Kingma and Ba, 2014) with learning rate $\eta = 1e^{-3}$ with $\eta \in \{1e^{-3}, 1e^{-4}, 1e^{-5}\}$, minimizing the binary cross-entropy using a batch size of 8 over 6 epochs with early stopping.

Text-only Transformers We fine-tune BERT and BERTweet for 20 epochs and choose the epoch with the lowest validation loss. We use the pre-trained base-uncased model for BERT (Vaswani et al., 2017; Devlin et al., 2019) from HuggingFace library (12-layer, 768-dimensional) (Wolf et al., 2019), and the base model for BERTweet (Nguyen et al., 2020) with a maximal sequence length of 128. We fine-tune BERT for 1 epoch, learning rate $\eta = 1e^{-5}$ and dropout $\delta = 0.05$; and BERTweet for 2 epochs, $\eta = 1e^{-5}$ and $\delta = 0.05$. For all models $\eta \in \{2e^{-5}, 1e^{-4}, 1e^{-5}\}$ and $\delta \in [0, 0.5]$, random search.

Text & Image Transformers We train MMBT (Kiela et al., 2019) for 1 epoch and $\eta = 1e^{-5}$ where $\eta \in \{1e^{-3}, 1e^{-4}, 1e^{-5}\}$ and dropout $\delta = 0.05$ (δ in $[0, 0.5]$, random search) before passing through the classification layer. ViLT (Kim et al., 2021c) is fine-tuned for 4 epochs and $\eta = 1e^{-5}$, vision layers are frozen. LXMERT (Tan and Bansal, 2019) is fine-tuned for 3 epochs with $\eta = 1e^{-5}$ and $\delta = 0.05$.

Aspect-Attention and ViT-BERTweet-Att We train Aspect-Attention and ViT-BERTweet-Att with BERTweet as text encoder and ViT as image encoder for 15 epochs and choose the epoch with the lowest validation loss. Aspect-Attention: 1 epoch with $\eta = 1e^{-5}$ and $\delta = 0.05$ and ViT-BERTweet-Att 3 epochs with $\eta = 1e^{-5}$ and $\delta = 0.05$; The dimensionality of the multimodal representation is 768. $\eta \in \{1e^{-3}, 1e^{-4}, 1e^{-5}\}$ and δ in $[0, 0.5]$, random search.

6 Results

Table 4 presents the performance on *commercial* and *non-commercial* influencer content prediction of all predictive models on our new multimodal influencer content dataset (MICD).

6.1 Unimodal Models

We first observe that the two image-only models obtain similar performance. Although both models surpass Most Freq. baseline and Flan-T5 prompting, the text-only models (BiLSTM-ATT, BERT and BERTweet) perform better than image-only models. This corroborates results from pre-

Model	F1	P	R
Most Freq.	31.15 _{0,0}	23.05 _{0,0}	48.01 _{0,0}
Prompting			
Flan-T5 (zero-shot)	42.98 _{0,0}	72.01 _{0,0}	53.51 _{0,0}
Flan-T5 (few-shot)	48.70 _{1,6}	62.07 _{0,9}	53.47 _{0,6}
GPT-3 (zero-shot)	63.91 _{0,0}	65.64 _{0,0}	64.81 _{0,0}
GPT-3 (few-shot)	69.57 _{1,5}	71.69 _{2,1}	70.01 _{0,8}
Image-only			
ResNet	59.59 _{0,5}	59.85 _{0,5}	59.60 _{0,5}
ViT	60.81 _{1,3}	61.58 _{0,9}	61.02 _{1,2}
Text-only			
BiLSTM-Att* (Zarei et al., 2020)	66.10 _{0,7}	66.48 _{0,8}	65.15 _{0,7}
BERT	74.32 _{0,6}	75.01 _{0,6}	74.43 _{0,7}
BERTweet	76.34 _{0,3}	76.80 _{0,3}	76.45 _{0,3}
Text & Image			
ViLT	68.46 _{0,9}	66.66 _{3,8}	66.66 _{3,8}
LXMERT	70.64 _{0,4}	71.00 _{0,3}	70.68 _{0,4}
MMBT	73.58 _{0,4}	73.79 _{0,6}	73.59 _{0,4}
Aspect-Att* (Kim et al., 2021b)	75.45 _{0,8}	77.42 _{1,1}	75.68 _{0,7}
ViT-BERTweet-Att (Ours)	77.50_{0,6}	78.46_{0,5}	77.61_{0,6}

Table 4: Weighted F1-Score, precision (P) and recall (R) for commercial influencer content prediction. † and ‡ indicates statistically significant improvement (t-test, $p < 0.05$) over BERTweet, and both BERTweet and Aspect-Att respectively. * denotes current state-of-the-art models for influencer commercial content detection. Subscripts denote standard deviations. Best results are in bold.

Model	F1	P	R
BERTweet	76.34 _{0,3}	76.80 _{0,3}	76.45 _{0,3}
ViT	60.81 _{1,3}	61.58 _{0,9}	61.02 _{1,2}
ViT-BERTweet-Concat	76.34 _{0,9}	78.10 _{0,5}	76.54 _{0,8}
ViT-BERTweet-Att (Ours)	77.50_{0,6}	78.46_{0,5}	77.61_{0,6}

Table 5: Comparison of each of the ViT-BERTweet-Att components including the removal of the Cross-Att layer (ViT-BERTweet-Concat). Subscripts denote standard deviations. Best results are in bold.

vious work in multimodal computational social science (Wang et al., 2020; Ma et al., 2021) and influencer content analysis (Kim et al., 2021b). We further note that BERT-based models (BERT and BERTweet) outperform GPT-3 prompting and BiLSTM-Att models over 4% across all metrics. Among the text-only models, BERTweet achieves the highest performance with 76.34, 76.80 and 76.45 weighted F1, precision and recall respectively.

6.2 Multimodal models

State-of-the-art pre-trained multimodal models, ViLT and LXMERT fail to outperform text-only transformers achieving only 68.46 and 70.64 weighted F1 respectively. This emphasizes the challenges for modeling multimodal influencer content. Specifically, ViLT and LXMERT are pretrained

on standard vision-language tasks including image captioning and visual question answering (Zhou et al., 2020; Lu et al., 2019) using data where text and image modalities share common semantic relationships. In contrast, social media advertising frequently employs various types of visual and text rhetoric (e.g., symbolism) to convey their message with no obvious relationship between text and image (Vempala and Preoȃiu-Pietro, 2019; Hessel and Lee, 2020; Sánchez Villegas and Aletras, 2021). Similar behavior is observed with MMBT which obtains comparable performance to BERT. This suggests it is more beneficial to use a text-only encoder (BERTweet) that has been pre-trained on the same domain, in this case Twitter, than fine-tuning a more complex out-of-the-box multimodal transformer model (e.g., ViLT, LXMERT, MMBT).

BERTweet and ViT are used by Aspect-Att (a state-of-the-art model for influencer commercial content prediction) and our model, ViT-BERTweet-Att, to obtain text and visual representations. However, only ViT-BERTweet-Att outperforms all text- and image-only models (77.50, 78.46, 77.61 weighted F1, precision, and recall), indicating that not only the choice of text and image encoders is important, but so is the fusion strategy for effectively modeling text-image relationships for identifying influencer *commercial* content.

6.3 Ablation Study

To analyze the contribution of each component of our ViT-BERTweet-Att in identifying *commercial* posts, Table 5 shows the performance of ViT, BERTweet, and ViT-BERTweet-Att with and without the Cross-Att layer (see Sec. 4). ViT-BERTweet-Att without the Cross-Att layer consists of simply concatenating text and image vectors (ViT-BERTweet-Concat). While the performance of BERTweet and ViT-BERTweet-Concat are comparable (BERTweet and ViT-BERTweet-Concat weighted F1: 76.34), ViT-BERTweet-Att (weighted F1: 77.50) outperforms BERTweet suggesting the Cross-Att layer successfully captures the relevant regions in images for identifying *commercial* posts.

6.4 Text-only Test Set Evaluation

Finally, previous work on text-image classification in *commercial* influencer content has only experimented with fully paired data where every post contains an image and text (Kim et al., 2021b). However, this requirement may not always hold

Model	F1	P	R
Most Freq.	78.55 _{0.0}	72.78 _{0.0}	85.31 _{0.0}
Flan-T5 (zero-shot)	81.02 _{0.0}	80.41 _{0.0}	84.88 _{0.0}
Flan-T5 (few-shot)	82.22 _{0.5}	81.72 _{0.6}	83.56 _{0.6}
GPT-3 (zero-shot)	77.26 _{0.0}	85.12 _{0.0}	73.79 _{0.0}
GPT-3 (few-shot)	84.03 _{3.0}	85.55 _{1.1}	83.68 _{4.8}
BERTweet	87.50 _{1.0}	88.58 _{0.4}	86.84 _{1.3}
ViT-BERTweet-Att (Ours)	88.69 _{0.2}	88.69 _{0.2}	88.93 _{0.5}

Table 6: Weighted F1-Score, precision (P) and recall (R) for commercial influencer content prediction for tweets containing text only. Subscripts denote standard deviations. Best results are in bold.

since not all posts contain both modalities. Thus, we further evaluate our models on our text-only test set (see Sec. 3.3). Table 6 shows the results obtained. We observe a consistent improvement of ViT-BERTweet-Att multimodal model over BERTweet text-only model, i.e., 88.69 versus 87.50. This suggests that multimodal modeling of influencer posts is beneficial for identifying text-only *commercial* posts.

7 Qualitative Analysis

We finally perform a qualitative analysis of the classification effectiveness between ViT-BERTweet-Att and the best text-only model (BERTweet). We analyze the strengths and limitations of each model.

Multimodal modeling helps to reduce the number of false positives. We find that 53% of BERTweet errors from the text-image test set are false positives, i.e., misclassifying *non-commercial* posts as *commercial*, which would be problematic for an automated regulatory compliance system. Our multimodal model, ViT-BERTweet-Att, on the other hand, correctly classifies 38% of BERTweet’s false positive mistakes such as the *non-commercial* post in Figure 1. Similarly, for text-only posts, we observe that 69% of BERTweet misclassifications correspond to false positive errors. 50.9% of these posts are correctly classified by ViT-BERTweet-Att.

Multimodal modeling errors. The most common error when distinguishing *commercial* posts (60%) by our multimodal model, ViT-BERTweet-Att, corresponds to cases where the post includes a standard natural or personal photo, rather than an image depicting products, as is more common in influencer *commercial* content (Kim et al., 2021b) and conventional online advertising (Al-Subhi, 2022). Figure 2 Post A depicts a post incorrectly

labeled as *non-commercial* by ViT-BERTweet-Att and correctly classified by BERTweet.

Multimodal modeling captures context beyond keyword-matching. To analyze if multimodal modeling improves over weak labels, we apply the keyword-based weak labeling approach¹⁴ to the test sets (see Sec. 3.2). We find that 20% and 80% of the weak labeling errors in the text-image and the text-only test sets respectively, are correctly classified by ViT-BERTweet-Att. This suggests that our multimodal model, ViT-BERTweet-Att captures stylistic differences and visual information relevant to identify *commercial* posts beyond keyword-matching. Indeed, most of the errors (85%) in both text-image and text-only posts are false positives (i.e., true label is *non-commercial*) and are mislabeled as *commercial* as they contain one of the keywords, although they are used in a different context. For example: *Just seen that Pepsi ad...awkward.*

Multimodal modeling aids in the discovery of undisclosed commercial posts Using ViT-BERTweet-Att we found undisclosed *commercial* posts (15%) in text-image posts such as the one depicted in Figure 1 (*commercial*) and Figure 2 Post B, as well as in text-only posts such as the next example: *if you love @USER pro-collagen then you might like the new ultra smart line.*

Challenging cases for text and multimodal models. We observe cases that remain challenging for both multimodal and text-only models. Previous work in influencer commercial content on Instagram (Zarei et al., 2020) highlights the difficulty of identifying commercial influencer posts promoting products given the use of *native advertising* (Chia, 2012). However, we find that the most common error (20%) when identifying *commercial* posts (in both text-image and text-only posts), are those that rather than promoting products, they describe their “personal” experiences, particularly while traveling, in both text and image as shown in Figure 2 Post C. These *commercial* posts are difficult to identify as they do not include any specific brand mention or product name and are accompanied by standard traveling images also common in *non-commercial* posts (Oliveira et al., 2020).




Post A	Post B	Post C
		
Combat the cold weather with these incredible @USER sheepskin boots	chunky knits and dainty jewels. This is my favorite vintage sweater #lovechupi	Cherry tree hill is hands down the best view in #Barbados. #VisitBarbados
Actual: C BERTweet: C ViT-BERTweet-Att: NC	Actual: C BERTweet: NC ViT-BERTweet-Att: C	Actual: C BERTweet: NC ViT-BERTweet-Att: NC

Figure 2: Examples of classifications of BERTweet and ViT-BERTweet-Att.

8 Conclusion

We introduced a novel dataset of multimodal influencer content consisting of tweets labeled as *commercial* or *non-commercial*. This is the first dataset to include high quality annotated posts by experts in advertising regulation. We conducted an extensive empirical study including vision, language and multimodal approaches as well as LLM prompting. Our results show that our proposed cross-attention approach to combine text and images, outperforms state-of-the-art multimodal models. Our new dataset can enable further studies on automatically detecting influencer hidden advertising as well as studies in computational linguistics (Sim et al., 2016; Sánchez Villegas et al., 2020; Mu and Aletras, 2020; Jin et al., 2022; Ao et al., 2022) for analysis of commercial language characteristics on a large scale. Future work includes modeling influencer content in multilingual settings.

Limitations

We experimented using only data in English. Influencer advertising strategies could differ across cultures and languages. We plan to address this research direction in future work. We have also presented the main limitations of our best performing model in Section 7.

Ethics Statement

Our work complies with Twitter data policy for research.¹⁵ Tweets were retrieved in August 2021. We received approval from our University Research Ethics Committee.

¹⁴Using the text before removing commercial keywords.

¹⁵See: <https://developer.twitter.com/en/developer-terms/agreement-and-policy>

Acknowledgments

DSV and NA are supported by the Leverhulme Trust under Grant Number: RPG#2020#148. NA is also supported by ESRC (ES/T012714/1). DSV is also supported by the Centre for Doctoral Training in Speech and Language Technologies (SLT) and their Applications funded by the UK Research and Innovation grant EP/S023062/1. CG is supported by the ERC Starting Grant research project HUMANads (ERC-2021-StG No 101041824) and the Spinoza grant of the Dutch Research Council (NWO), awarded in 2021 to José van Dijck, Professor of Media and Digital Society at Utrecht University. We would also like to thank the members of the Competition and Markets Authority in the UK who contributed to the enhancement of the list of terms for our initial keyword-based strategy (refer to Section 3.2). Additionally, we extend our gratitude to the annotators who actively participated in our human annotation labeling task. We would like to thank Mali Jin, Yida Mu, Katerina Margatina, Constantinos Karouzos, Panayiotis Karachristou and all reviewers for their valuable feedback.

References

- Aisha Saadi Al-Subhi. 2022. Metadiscourse in online advertising: Exploring linguistic and visual metadiscourse in social media advertisements. *Journal of Pragmatics*, 187:24–40.
- Rachidatou Alassani and Julia Göretz. 2019. Product placements by micro and macro influencers on instagram. In *International conference on human-computer interaction*, pages 251–267. Springer.
- Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086.
- Xiao Ao, Danae Sanchez Villegas, Daniel Preotiuc-Pietro, and Nikolaos Aletras. 2022. **Combining humor and sarcasm for improving political parody detection**. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1800–1807, Seattle, United States. Association for Computational Linguistics.
- Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational linguistics*, 34(4):555–596.
- Sophie C Boerman and Eva A van Reijmersdal. 2016. Informing consumers about “hidden” advertising: A literature review of the effects of disclosing sponsored content. *Advertising in new formats and media*.
- Duncan Brown and Nick Hayes. 2008. *Influencer marketing*. Routledge.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- Aleena Chia. 2012. Welcome to me-mart: The politics of user-generated content in personal blogs. *American Behavioral Scientist*, 56(4):421–438.
- Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.
- Competition and Markets Authority CMA. 2020. Influencers’ guide to making clear that ads are ads. https://www.ftc.gov/system/files/documents/plain-language/1001a-influencer-guide-508_1.pdf.
- Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.
- Giovanni De Gregorio and Catalina Goanta. 2020. The influencer republic: Monetizing political speech on social media. *Available at SSRN*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. **BERT: Pre-training of deep bidirectional transformers for language understanding**. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.
- Rossana Ducato. 2020. One hashtag to rule them all? mandated disclosures and design duties in influencer marketing practices. In *The Regulation of Social Media Influencers*. Edward Elgar Publishing.
- Daniel Ershov and Matthew Mitchell. 2020. The effects of influencer advertising disclosure regulations: Evidence from instagram. In *Proceedings of the 21st ACM Conference on Economics and Computation*, pages 73–74.

- Nathaniel J Evans, Joe Phua, Jay Lim, and Hyoyeun Jun. 2017. Disclosing instagram influencer advertising: The effects of disclosure language on advertising recognition, attitudes, and behavioral intent. *Journal of interactive advertising*, 17(2):138–149.
- Xing Fang and Tianfu Wang. 2022. Using natural language processing to identify effective influencers. *International Journal of Market Research*, 64(5):611–629.
- Roshan Fernandes, Anisha P Rodrigues, and Bhuvaneshwari Shetty. 2022. **Influencers analysis from social media data**. In *2022 International Conference on Artificial Intelligence and Data Engineering (AIDE)*, pages 217–222.
- Federal Trade Commission FTC. 2019. Disclosures 101 for social media influencers. https://www.ftc.gov/system/files/documents/plain-language/1001a-influencer-guide-508_1.pdf.
- Eric Gilbert, Saeideh Bakhshi, Shuo Chang, and Loren Terveen. 2013. "i need to try this"? a statistical overview of pinterest. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 2427–2436.
- Jana Gross and Florian V Wangenheim. 2018. The big four of influencer marketing. a typology of influencers. *Marketing Review St. Gallen*, 2:30–38.
- Jinda Han, Qinglin Chen, Xilun Jin, Weikai Xu, Wanxian Yang, Suhansan Kumar, Li Zhao, Hari Sundaram, and Ranjitha Kumar. 2021. **Fitnet: Identifying fashion influencers on twitter**. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW1).
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.
- Jack Hessel and Lillian Lee. 2020. **Does my multimodal model learn cross-modal interactions? it's harder to tell than you might think!** In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 861–877, Online. Association for Computational Linguistics.
- Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.
- Jidong Huang, Rachel Kornfield, Glen Szczypka, and Sherry L Emery. 2014. A cross-sectional examination of marketing of electronic cigarettes on twitter. *Tobacco control*, 23(suppl 3):iii26–iii30.
- Yosra Jarrar, Ayodeji Olalekan Awobamise, and Adedola Adewunmi Aderibigbe. 2020. Effectiveness of influencer marketing vs social media sponsored advertising. *Utopia y Praxis Latinoamericana*, 25(12):40–54.
- Mali Jin, Daniel Preotiuc-Pietro, A. Seza Doğruöz, and Nikolaos Aletras. 2022. **Automatic identification and classification of bragging in social media**. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3945–3959, Dublin, Ireland. Association for Computational Linguistics.
- Samantha Kay, Rory Mulcahy, and Joy Parkinson. 2020. When less is more: the impact of macro and micro social media influencers' disclosure. *Journal of Marketing Management*, 36(3-4):248–278.
- Edward Keller and Jonathan Berry. 2003. *The influentials: One American in ten tells the other nine how to vote, where to eat, and what to buy*. Simon and Schuster.
- Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, Ethan Perez, and Davide Testuggine. 2019. Supervised multimodal bitransformers for classifying images and text. *arXiv preprint arXiv:1909.02950*.
- Seungbae Kim, Xiushi Chen, Jyun-Yu Jiang, Jinyoung Han, and Wei Wang. 2021a. Evaluating audience loyalty and authenticity in influencer marketing via multi-task multi-relational learning. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pages 278–289.
- Seungbae Kim, Jinyoung Han, Seunghyun Yoo, and Mario Gerla. 2017. How are social influencers connected in instagram? In *International Conference on Social Informatics*, pages 257–264. Springer.
- Seungbae Kim, Jyun-Yu Jiang, Masaki Nakada, Jinyoung Han, and Wei Wang. 2020. Multimodal post attentive profiling for influencer marketing. In *Proceedings of The Web Conference 2020*, pages 2878–2884.
- Seungbae Kim, Jyun-Yu Jiang, and Wei Wang. 2021b. Discovering undisclosed paid partnership on social media via aspect-attentive sponsored post learning. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pages 319–327.
- Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021c. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR.
- Gary King and Langche Zeng. 2001. Logistic regression in rare events data. *Political analysis*, 9(2):137–163.
- Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- Paul Lagrée, Olivier Cappé, Bogdan Cautis, and Silviu Maniu. 2018. Algorithms for online influencer marketing. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 13(1):1–30.

- Jung Ah Lee, Sabitha Sudarshan, Kristen L Sussman, Laura F Bright, and Matthew S Eastin. 2022. Why are consumers following social media influencers on instagram? exploration of consumers' motives for following influencers and the role of materialism. *International Journal of Advertising*, 41(1):78–100.
- Paddy Leerssen, Jef Ausloos, Brahim Zarouali, Natali Helberger, and Claes H. de Vreese. 2019. Platform ad archives: promises and pitfalls. *Internet Policy Review*, 8(4).
- Chen Lou, Sang-Sang Tan, and Xiaoyu Chen. 2019. Investigating consumer engagement with influencer-vs. brand-promoted ads: The roles of source and disclosure. *Journal of Interactive Advertising*, 19(3):169–186.
- Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. **Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks**. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.
- Chunpeng Ma, Aili Shen, Hiyori Yoshikawa, Tomoya Iwakura, Daniel Beck, and Timothy Baldwin. 2021. **On the (in)effectiveness of images for text classification**. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 42–48, Online. Association for Computational Linguistics.
- Arunesh Mathur, Arvind Narayanan, and Marshini Chetty. 2018. Endorsements on social media: An empirical study of affiliate marketing disclosures on youtube and pinterest. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–26.
- Masoud Mazloom, Robert Rietveld, Stevan Rudinac, Marcel Worring, and Willemijn Van Dolen. 2016. Multimodal popularity prediction of brand-related social media posts. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 197–201.
- Yida Mu and Nikolaos Aletras. 2020. **Identifying twitter users who repost unreliable news sources with linguistic information**. *PeerJ Computer Science*, 6:e325.
- Vaibhavi Nandagiri and Leena Philip. 2018. Impact of influencers from instagram and youtube on their followers. *International Journal of Multidisciplinary Research and Modern Education*, 4(1):61–65.
- Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. **BERTtweet: A pre-trained language model for English tweets**. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14, Online. Association for Computational Linguistics.
- Tiago Oliveira, Benedita Araujo, and Carlos Tam. 2020. Why do people share their travel experiences on social media? *Tourism Management*, 78:104041.
- Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. **GloVe: Global vectors for word representation**. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.
- Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2016. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149.
- Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252.
- Danae Sánchez Villegas and Nikolaos Aletras. 2021. **Point-of-interest type prediction using text and images**. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7785–7797, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Danae Sánchez Villegas, Saeid Mokaram, and Nikolaos Aletras. 2021. **Analyzing online political advertisements**. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3669–3680, Online. Association for Computational Linguistics.
- Danae Sánchez Villegas, Daniel Preotiuc-Pietro, and Nikolaos Aletras. 2020. **Point-of-interest type inference from social media text**. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 804–810, Suzhou, China. Association for Computational Linguistics.
- H. Andrew Schwartz, Salvatore Giorgi, Maarten Sap, Patrick Crutchley, Lyle Ungar, and Johannes Eichstaedt. 2017. **DLATK: Differential language analysis ToolKit**. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 55–60, Copenhagen, Denmark. Association for Computational Linguistics.
- Yanchuan Sim, Bryan Routledge, and Noah A. Smith. 2016. **Friends with motives: Using text to infer influence on SCOTUS**. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1724–1733, Austin, Texas. Association for Computational Linguistics.
- Hao Tan and Mohit Bansal. 2019. **LXMERT: Learning cross-modality encoder representations from transformers**. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages

- 5100–5111, Hong Kong, China. Association for Computational Linguistics.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.
- Alakananda Vempala and Daniel Preoțiuc-Pietro. 2019. [Categorizing and inferring the relationship between the text and image of Twitter posts](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2830–2840, Florence, Italy. Association for Computational Linguistics.
- Yue Wang, Jing Li, Michael Lyu, and Irwin King. 2020. [Cross-media keyphrase prediction: A unified framework with multi-modality multi-head attention and image wordings](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3311–3324, Online. Association for Computational Linguistics.
- Philip Weber, Thomas Ludwig, Sabrina Brodesser, and Laura Grönewald. 2021. “it’s a kind of art!”: Understanding food influencers as influential content creators. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–14.
- Bartosz W Wojdyski. 2016. Native advertising: Engagement, deception, and implications for theory. *The new advertising: Branding, content and consumer relationships in a data-driven social media era*, pages 203–236.
- Bartosz W Wojdyski, Nathaniel J Evans, and Mariea Grubbs Hoy. 2018. Measuring sponsorship transparency in the age of native advertising. *Journal of Consumer Affairs*, 52(1):115–137.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface’s transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.
- Xiao Yang, Seungbae Kim, and Yizhou Sun. 2019. How do influencers mention brands in social media? sponsorship prediction of instagram posts. In *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 101–104.
- Koosha Zarei, Damilola Ibosiola, Reza Farahbakhsh, Zafar Gilani, Kiran Garimella, Noël Crespi, and Gareth Tyson. 2020. Characterising and detecting sponsored influencer posts on instagram. In *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 327–331. IEEE.
- Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. 2020. [Unified Vision-Language Pre-Training for Image Captioning and VQA](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):13041–13049.

A Influencer Marketing Keywords

We extract keywords strongly associated with influencer marketing from the guidelines provided by the Federal Trade Commission (FTC, 2019) in the US, and the Advertisements Standards Authority and Competition and Markets Authority in the UK (CMA, 2020). The keywords in these guidelines are based on regulatory standards for digital enforcement which are meant to create objective and transparent expectations regarding the disclosure of *native advertising* on social media. Thus, our list of keywords include sponsorship disclosure terms that are relevant to different business models (i.e., market practices based on the obligations of the parties). A complete list of keywords is presented in Table 7.

B Annotation Guidelines

Purpose of the study This annotation effort is part of a study that aims to characterize and identify commercial content on Twitter. Commercial content is an umbrella term for communications that relate to commercial transactions, or in other words, content that is monetized. For influencers, that may entail various business models:

- Endorsements: an influencer receives money in order to promote a product or service.
- Affiliate marketing: the influencer is paid a percentage of referral sales, often identified through discount codes.
- Barter: exchange of goods or services from a brand or its representatives against an advertising service offered by the influencer.
- Direct selling: influencers can also choose to create their own products, branded products, and/or services, and link to their web shops.

Task Description The task is to annotate whether a given influencer's Twitter post is perceived to contain commercial content or not given only its text and image content (if available). If annotators perceive that the tweet contains commercial content, then it should be annotated as commercial, otherwise as non-commercial. If it is not clear whether the Tweet is perceived to contain commercial content, it should be labeled as unclear. The details of each category are as follows:

- Commercial: posts refer to any of the business models mentioned above. This category includes promoting or endorsing a brand or its products/services, a free loan of a product/service, a free product/service (whether requested or received out of the blue), or any other incentive. This can be noted by the use of terms or hashtags such as #gifted, #ad, @mentions of the brand, hashtags including the name of the brand and/or campaign slogans.
- Non-Commercial: Organic content such as personal ideas, personal comments and life updates, and that does not seem monetized through any of the business models mentioned above.
- Unclear: This option should be chosen when it is not clear whether the Tweet contains commercial content or not (e.g., commenting about a brand without using hashtags or @mentioning the brand).

Instructions

1. For each post, read the text, look at the image (if available), and select one of the categories (Commercial, Non-commercial, Unclear).
2. If the post is annotated as Commercial, then in the "Brand Cues" section write down the term(s) or hashtag(s) that support your decision such as: #gifted, #ad, @mentions, hashtags including the name of the brand and/or campaign slogans. Use the "Brand Cues" column that corresponds to the location of them: "Brand Cues Text" if the brand cues are found in the text and/or "Brand Cues Image" if they are located in the image. Select the option(s) (Text, Image) used to make your annotation (e.g., if the brand cues are in the text then select Text, if the post was annotated as non-commercial choose the option that you looked at to make your decision).
3. If the post was annotated as Unclear, then: select the "Other" option and click on the Tweet Link. If you find any brand cues in the Tweet's page, write them down in the column "Brand cues Other". If it is still unclear whether the Tweet is commercial or not keep the label "Unclear", otherwise select the appropriate label (Commercial/Non-Commercial).

Type	Description	Commercial Keywords
Guidelines	Keywords retrieved from relevant guidelines Recommended and not recommended terms.	#ad, ad, #advert #collab, collab, #spon, #sponsored, spon, #sp, sponsored, 'thanks to' / 'funded by' / 'supported by' / 'in association with' @USER
Endorsements	An influencer receives money to promote a product or service.	#ambassador, ambassador
Barter	Exchange of goods or services from a brand or its representatives against an advertising service offered by the influencer.	#gift, gift, #giveaway, giveaway unpaid sample
Affiliate Marketing	The influencer is paid a percentage of referral sales, often identified through discount codes.	#aff, aff, #affiliate, affiliate, discount code

Table 7: Commercial keywords. @USER refers to an @-mention of a brand account.

Dataset	No. of Commercial Keywords	Commercial Keywords
Han et al. (2021)	0	-
Zarei et al. (2020)	7	#ad, #advert, #sponsored #advertising, #giveaway, #spon, #sponsor
Yang et al. (2019)	3	#ad, #sponsored, #paidAD
Kim et al. (2021b)	3	#ad, #sponsored, #paidA
Kim et al. (2020)	1	#ad
MICD (Ours)	26	#ad, ad, #advert, #sponsored, #collab, collab, spon, #sp, sponsored, #aff, aff, 'thanks to' / 'funded by' /, unpaid sample, 'supported by' / 'in association with' @USER, #ambassador, ambassador, discount code #gift, gift, #giveaway, giveaway, #spon #affiliate, affiliate,

Table 8: Comparison of commercial keywords used in existing datasets and in ours (MICD)


Tweet Text	Tweet Image	Class	Brand Cues Text	Brand Cues Image	Text	Image	Other	Tweet Link	Brand Cues Other
we had such an amazing experience driving the @USER tucson 2022 in tucson! we picked some features we love about it , all the new #design , #tech & #safety we know you'll love as well . for now , check it out : 📍👉 HTTPURL #hyundaitucson #boldchange #mediadrive		Commercial	#hyundaitucson		<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	LINK	

Figure 3: Example of Annotation

Annotator Details All annotators were senior law school students (third year bachelor and masters level) who study comparative and international law. The students have a background in law, which entails a good grasp of consumer protection disclosures. In addition, their profiles were also particularly interesting for annotation since they had spent 6 months of their study being trained under an extracurricular Influencer Law Clinic honors programme. The training consisted in multidisciplinary workshops and hands-on research on influencer-related legal topics. The annotators come from a wide range of socio-economic back-

grounds and are fluent in English. The majority of annotators are female. However, the emphasis in the annotation process has been on the understanding of market practices in the light of legal frameworks, which mitigates any potential gender imbalance in the annotator pool. All annotators expressed their written consent and were informed about how data would be used following ethics guidelines from our Institution.

C Predictive Performance

Table 9 and Table 10 present the macro-averaged results of *commercial* content prediction.

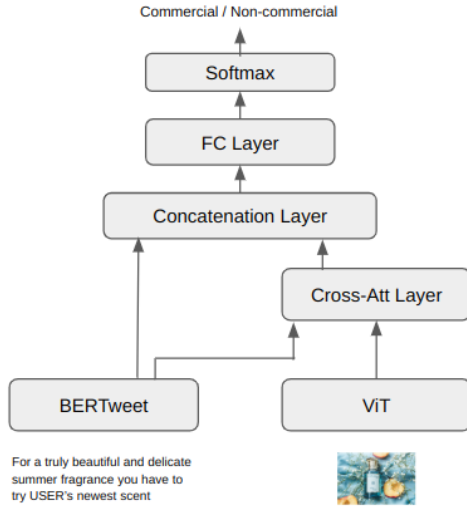


Figure 4: ViT-BERTweet-Att model for detecting *commercial* content. FC: fully-connected layer.

D Prompt Templates

D.1 Zero-shot Prompting

For zero-shot prompting we use the following prompt:

Label the next text as 'commercial' or 'not commercial'. Text: <TWEET>.

We map responses to the corresponding *commercial* or *non-commercial* class and report results for each model.

D.2 Few-shot Prompting

We experiment with few-shot prompting by appending four randomly selected training examples (two examples from each class) before each prompt. We run this three times with a different set examples. Table 4 shows average and standard deviation performance. The few-shot prompt follows the next template:

Label the next text as 'commercial' or 'not commercial'. Text: <TWEET-TRAIN> // <LABEL-TRAIN>

Label the next text as 'commercial' or 'not commercial'. Text: <TWEET-TRAIN> // <LABEL-TRAIN>

Label the next text as 'commercial' or 'not commercial'. Text: <TWEET-TRAIN> // <LABEL-TRAIN>

Label the next text as 'commercial' or 'not commercial'. Text: <TWEET-TRAIN> // <LABEL-TRAIN>

Label the next text as 'commercial' or 'not commercial'. Text: <TWEET> //

Model	F1	P	R
Most Freq.	32.44 _{0.0}	24.01 _{0.0}	50.00 _{0.0}
Prompting			
Flan-T5 (zero-shot)	43.90 _{0.0}	71.20 _{0.0}	55.25 _{0.0}
Flan-T5 (few-shot)	32.91 _{1.0}	41.14 _{0.6}	36.53 _{0.4}
GPT-3 (zero-shot)	63.65 _{0.0}	65.76 _{0.0}	64.20 _{0.0}
GPT-3 (few-shot)	69.32 _{1.7}	72.12 _{2.2}	70.24 _{0.4}
Image-only			
ResNet	59.60 _{0.5}	59.75 _{0.5}	59.73 _{0.5}
ViT	60.96 _{1.2}	61.62 _{0.7}	61.35 _{0.8}
Text-only			
BiLSTM-Att* (Zarei et al., 2020)	66.10 _{0.7}	66.37 _{0.7}	66.27 _{0.7}
BERT	74.35 _{0.6}	74.84 _{0.6}	74.61 _{0.6}
BERTweet	76.68 _{0.7}	76.86 _{0.5}	76.76 _{0.6}
Text & Image			
ViLT	68.44 _{0.8}	68.65 _{0.6}	68.55 _{0.7}
LXMERT	66.10 _{0.7}	66.37 _{0.7}	66.27 _{0.7}
MMBT	73.38 _{0.6}	73.89 _{0.6}	73.46 _{0.7}
Aspect-Att* (Kim et al., 2021b)	75.52 _{0.8}	77.13 _{1.1}	75.80 _{1.0}
ViT-BERTweet-Att (Ours)	77.75 _{0.5}	78.60 _{0.2}	77.97 _{0.1}

Table 9: Macro F1-Score, precision (P) and recall (R) for commercial influencer content prediction. * denotes current state-of-the-art models for influencer commercial content detection. Subscripts denote standard deviations. Best results are in bold.

Model	F1	P	R
Most Freq.	46.04 _{0.0}	42.66 _{0.0}	50.00 _{0.0}
Flan-T5 (zero-shot)	55.43 _{0.0}	65.60 _{0.0}	54.81 _{0.0}
Flan-T5 (few-shot)	40.77 _{1.1}	43.68 _{0.4}	39.52 _{1.1}
GPT-3 (zero-shot)	63.96 _{0.0}	63.38 _{0.0}	73.64 _{0.0}
GPT-3 (few-shot)	70.95 _{0.7}	74.81 _{6.4}	69.82 _{4.4}
BERTweet	76.48 _{1.3}	74.41 _{2.0}	79.66 _{0.4}
ViT-BERTweet-Att (Ours)	77.69 _{0.1}	77.41 _{0.7}	78.00 _{0.6}

Table 10: Macro F1-Score, precision (P) and recall (R) for commercial influencer content prediction for tweets containing text only. Subscripts denote standard deviations. Best results are in bold.

<Label-TRAIN> corresponds to the true label of the <TWEET-TRAIN> training example (*commercial* or *non-commercial*), <TWEET> refers to a testing example. We remove punctuation and spaces and map the output of each model (FLAN-T5 or GPT-3) to the corresponding label (*commercial* or *non-commercial*).