The University of Sheffield

Beyond Words: Analyzing Social Media with Text and Images

A thesis submitted for the degree of Doctor of Philosophy

in the

Department of Computer Science

Danae Sánchez Villegas

Supervisor: Professor Nikolaos Aletras

December 2023

Declaration

I, Danae Sánchez Villegas, hereby declare that this thesis has been composed solely by myself and that it has not been submitted, in whole or in part, in any previous application for a degree. Except where states otherwise by reference or acknowledgement, the work presented is entirely my own.

Acknowledgments

First, I would like to thank my supervisor, Professor Nikolaos Aletras, for his encouragement, guidance and support, throughout my PhD journey. Nikos has helped me reach academic, and professional goals I never thought I could accomplish. It has been an honor to collaborate with him over the years.

I would also like to thank my examiners Professor Andreas Vlachos and Dr. Carolina Scarton. I extend my gratitude to the panel committee members, Dr. Chenghua Lin, Dr. Kevin Li Sun, and Dr. Diana Maynard for their advice during my initial stages of study. Additionally, I would like to thank the academics in the Computer Science Department: Professor Rob Gaizauskas, Professor Heidi Christensen and Dr. Löic Barrault, for their guidance during my PhD journey.

My gratitude also extends to my research project collaborators: Dr. Daniel Preoţiuc-Pietro, Dr. Catalina Goanta and Dr. Saeid Mokaram. Their advice, ideas, and professional domain expertise motivate me to become a better researcher.

Many thanks to all my colleagues in the Natural Language Processing group for creating such a pleasant working atmosphere: Dr. George Chrysostomou, Katerina Margatina, Dr. Mali Jin, Dr. Yida Mu, Ahmed Alajrami, Constantinos Karouzos, Huiyin Xue, Dr. Samuel Mensah, Dr. Cass Zhao, Iknoor Singh, Varvara Papazoglou, Dr. Hardy Hardy, Dr. Zeerak Talat, Dr. Xutan Peng, Dr. Harish Tayyar Madabushi and Dr. Fernando Alva-Manchego; as well as CDT members: Jasivan Sivakumar, Tom Green, Tomas Goldsack, Edward Gow-Smith, Cliodhna Hughes and Robbie Sutherland; and friends: Panayiotis Karachristou, Emily Lau and Mariya Hendriksen, I value the insightful conversations and after-work events.

Finally, I would like to thank my family, especially Aleida Sánchez, my grandparents: Elsa, Delia and Isidro, and my deepest thanks to my parents, Vero and David, for their love and support throughout my life.

This work is 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.

Abstract

People express their opinions and experiences through text and images in social media platforms. Analyzing social media content has several applications in natural language processing such as sentiment analysis, hate speech detection, fact checking and sarcasm detection. Combining text and images from social media posts is challenging due to weak visual-text relationships. For instance, a post with the text: *Feeling on top of the world after acing my final exams!* and a picture of a group of friends at the beach. The image and the text are weakly related as the image does not directly align with the academic context, potentially leading to confusion or misinterpretation of the intended message. Thus, effectively modeling text and images from social media posts is crucial for advancing natural language understanding. This thesis proposes a number of new challenging multimodal classification tasks: point-of-interest (POI) type prediction, political advertisements analysis, and influencer content analysis. First, we introduce POI type prediction which consists of inferring the type of location from which a social media message was posted such as a park or a restaurant. This task is relevant to study a place's identity and has applications such as POI visualization and recommendation. Second, we analyze political advertisements by introducing two new datasets containing political ads labeled by the sponsor's ideology (*conservative*, *liberal*), and the sponsor type (*political party*, *third party*); and we experiment with multimodal models for advertisement classification. Analyzing political ads is important for researching the characteristics of online campaigns (e.g. voter targeting, non-party campaigns and misinformation) on a large scale. Next, we perform an extensive analysis of influencer content including multimodal approaches for identifying commercial posts, i.e., content that is monetized. Automatically detecting influencer commercial posts is of utmost importance for addressing issues related to transparency and regulatory compliance, such as misleading advertising. Finally, this thesis also presents novel methods for tackling the challenges of modeling text and visual content in social media. We propose two auxiliary losses, Image-Text Contrastive which encourages the model to capture the underlying dependencies in multimodal posts; and Image-Text Matching to enable visual and language alignment.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Social media platforms serve as hubs for users to share their thoughts, opinions, and experiences. Analyzing social media content has different applications in natural language processing (NLP) such as sentiment analysis (Nguyen and Shirai, 2015; Chambers et al., 2015; Nakov et al., 2016; Kruspe et al., 2020; Barbieri et al., 2022), rumor detection and fact checking (Ma et al., 2018; Thorne and Vlachos, 2018; Li et al., 2019b; Zhou et al., 2019; Li and Scarton, 2020; Santos et al., 2020; Tian et al., 2022; Guo et al., 2022), and political discourse and biased language analysis (Johnson and Goldwasser, 2018; Huguet Cabot et al., 2020; Mendelsohn et al., 2021; De Kock and Vlachos, 2022). Multimodal posts, consisting of images and text, offer a creative and engaging means of communication for users and enrich the narrative for readers. Furthermore, they highlight the necessity of automated vision and language understanding in addressing diverse multimodal classification tasks.

Combining text and images has been largely studied for modeling vision-and-language tasks such as visual question answering (Antol et al., 2015; Fukui et al., 2016; Ray et al., 2019; Si et al., 2022) and image captioning (Devlin et al., 2015; Johnson et al., 2016; Li et al., 2020b, 2022b; Dai et al., 2023) where strong image-text connections are assumed, i.e., captions that explicitly describe a corresponding image (Hessel and Lee, 2020; Xu and Li, 2022). However, modeling text-image pairs from social media posts presents additional challenges. For instance, capturing cross-modal semantics that are not immediately apparent is difficult (Vempala and Preoţiuc-Pietro, 2019). Figure 1.1 (top) shows an example where the text refers specifically to the mood of the person in the photo (i.e., "unhappy feeling" *when @USER gets more followers...*). Moreover, cases when the visuals are weakly related to the text are also common (Xu et al., 2022). For example, Figure 1.1 (bottom) shows an

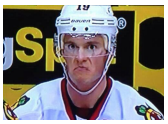
Image	Text (Post)	Image-Text Relation in Post	Image Caption
	When @USER gets more followers than you in 12 hours	The image complements the text to provide meaning of the post	A close up of a hockey player wearing a helmet
	My baby approves	The image does not add to the meaning of the post and the text does not provide a description of the image	A gray and white chicken standing in the dirt

Figure 1.1: Examples of image-text relations in social media posts from Vempala and Preoţiuc-Pietro (2019) and corresponding image captions automatically generated with InstructBLIP (Dai et al., 2023). While image captions have a clear visual-language connection, image-text relationships in social media posts may no be apparent.

image of a hen accompanied by the text *My baby approves*. Without any further background, it is difficult to draw a straight relationship between the two. Another challenge that arises when modeling multimodal posts, is that the image type distribution is diverse. Image types include screenshots, natural photos, posters, and drawing pictures (Wang et al., 2020); and they may contain text wordings which have proved to be beneficial to model tasks such as inferring the topic (e.g. cars, electronics) and sentiment of online commercial advertisements (Hussain et al., 2017; Kalra et al., 2020) and identifying hateful messages (Pramanick et al., 2021; Cao et al., 2022).

Effectively modeling textual and visual information is crucial to natural language understanding as incorporating both modalities enhances the understanding of the user's intentions, emotions, and opinions. For instance, it can aid in disambiguating the intended meaning, as images often provide visual context that clarifies the text's tone and intent. Visual context can also help handling noisy textual data (e.g., abbreviations and typos), predominant in social media, by providing additional background. Content of both text and images has been widely used to improve upon single modality results in various downstream tasks such as sentiment analysis (Niu et al., 2016; Ju et al., 2021), hate speech detection (Botelho et al., 2021; Hossain et al., 2022; Cao et al., 2022), sarcasm detection (Cai et al., 2019; Xu et al., 2020; Liang et al., 2022), and named entity recognition (Moon et al., 2018b; Sun et al., 2020).

In this thesis, we focus on three under-explored multimodal classification tasks: point-of-interest (POI) type prediction, political advertisements analysis, and influencer content

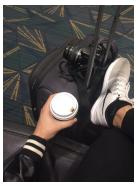
Task	POI Type Prediction	Online Political Ads Analysis	Influencer Content Analysis
Image			
Text	Next stop: NYC ✈️	WE CAN'T LET JOE BIDEN WIN! VOTE EARLY	Cherry tree hill is hands down the best view in #Barbados. #VisitBarbados
Description	Users share content that is relevant to their experiences and feelings in the location.	Example of an online political advertisement and corresponding text.	Influencers share commercial and non-commercial content in social media.

Figure 1.2: Representative examples of text and image content of social media posts for three tasks: POI type prediction, online political advertisements analysis and influencer content analysis.

analysis. Moreover, we aim to study the intricate relationships between text and images in social media posts, delving into how they complement, reinforce, or even contradict each other to convey complex messages.

POI type prediction The content of social media posts shared by users from specific places such as restaurants, shops, and parks, contributes to shaping a place's identity, by offering information about feelings elicited by participating in an activity or living an experience in that place. For instance, Figure 1.2 (second column) shows an example of a post sent from a specific place or POI. It consists of the text *Next stop: NYC* along with a picture of descriptive items that people carry at an airport such as luggage, a camera and a takeaway coffee cup. In this thesis we introduce the task of POI type prediction defined as a classification task where given the content of a post, the goal is to classify it in one of the POI categories. Inferring the type of place from a user's post using text and visual information, is useful for cultural geographers to study a place's identity (Tuan, 1991) and has downstream geosocial applications such as POI visualization (McKenzie et al., 2015; Yaqub et al., 2020; McKitrick et al., 2023) and recommendation (Alazzawi et al., 2012; Yuan et al., 2013; Preoţiuc-Pietro and Cohn, 2013; Gao et al., 2015; Zeng et al., 2020; Yang et al., 2023).

Online Political Advertising Analysis Political advertising is defined as *'any controlled message communicated through any channel designed to promote the political interests of*

individuals, parties, groups, government, or other organizations' (Kaid and Holtz-Bacha, 2006). Figure 1.2 (third column) shows an example of an online political ad. Automatically analyzing political ads is important in political science for researching the characteristics of online campaigns (e.g. voter targeting, sponsors, non-party campaigns, privacy, and misinformation) on a large scale (Scammell and Langer, 2006; Johansson and Holtz-Bacha, 2019; Biamby et al., 2022). Moreover, computational methods for political ads analysis can help linguists to study features of political discourse and communication (Kenzhekanova, 2015; Skorupa and Dubovičienė, 2015; Mancini et al., 2022).

Influencer Content Analysis Social media influencers are content creators who have established credibility in a specific domain (e.g., fitness, technology), are 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 in social media) has grown in popularity as an alternative to traditional advertising (e.g., magazines, television, billboards) and mainstream digital marketing such as pop-up and platform ads for reaching a larger and more targeted audience (Leerssen et al., 2019; Nandagiri and Philip, 2018; Gross and Wangenheim, 2018; Lou et al., 2019; Jarrar et al., 2020; Fang and Wang, 2022).

Figure 1.2 (fourth column) shows an example of a *commercial* post, i.e., content that is monetized. However, this post rather than promoting a specific product as they normally do, it contains a description of their "personal" experience. This type of posts are also common in *non-commercial* posts (Oliveira et al., 2020) making 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; Wojdynski et al., 2018; Ducato, 2020; Ershov and Mitchell, 2020).

1.1 Research Aims and Objectives

This thesis focuses on modeling three under-explored multimodal social media tasks namely point-of-interest type prediction, online political advertisements analysis and influencer content analysis using machine learning methods. We aim to achieve the next research objectives:

- Previous work on social media analysis related to POIs, is mainly focused on the popular task of geolocation prediction (Cheng et al., 2010; Eisenstein et al., 2010; Han et al., 2012; Roller et al., 2012; Rahimi et al., 2015; Dredze et al., 2016; Mishra, 2020; Khanal et al., 2022), which consists of inferring the exact geographical location of a post using language variation and geographical cues. However, inferring the place's type using textual and visual information to uncover the geographic agnostic features associated with locations of different types has yet to be studied. Therefore, *we aim to develop new data resources and models for studying POI type prediction at a large scale in computational social science.*

- Previous work in NLP related to online advertising, has explored tasks such as predicting the category (e.g. politics, cars, electronics) and sentiment of an ad in the *commercial* domain (Hussain et al., 2017; Kalra et al., 2020). Moreover, large-scale studies of online political advertising have so far focused on understanding targeting strategies rather than developing predictive models for analyzing its content (Edelson et al., 2019; Medina Serrano et al., 2020). *By conducting a systematic study of online political ads consisting of text and images, the aim is to uncover linguistic and visual cues across political ideologies (liberal or conservative) and sponsor types (political party or third-party) using computational methods.*

- 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. Therefore, *we aim to develop new expert annotated data as well as an extensive empirical study on influencer multimodal content focused on analyzing the contribution of text and image modalities to commercial and non-commercial posts.*

- Previous work on multimodal social media analysis has shown that combining text and image information is challenging because of the idiosyncratic cross-modal semantics with hidden or complementary information present in matching image-text pairs (Vempala and Preoţiuc-Pietro, 2019; Kruk et al., 2019; Xu et al., 2022). In this thesis, *we aim to directly model this by proposing auxiliary losses that can be used jointly with any downstream classification task when fine-tuning pre-trained multimodal models.*

Additionally, as we explore these multimodal social media tasks, this thesis seeks to study the following research questions:

- **Q1: What are the various methodologies available for extracting visual information from social media posts, and how can these methodologies be effectively used to enhance classification models?**
- **Q2: Can pre-trained multimodal models be directly applied to classify social media posts, or how can these models be adapted to account for the unique characteristics of social media posts?**
- **Q3: To what extent does multimodal commercial content exist in social media beyond traditional forms of paid product advertising? Moreover, how transparent are these types of advertising to social media users?**

1.2 Thesis Overview: Publications and Contributions

This section lists the contributions made throughout this thesis. It follows a *thesis by publications* format and consists of a collection of five papers where each paper corresponds to an individual chapter.

Publication I: *Point-of-Interest Type Inference from Social Media Text* In this paper, we conduct an analysis to uncover linguistic features specific to place types and train predictive models to infer the place or POI category using text and posting time. The contributions of this publication are as follows:

- We provide the first study of POI type prediction in computational linguistics.
- A large dataset made out of tweets (text and posting time) linked to particular POI categories is developed and made publicly available.
- We provide a linguistic and temporal analyses related to the place the text was posted from.
- Predictive models using text and temporal information.

This work has been published 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 (AACL 2020) (Sánchez Villegas et al., 2020).

My contributions to the work: conceptualization, data collection, methodology, software, validation and writing.

Publication II: *Point-of-Interest Type Prediction using Text and Images* In this work, we enrich the dataset developed in Publication I with images and we propose a multimodal model to tackle the task of POI type prediction as a multimodal classification task. The contributions of this work are as follows:

- We enrich our dataset introduced in Publication I with images.
- We propose a multimodal model that combines text and images in two levels using: (i) a modality gate to control the amount of information needed from the text and image; (ii) a cross-attention mechanism to learn cross-modal interactions.
- We provide an in-depth analysis to uncover the limitations of our model and uncover cross-modal characteristics of POI types.

This work has been published in Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP 2021) (Sánchez Villegas and Aletras, 2021).

My contributions to the work: conceptualization, data collection, methodology, software, validation and writing.

Publication III: *Analyzing Online Political Advertisements* This paper presents the first study in NLP for analyzing the language of political ads. We define two tasks as advertisement-level binary classification tasks and evaluate a variety of approaches, including textual, visual and multimodal models. The contributions of this work are as follows:

- A new classification task for predicting the political ideology (conservative or liberal) of an ad. We collect 5,548 distinct political ads in English from 242 different advertisers in the U.S., and label them according to the dominant political ideology of the respective sponsor's party affiliation (*Liberal* or *Conservative*).
- A new classification task to automatically classify ads that were sponsored by official political parties and third-party organizations, such as businesses and non-profit organizations. For this task, we extract 15,116 advertisements in English from 665 distinct

advertisers in the U.S., and label them as *Political Party* (i.e. officially registered) and *Third-Party* (i.e. other organizations) following Fowler et al. (2020b).

- Experiments with text-based and multimodal (text and images) models for political ideology prediction and sponsor type classification.
- Analysis of textual and visual features of online political ads and error analysis to understand model limitations.

This work has been published in Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021 (ACL Findings 2021) (Sánchez Villegas et al., 2021).

My contributions to the work: conceptualization, data collection, methodology, software, validation and writing.

Publication IV: *A Multimodal Analysis of Influencer Content on Twitter* In this paper, we conduct an empirical study of influencer content. We introduce 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. In this publication we also experiment with an extensive set of predictive models that combine text and visual information and conduct a thorough analysis of strengths and limitations of our models. The contributions of this work 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;
- 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.

This work has been published in the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association

for Computational Linguistics (AACL 2023) ***Area Chair Award (Society and NLP)*** (Sánchez Villegas et al., 2023).

My contributions to the work: conceptualization, data collection, methodology, software, validation and writing.

Publication V: ***Improving Multimodal Classification of Social Media Posts by Leveraging Image-Text Auxiliary Tasks*** In this work, we propose the use of two auxiliary losses when fine-tuning pre-trained multimodal models for social media classification: Image-Text Contrastive (ITC) which encourages the model to capture the underlying dependencies in multimodal posts; and Image-Text Matching (ITM) to improve visual and language alignment. Our results show consistent improvement in predictive performance upon the inclusion of these objectives on four different tasks. The contributions of this work are as follows:

- We present an extensive study on comparing multimodal models jointly fine-tuned with ITC and ITM.
- We show that models using ITC and ITM as auxiliary losses consistently improve their performance on four popular multimodal social media classification datasets.
- We provide a comprehensive analysis that sheds light on the effectiveness of each auxiliary task and their combination.

This work is under review in ACL Rolling Review (ARR).

My contributions to the work: conceptualization, methodology, software, validation and writing.

Conclusions In Chapter 7 we summarize the findings and contributions of this thesis and suggest possible research directions for future work.

Chapter 2

Publication I: Point-of-Interest Type Inference from Social Media Text

Point-of-Interest Type Inference from Social Media Text

Danae Sánchez Villegas[α], *Daniel Preoţiuc-Pietro*[β], *Nikolaos Aletras*[α]

[α] Computer Science Department, University of Sheffield, UK

[β] Bloomberg

Abstract

Physical places help shape how we perceive the experiences we have there. We study the relationship between social media text and the type of the place from where it was posted, whether a park, restaurant, or someplace else. To facilitate this, we introduce a novel data set of ∼200,000 English tweets published from 2,761 different points-of-interest in the U.S., enriched with place type information. We train classifiers to predict the type of the location a tweet was sent from that reach a macro F1 of 43.67 across eight classes and uncover the linguistic markers associated with each type of place. The ability to predict semantic place information from a tweet has applications in recommendation systems, personalization services and cultural geography.[1]

[1]Data is available here: https://archive.org/details/poi-data

2.1 Introduction

Social networks such as Twitter allow users to share information about different aspects of their lives including feelings and experiences from places that they visit, from local restaurants to sport stadiums and parks. Feelings and emotions triggered by performing an activity or living an experience in a Point-of-Interest (POI) can give a glimpse of the atmosphere in that place (Tanasescu et al., 2013).

In particular, the language used in posts from POIs is an important component that contributes toward the place's identity and has been extensively studied in the context of social and cultural geography (Tuan, 1991; Scollon and Scollon, 2003; Benwell and Stokoe, 2006). Social media posts from a particular location are usually focused on the person posting the content, rather than on providing explicit information about the place. Table 2.1 displays example Twitter posts from different POIs. Users express their feelings related to a certain place ('this places gives me war flashbacks'), comments and thoughts associated with the place they are in ('few of us dressed appropriately') or activities they are performing ('leaving the news station', 'on the way to the APCE Annual').

In this paper, we aim to study the language that people on Twitter use to share information about a specific place they are visiting. Thus, we define the prediction of a POI type given a post (i.e. tweet) as a multi-class classification task using only information available at posting time. Given the text from a user's post, our goal is to predict the correct type of the location it was posted, e.g. park, bar or shop. Inferring the type of place from a user's post using linguistic information, is useful for cultural geographers to study a place's identity (Tuan, 1991) and has downstream geosocial applications such as POI visualisation (McKenzie et al., 2015) and recommendation (Alazzawi et al., 2012; Yuan et al., 2013; Preoţiuc-Pietro and Cohn, 2013; Gao et al., 2015).

Predicting the type of a POI is inherently different to predicting the POI type from comments or reviews. The role of the latter is to provide opinions or descriptions of the places, rather than the activities and feelings of the user posting the text (McKenzie et al., 2015), as illustrated in Table 2.1. This is also different, albeit related, to the popular task of geolocation prediction (Cheng et al., 2010; Eisenstein et al., 2010; Han et al., 2012; Roller et al., 2012; Rahimi et al., 2015; Dredze et al., 2016), as this aims to infer the exact geographical location of a post using language variation and geographical cues or GPS coordinates rather than inferring the place's type. Our task aims to uncover the geographic agnostic features associated with

POIs of different types. Moreover, while GPS provides crucial location information, extracting insights from social media content augments our understanding of POIs by capturing user experiences, sentiments, and contextual details that go beyond spatial coordinates.

Our contributions are as follows: (1) We provide the first study of POI type prediction in computational linguistics; (2) A large data set made out of tweets linked to particular POI categories; (3) Linguistic and temporal analyses related to the place the text was posted from; (4) Predictive models using text and temporal information reaching up to 43.67 F1 across eight different POI types.

2.2 Point-of-Interest Type Data

We define the POI type prediction as a multi-class classification task performed at the social media post level. Given a post T, defined as a sequence of tokens $T = \{t_1, ..., t_n\}$, the goal is to label T as one of the M POI categories. We create a novel data set for POI type prediction containing text and the location type it was posted from as, to the best of our knowledge, no such data set is available. We use Twitter as our data source because it contains a large variety of linguistic information such as expression of thoughts, opinions and emotions (Java et al., 2007; Kouloumpis et al., 2011).

2.2.1 Types of POIs

Foursquare is a location data platform that manages 'Places by Foursquare', a database of more than 105 million POIs worldwide. The place information includes verified metadata such as name, geo-coordinates and categories as well as other user-sourced metadata such as tags, comments or photos. POIs are organized into 9 top level primary categories with multiple subcategories. We only focus on 8 primary top-level POI categories since the category 'Residence' has a considerably smaller number of tweets compared to the other categories (0.78% tweets from the total). We leave finer-grained place category inference as well as using other metadata for future work since the scope of this work is to study the language of posts associated with semantic type places.

Category	Sample Tweet	Train	Dev	Test	Tokens
Arts & Entertainment	i'm back in central park . this place gives me war flashbacks now lol	40,417	4,755	5,284	14.41
College & University	currently visiting my dream school 😘 ♥	21,275	2,418	2,884	15.52
Food	Some Breakfast, it's only right! #LA	6,676	869	724	14.34
Great Outdoors	Sorry Southport, Billy is dishing out donuts at #donutfest today. See you next weekend!	27,763	4,173	3,653	13.49
Nightlife Spot	Chicago really needs to step up their Aloha shirt game. Only a few of us dressed "appropriately" tonight. :) 🗿🌴🌺	5,545	876	656	15.46
Professional & Other Places	Leaving the news station after a long day	30,640	3,381	3,762	16.46
Shop & Service	Came to get an old fashioned tape measures and a button for my coat	8,285	886	812	15.31
Travel & Transport	Shoutout to anyone currently on the way to the APCE Annual Event in Louisville, KY! #APCE2018	16,428	2,201	1,872	14.88

Table 2.1: Place categories with sample tweets and data set statistics.

2.2.2 Associating Tweets with POI Types

Twitter users can tag their tweets to the locations they are posted from by linking to Foursquare places.[2] In this way, we collect tweets assigned to the POIs and associated metadata (see Table 2.1). We select a broad range of locations for our experiments. There is no public list of all Foursquare locations that can be used through Twitter and can be programmatically accessed. Hence, in order to discover Foursquare places that are actually used in tweets, we start with all places found in a 1% sample of the Twitter feed between 31 July 2016 and 24 January 2017 leading us to a total of 9,125 different places. Then, we collect all tweets from these places between 17 August 2016 and 1 March 2018 using the Twitter Search API[3]. We collect the place metadata from the public Foursquare Venues API. This resulted in a total data set of 1,648,963 tweets tagged to a Foursquare place. In order to extract metadata about each location, we crawled the Twitter website to identify the corresponding Foursquare Place ID of each Twitter place. Then, we used the public Foursquare Venues API[4] to download all the place metadata.

2.2.3 Data Filtering

To limit variation in our data, we filter out all non-English tweets and non-US places, as these were very limited in number. We keep POIs with at least 20 tweets and randomly

[2] https://developer.foursquare.com/places
[3] https://developer.twitter.com/en/docs/tweets/search/guides/tweets-by-place
[4] https://developer.foursquare.com/overview/venues.html

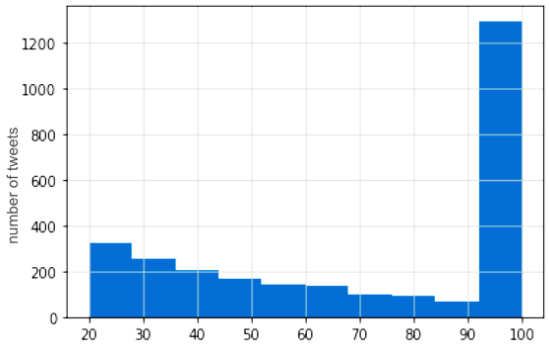


Figure 2.1: Distribution of POIs in our dataset.

subsample 100 tweets from POIs with more tweets to avoid skewing our data. Fig. 2.1 shows the distribution of POIs in our dataset. Our final data set consists of 196,235 tweets from 2,761 POIs.

2.2.4 Data Split

We create our data split at a location-level to ensure that our models are robust and generalize to locations held-out in training. We split the locations in train (80%), development (10%) and test (10%) sets and assign tweets to one of the three splits based on the location they were posted from (see Table 2.1 for detailed statistics).

2.2.5 Text Processing

We lower-case text and replace all URLs and mentions of users with placeholders. We preserve emoticons and punctuation and replace tokens that appear in less than five tweets with an 'unknown' token. We tokenize text using a Twitter-aware tokenizer (Schwartz et al., 2017).

2.3 Analysis

We first analyze our data set to understand the relationship between location type, language and posting time.

Arts		College		Food		Outdoors		Nightlife		Professional		Shop		Travel	
Feature	χ^2	Feature	χ^2	Feature	χ^2	Feature	χ^2	Feature	χ^2	Feature	χ^2	Feature	χ^2	Feature	χ^2
concert	167.20	campus	298.74	chicken	375.52	beach	591.81	#craftbeer	425.97	school	87.46	mall	462.03	airport	394.20
museum	152.14	college	266.63	#nola	340.64	🏄	239.00	🍻	311.68	students	79.93	store	403.00	🛫	343.30
show	134.39	university	155.65	lunch	255.98	hike	227.91	beer	203.57	grade	66.05	shopping	359.00	flight	292.94
night	104.48	class	112.23	fried	216.49	lake	193.58	bar	93.90	vote	65.80	shop	132.39	hotel	168.38
tonight	80.76	semester	103.19	dinner	203.65	park	165.92	🍺	67.00	our	63.12	🌹	126.07	conference	141.74
game	73.56	football	59.24	🍔	195.41	island	151.45	🍾	56.94	jv	60.64	🛍	95.32	landed	118.05
art	69.77	student	57.86	pizza	190.83	sunset	142.44	dj	56.56	church	52.97	apple	88.74	plane	88.42
USER	66.14	classes	57.37	shrimp	188.77	hiking	137.74	tonight	53.39	hs	50.63	market	76.60	bound	78.43
zoo	66.09	students	56.98	🍕	179.39	beautiful	109.45	ale	52.62	senior	50.05	auto	73.52	heading	62.09
baseball	62.90	camp	44.19	😊	151.00	bridge	108.56	party	51.14	ss	44.46	stock	72.31	headed	57.12

Table 2.2: Unigrams associated with each category, sorted by χ^2 value computed between the normalized frequency of each feature and the category label across all tweets in the training set ($p < 0.001$).

2.3.1 Linguistic Analysis

We analyze the linguistic features specific to each category by ranking unigrams that appear in at least 5 different locations, such that these are representative of the larger POI category rather than a few specific places. Features are normalized to sum up to unit for each tweet, then we compute the (Pearson) χ^2 coefficient independently between its distribution across posts and the binary category label of the post similar to the approach followed by Maronikolakis et al. (2020a) and Preoţiuc-Pietro et al. (2019a). Table 2.2 presents the top unigram features for each category.

We note that most top unigrams specific of a category naturally refer to types of places (e.g. 'campus', 'beach', 'mall', 'airport') that are part of that category. All categories also contain words that refer to activities that the poster of the tweet is performing or observing while at a location (e.g. 'camp' and 'football' for College, 'concert' and 'show' for Arts & Entertainment, 'party' for Nightlife Spot, 'landed' for Travel & Transport, 'hike' for Greater Outdoors). Nightlife Spot and Food categories are represented by types of food or drinks that are typically consumed at these locations. Beyond these typical associations, we highlight that usernames are more likely mentioned in the Arts & Entertainment category, usually indicating activities involving groups of users, emojis indicative of the user state (e.g. happy emoji in Food places) and adjectives indicative of the user's surroundings (e.g. 'beautiful' in Greater Outdoors places). Finally, we also uncover words indicative of the time the user is at a place, such as 'tonight' for Arts & Entertainment, 'sunset' for the Greater Outdoors and 'night' for Nightlife Spots and Arts & Entertainment.

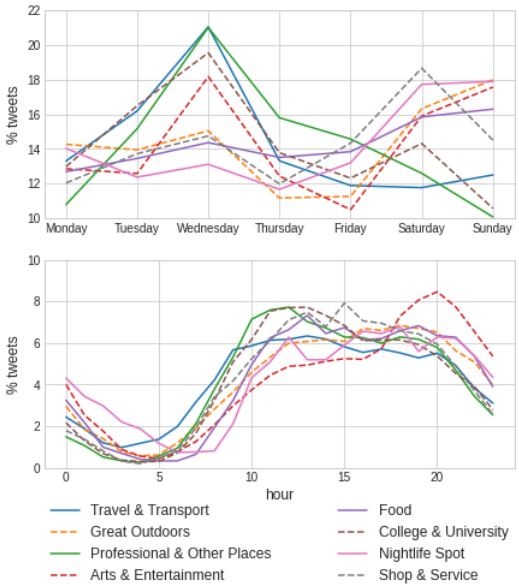


Figure 2.2: Percentage of tweets by day of week (top) and by hour of day (bottom).

2.3.2 Temporal Analysis

We further examine the relationship between the time a tweet was posted and the POI type it was posted from. Figure 2.2 shows the percentage of tweets by day of week (top) and hour of day (bottom).

We observe that tweets posted from the 'Professional & Other Places', 'Travel & Transport' and 'College & University' categories are more prevalent on weekdays, peaking on Wednesday, while on weekends more tweets are posted from the 'Great Outdoors', 'Arts & Entertainment', 'Nightlife & Spot' and 'Food' categories when people focus less on professional activities and dedicate more time to leisure as expected. The hour of day pattern follows the daily human activity rhythm, but the differences between categories are less prominent, perhaps with the exception of the 'Arts & Entertainment' category peaks around 8PM and 'Nightlife Spots' that see a higher percent of tweets in the early hours of the day (between 1-5am) than other categories.

2.4 Predicting POI Types of Tweets

2.4.1 Methods

Logistic Regression We first experiment with logistic regression using a standard bag of n-grams representation of the tweet (**LR-W**), including unigrams to trigrams weighted using TF-IDF. We identified in the analysis section that temporal information about the tweet may be useful for classification. Hence, to add temporal information extracted from a tweet, we create a 31-dimensional vector encoding the hour of the day and the day of the week it was sent from. We experiment with only using the temporal features (**LR-T**) and in combination with the text features (**LR-W+T**). We use L1 regularization (Hoerl and Kennard, 1970) with hyperparameter $\alpha = .01$ (selected based on dev set from {.001, .01, .1}).

BiLSTM We train models based on bidirectional Long-Short Term Memory (LSTM) networks (Hochreiter and Schmidhuber, 1997), which are popular in text classification tasks. Tokens in a tweet are mapped to embeddings and passed through the two LSTM networks, each processing the input in opposite directions. The outputs are concatenated and passed to the output layer using a softmax activation function (**BiLSTM**). We extend the BiLSTM to encode temporal one-hot representation by: (a) concatenating the temporal vector to the tweet representation (**BiLSTM-TC**); and (b) projecting the time vector into a dense representation using a fully connected layer which is added to the tweet representation before passing it through the output layer using a softmax activation function (**BiLSTM-TS**). We use 200-dimensional GloVe embeddings (Pennington et al., 2014) pre-trained on Twitter data. The maximum sequence length is set to 26, covering 95% of the tweets in the training set. The LSTM size is $h = 32$ where $h \in \{32, 64, 100, 300\}$ with dropout $d = 0.5$ where $d \in \{.2, .5\}$. We use Adam (Kingma and Ba, 2014) with default learning rate, minimizing cross-entropy using a batch size of 32 over 10 epochs with early stopping.

BERT Bidirectional Encoder Representations from Transformers (BERT) is a pre-trained language model based on transformer networks (Vaswani et al., 2017; Devlin et al., 2019). BERT consists of multiple multi-head attention layers to learn bidirectional embeddings for input tokens. The model is trained on masked language modeling, where a fraction of the input tokens in a given sequence is replaced with a mask token, and the model attempts to predict the masked tokens based on the context provided by the non-masked tokens in the

sequence. We fine-tune BERT for predicting the POI type of a tweet by adding a classification layer with softmax activation function on top of the Transformer output for the 'classification' $[CLS]$ token (**BERT**). Similarly to the previous models, we extend BERT to make use of the time vector in two ways, by concatenating (**BERT-TC**), and by adding it (**BERT-TS**) to the output of the Transformer before passing it to through the classification layer with softmax activation function. We use the base model (12-layer, 110M parameters) trained on lower-cased English text. We fine-tune it for 2 epochs with a learning rate $l = 2e^{-5}$, $l \in \{2e^{-5}, 3e^{-5}, 5e^{-5}\}$ and a batch size of 32.

2.4.2 Results

Table 2.3 presents the results of POI type prediction measured using accuracy, macro F1, precision and recall across three runs. In general, we observe that we can predict POI types of tweets with good accuracy, considering the classification is across eight relatively well balanced classes.

Best results are obtained using BERT-based models (BERT, BERT-TC and BERT-TS), with the highest accuracy of 49.17 (compared to 26.89 majority class) and highest macro-F1 of 43.67 (compared to 12.64 random). We observe that BERT models outperform both BiLSTM and linear methods across all metrics, with over 4% improvement in accuracy and 5 points F1. The BiLSTM models perform marginally better than the linear models. Temporal features alone are marginally useful when models are evaluated using accuracy (+0.28 BERT, +0.34 for BiLSTMs, +0.69 for LR) and perform similarly on F1, with the notable exception of the BiLSTM models. We find that adding these features is more beneficial than concatenating them, with concatenation hurting performance on accuracy for both BiLSTM and BERT.

Figure 2.3 shows the confusion matrix of our best performing model, BERT, according to the macro-F1 score. The confusion matrix is normalized over the actual values (rows). The category 'Arts & Entertainment' has the greatest percentage (62%) of correctly classified tweets, followed by the 'Great Outdoors' category with 54%, and the 'College & University' category with 44%. On the other hand, the categories 'Nightlife Spot' and 'Shop & Service' have the lowest results, where 30% of the tweets predicted as each of these classes is correctly classified. Most common error is when the model classifies tweets from the category 'College & University' as 'Professional & Other Places', as tweets from these places contain similar terms such as 'students' or 'class'. It is important to note the distinction between these two classes. While 'Professional & Other Places' encompasses a broad spectrum of locations,

Model	Acc	F1	P	R
Major. Class	26.89	5.30	3.36	12.50
Random	13.63	12.64	13.63	15.68
LR-T	27.93	14.01	15.78	16.06
LR-W	43.04	37.33	37.06	38.03
LR-W+T	43.73	37.83	37.68	38.37
BiLSTM	44.38	35.77	45.29	33.78
BiLSTM-TC	44.01	38.07	41.51	36.46
BiLSTM-TS	44.72	38.26	42.91	36.30
BERT	48.89	**43.67**	**48.44**	**41.33**
BERT-TC	46.13	41.19	46.81	39.03
BERT-TS	**49.17**	43.47	48.40	41.26

Table 2.3: Accuracy (Acc), Macro-F1 Score (F1), Precision macro (P), and Recall macro (R) for POI type prediction (all std. dev < 0.01). Best results are in bold.

including convention centers, libraries, offices, spiritual centers, and schools, 'College & University' is specifically confined to colleges and universities. As part of future enhancements, it is recommended to introduce a distinct category for educational establishments. This expanded category would not only encompass colleges and universities but also include schools, addressing the need for a more nuanced classification system.

2.5 Conclusion

We presented the first study on predicting the POI type a social media message was posted from and developed a large-scale data set with tweets mapped to their POI category. We conducted an analysis to uncover features specific to place type and trained predictive models to infer the POI category using only tweet text and posting time with accuracy close to 50% across eight categories. Future work will focus on using other modalities such as network (Aletras and Chamberlain, 2018; Tsakalidis et al., 2018) or image information (Vempala and Preoţiuc-Pietro, 2019; Alikhani et al., 2019) and prediction at a more granular level of POI types.

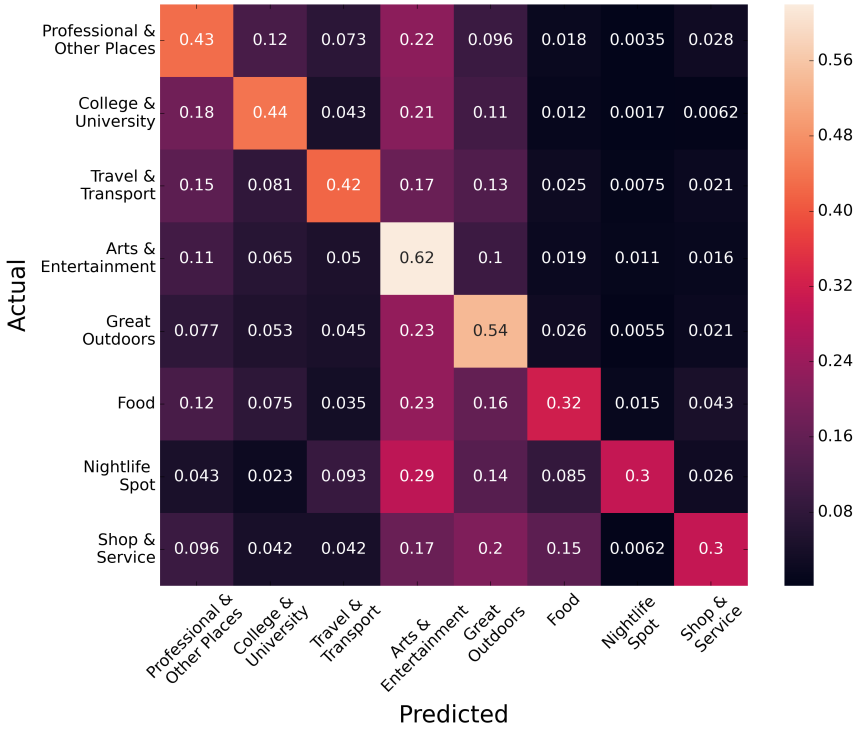


Figure 2.3: Confusion Matrix of the best performing model (BERT).

Acknowledgments

DSV is 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. NA is supported by ESRC grant ES/T012714/1.

Chapter 3

Publication II: Point-of-Interest Type Prediction using Text and Images

Point-of-Interest Type Prediction using Text and Images

Danae Sánchez Villegas, Nikolaos Aletras
Computer Science Department, University of Sheffield, UK

Abstract

Point-of-interest (POI) type prediction is the task of inferring the type of a place from where a social media post was shared. Inferring a POI's type is useful for studies in computational social science including sociolinguistics, geosemiotics, and cultural geography, and has applications in geosocial networking technologies such as recommendation and visualization systems. Prior efforts in POI type prediction focus solely on text, without taking visual information into account. However in reality, the variety of modalities, as well as their semiotic relationships with one another, shape communication and interactions in social media. This paper presents a study on POI type prediction using multimodal information from text and images available at posting time. For that purpose, we enrich a currently available data set for POI type prediction with the images that accompany the text messages. Our proposed method extracts relevant information from each modality to effectively capture interactions between text and image achieving a macro F1 of 47.21 across eight categories significantly outperforming the state-of-the-art method for POI type prediction based on text-only methods. Finally, we provide a

detailed analysis to shed light on cross-modal interactions and the limitations of our best performing model.[1]

3.1 Introduction

A place is typically described as a physical space infused with human meaning and experiences that facilitate communication (Tuan, 1977). The multimodal content of social media posts (e.g. text, images, emojis) generated by users from specific places such as restaurants, shops, and parks, contribute to shaping a place's identity, by offering information about feelings elicited by participating in an activity or living an experience in that place (Tanasescu et al., 2013).

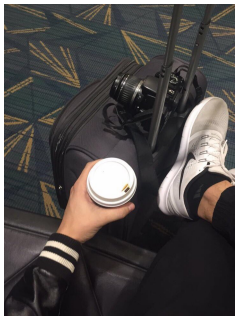
Figure 3.1 shows examples of Twitter posts consisting of image-text pairs, shared from two different places or Point-of-Interests (POIs). Users share content that is relevant to their experience in the location. For example, the text *imagine all the people sharing all the world* which is accompanied by a photograph of the Imagine Mosaic in Central Park; and the text *Next stop: NYC* along with a picture of descriptive items that people carry at an airport such as luggage, a camera and a takeaway coffee cup.

Developing computational methods to infer the type of a POI from social media posts (Liu et al., 2012; Sánchez Villegas et al., 2020) is useful for complementing studies in computational social science including sociolinguistics, geosemiotics, and cultural geography (Kress et al., 1996; Scollon and Scollon, 2003; Al Zydjaly, 2014), and has applications in geosocial networking technologies such as recommendation and visualization systems (Alazzawi et al., 2012; Zhang and Cheng, 2018; van Weerdenburg et al., 2019; Liu et al., 2020b).

Previous work in natural language processing (NLP) has investigated the language that people use in social media from different locations, by inferring the type of a POI of a given social media post using only text and posting time, ignoring the visual context (Sánchez Villegas et al., 2020). However, communication and interactions in social media are naturally shaped by the variety of available modalities and their semiotic relationships (i.e. how meaning is created and communicated) with one another (Georgakopoulou and Spilioti, 2015; Kruk et al., 2019; Vempala and Preoţiuc-Pietro, 2019).

[1]Code and data are available here: https://github.com/danaesavi/poi-type-prediction

imagine all the people
sharing all the world ~



Next stop: NYC ✈️

Figure 3.1: Example of text and image content of sample tweets. Users share content that is relevant to their experiences and feelings in the location.

In this paper, we propose POI type prediction using multimodal content available at posting time by taking into account textual and visual information. Our contributions are as follows:

- We enrich a publicly available data set of social media posts and POI types with images;
- We propose a multimodal model that combines text and images in two levels using: (i) a modality gate to control the amount of information needed from the text and image; (ii) a cross-attention mechanism to learn cross-modal interactions. Our model significantly outperforms the best state-of-the-art method proposed by Sánchez Villegas et al. (2020);
- We provide an in-depth analysis to uncover the limitations of our model and uncover cross-modal characteristics of POI types.

3.2 Related Work

3.2.1 POI Analysis

POIs have been studied to classify functional regions (e.g. residential, business, and transportation areas) and to analyze activity patterns using social media check-in data and geo-referenced images (Zhi et al., 2016; Liu et al., 2020a; Zhou et al., 2020a; Zhang et al.,

2020). Zhou et al. (2020a) present a model for classifying POI function types (e.g. bank, entertainment, culture) using POI names and a list of results produced by searching for the POI name in a web search engine. In contrast, our research is focused on the classification of POI types using textual and visual content extracted from social media posts. Unlike the approach of relying on POI names to fetch information from search engines, our emphasis is on analyzing the inherent information within social media data to categorize points-of-interest. Zhang et al. (2020) makes use of social media check-ins and street-level images to compare the different activity patterns of visitors and locals, and uncover inconspicuous but interesting places for them in a city. A framework for extracting emotions (e.g. joy, happiness) from photos taken at various locations in social media is described in Kang et al. (2019).

3.2.2 POI Type Prediction

POI type prediction is related to geolocation prediction of social media posts that has been widely studied in NLP (Eisenstein et al., 2010; Roller et al., 2012; Dredze et al., 2016). However, while geolocation prediction aims to infer the exact geographical location of a post using language variation and geographical cues, POI type prediction is focused on identifying the characteristics associated with each type of place, regardless of its geographic location.

Previous work on POI type prediction from social media content has used Twitter posts (text and posting time), to identify the POI type from where a post was sent from (Liu et al., 2012; Sánchez Villegas et al., 2020). Liu et al. (2012) incorporate text, temporal features (posting hour) and user history information into probabilistic text classification models. Rather than a user-based study, our research aims to uncover the characteristics associated with various types of POIs. Sánchez Villegas et al. (2020) analyze semantic place information of different types of POIs by using text and temporal information (hour, and day of the week) of a Twitter's post. To the best of our knowledge, this is the first study to combine textual and visual features to classify POI types (e.g. arts & entertainment, nightlife spot) from social media messages, regardless of its geographic location.

3.2.3 Social Media Analysis using Text and Images

The combination of text and images of social media posts has been largely used for different applications such as sentiment analysis, (Nguyen and Shirai, 2015; Chambers et al., 2015),

sarcasm detection (Cai et al., 2019) and text-image relation classification (Vempala and Preoţiuc-Pietro, 2019; Kruk et al., 2019). Moon et al. (2018a) propose a model for recognizing named entities from short social media texts using image and text. Cai et al. (2019) use a hierarchical fusion model to integrate image and text context with an attention-based fusion. Chinnappa et al. (2019) examine the possession relationships from text-image pairs in social media posts. Wang et al. (2020) use texts and images for predicting the keyphrases (i.e. representative terms) for a post by aligning and capturing the cross-modal interactions via cross-attention. Previous text-image classification in social media requires that the data is fully paired, i.e. every post contains an image and a text. This becomes limiting when faced with missing data, as not all posts contain both modalities[2]. To address this limitation, in this work we propose a model that can handle both scenarios:(1) when all modalities (text-image pairs) are present and (2) when only text is available. This dual-capability enables the model to perform effectively in situations where information in one modality is absent.

Social media analysis research has also looked at the semiotic properties of text-image pairs in posts (Alikhani et al., 2019; Vempala and Preoţiuc-Pietro, 2019; Kruk et al., 2019). Vempala and Preoţiuc-Pietro (2019) investigate the relationship between text and image content by identifying overlapping meaning in both modalities, those where one modality contributes with additional details, and cases where each modality contributes with different information. Kruk et al. (2019) analyze the relationship between the text-image pairs and find that when the image and caption diverge semiotically, the benefit from multimodal modeling is greater.

3.3 Task & Data

Sánchez Villegas et al. (2020) define POI type prediction as a multi-class classification task where given the text content of a post, the goal is to classify it in one of the M POI categories. In this work, we extend this task definition to include images in order to capture the semiotic relationships between the two modalities. For that purpose, we consider a social media post P (e.g. tweet) to comprise of a text and image pair (x^t, x^v), where $x^t \in \mathbb{R}^{d_t}$ and $x^v \in \mathbb{R}^{d_v}$ are the textual and visual vector representations respectively.

[2]See: https://buffer.com/resources/twitter-data-1-million-tweets/

Category	Train		Dev		Test		
	# Tweets	# Images	# Tweets	# Images	# Tweets	# Images	Tokens
Arts & Entertainment	40,417	20,711	4,755	2,527	5,284	2,740	14.41
College & University	21,275	9,112	2,418	1,057	2,884	1,252	15.52
Food	6,676	2,969	869	351	724	280	14.34
Great Outdoors	27,763	13,422	4,173	2,102	3,653	1,948	13.49
Nightlife Spot	5,545	2,532	876	385	656	353	15.46
Professional & Other Places	30,640	13,888	3,381	1,499	3,762	1,712	16.46
Shop & Service	8,285	3,455	886	266	812	353	15.31
Travel & Transport	16,428	6,681	2,201	829	1,872	789	14.88
All	157,029	72,679 (46.28%)	19,559	9,006 (46.05%)	19,647	9,410 (47.90%)	14.92

Table 3.1: POI categories and data set statistics showing the number of tweets for each category, and number (%) of tweets having an accompanying image

3.3.1 POI Data

We use the data set introduced by Sánchez Villegas et al. (2020) which contains 196,235 tweets written in English, labeled with one out of the eight POI broad type categories shown in Table 3.1, which correspond to the 8 primary top-level POI categories in 'Places by Foursquare', a database of over 105 million POIs worldwide managed by Foursquare. To generalize to locations not present in the training set, we use the same location-level data splits (train, dev, test) as in Sánchez Villegas et al. (2020), where each split contains tweets from different locations.

3.3.2 Image Collection

We use the Twitter API to collect the images that accompany each textual post in the data set. For the tweets that have more than one image, we select the first available only. This results in 91,224 tweets with at least one image. During the image processing (see Section 3.5.3) we removed 129 images because we found they were either damaged, absent[3], or no objects were detected, resulting in 91,095 text-image pairs (see Table 3.1 for data statistics). In order to deal with the rest of the tweets with no associated image, we pair them with a single 'average' image computed over all images in the train set: $x^v = avg(x^v_{tr})$. The intuition behind this approach is to generate a 'noisy' image that is not related and does not add to

[3]Removed by Twitter due to violations to the Twitter Rules and Terms of Service.

the meaning (Vempala and Preoţiuc-Pietro, 2019).[4]

3.3.3 Exploratory Analysis of Image Data

To shed light on the characteristics of the collected images, we apply object detection on the images collected using Faster-RCNN (Ren et al., 2016) pretrained on Visual Genome (Krishna et al., 2017; Anderson et al., 2018). Table 3.2 shows the most common objects for each specific category. We observe that most objects are related to items one would find in each place category (e.g. 'spoon', 'meat', 'knife' in *Food*). Clothing items are common across category types (e.g. 'shirt', 'jacket', 'pants') suggesting the presence of people in the images. A common object tag of the *Shop & Service* category is 'letters', which concerns images that contain embedded text. Finally, the category *Great Outdoors* includes object tags such as 'cloud', 'hill', and 'grass', words that describe the landscape of this type of place.

Building upon the insights detailed in Sánchez Villegas and Aletras (2021) (see Section 2.3.1), a prior study which provides a linguistic analysis on the same data set, we observe a convergence in the exploratory analyses of both image and text data. While the text analysis shows prevalent words and expressions linked to both the place categories and the activities users engage in at these locations, the object detection applied to images reveals common objects specific to each place category, providing visual context that aligns with the types of activities and environments associated with those locations. Thus, we suspect that integrating visual information to the models will enhance their predicting capability by including the actual objects and scenes present in these locations, and thus providing a more comprehensive understanding of the diverse aspects related to different types of POIs.

3.4 Multimodal POI Type Prediction

3.4.1 Text and Image Representation

Given a text-image post $P = (x^t, x^v)$, $x^t \in \mathbb{R}^{d_t}$, $x^v \in \mathbb{R}^{d_v}$, we first compute text and image representations.

[4]Early experimentation with associating tweets with the image of the most similar tweet that contains a real image from the training data yielded similar performance.

Category	Common Objects in Images
Arts & Entertainment	light, pants, shirt, arm, picture, hair, glasses, line, girl, jacket
College & University	pants, shirt, line, hair, arm, picture, light, glasses, girl, trees
Food	cup, picture, spoon, meat, knife, arm, glasses, shirt, pants, handle
Great Outdoors	trees, arm, pants, cloud, hill, line, shirt, grass, picture, glasses
Nightlife Spot	arm, picture, shirt, light, hair, pants, glasses, mouth, girl, cup
Professional & Other Places	pants, shirt, picture, light, hair, screen, line, arm, glasses, girl
Shop & Service	picture, pants, arm, shirt, glasses, light, hair, line, girl, letters
Travel & Transport	pants, shirt, light, screen, arm, hair, glasses, picture, chair, line

Table 3.2: Most common objects for each POI category.

Text We use Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019) to obtain the text representations f^t.

Image For encoding the images, we use Xception (Chollet, 2017) pre-trained on ImageNet (Deng et al., 2009).[5] We extract convolutional feature maps for each image and we apply average pooling to obtain the image representation f^v.

3.4.2 MM-Gate

Given the complex semiotic relationship between text and image, we need a weighting strategy that assigns more importance to the most relevant modality while suppressing irrelevant information. Thus, a first approach is to use gated multimodal fusion (MM-Gate), similar to the approach proposed by Arevalo et al. (2020) to control the contribution of text and image to the POI type prediction. Given f^t, f^v the text and visual representations, we obtain the

[5]Early experimentation with ResNet101 (He et al., 2016) and EfficientNet (Tan and Le, 2019) yielded similar results.

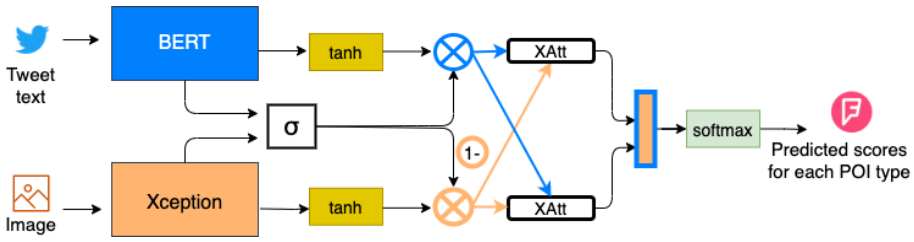


Figure 3.2: Overview of our MM-Gated-XAtt model which combines features from text and image modalities for POI type prediction.

multimodal representation h of a post P as follows:

$$h^t = tanh(W^t f^t + b^t) \tag{3.1}$$

$$h^v = tanh(W^v f^v + b^v) \tag{3.2}$$

$$z = \sigma(W^z [f^t; f^v] + b^z) \tag{3.3}$$

$$h = z * h^t + (1 - z) * h^v \tag{3.4}$$

where $W^t \in \mathbb{R}^{d_t}$, $W^v \in \mathbb{R}^{d_v}$ and $W^z \in R^{d_t+d_v}$ are learnable parameters, $tanh$ is the activation function and $h^t, h^v \in \mathbb{R}$ are projections of f^t and f^v. $[;]$ denotes concatenation and σ is the sigmoid activation function. h is a weighted combination of the textual and visual information h^t and h^v respectively. We fine-tune the entire model by adding a classification layer with a softmax activation function for POI type prediction

3.4.3 MM-XAtt

The MM-Gate model does not capture interactions between text and image that might be beneficial for learning semiotic relationships. To model cross-modal interactions, we adapt the cross-attention mechanism (Tsai et al., 2019; Tan and Bansal, 2019) to combine text and image information for multimodal POI type prediction (MM-XAtt). Cross-attention consists of two attention layers, one from textual to visual features, and one from visual to textual features. We first linearly project the text and visual representations to obtain the same dimensionality (d_{proj}). Then, we compute the scaled dot attention ($a = softmax\frac{(Q(K)^T)}{\sqrt{d_{proj}}}V$) with the projected text representation as query (Q), and the projected image representation as the key (K) and values (V), and vice versa. The multimodal representation h is the sum of the resulting attention layers. The entire model is fine-tuned by adding a classification layer with a softmax activation function.

3.4.4 MM-Gated-XAtt

Vempala and Preoţiuc-Pietro (2019) have demonstrated that the relationship between the text and image in a social media post is complex. Images may or may not add meaning to the post and the text content (or meaning) may or may not correspond to the image. We hypothesize that this might actually happen in posts made from particular locations, i.e. language and visual information may or may not be related. To address this, we propose (1) using gated multimodal fusion to manage the flow of information from each modality, and (2) also learn cross-modal interactions by using cross-attention on top of the gated multimodal mechanism. Figure 3.2 shows an overview of our model architecture (MM-Gated-XAtt). Given the text and image representations f^t, f^v respectively, we compute h^t, h^v, and z as in Equation 3.1, 3.2 and 3.3. Next, we apply cross-attention using two attention layers where the query and context vectors are the weighted representations of the text and visual modalities, $z * h^t$ and $(1 - z) * h^v$, and vice versa. The multimodal context vector h is the sum of the resulting attention layers. Finally, we fine-tune the model by passing h through a classification layer for POI type prediction with a softmax activation function.

3.5 Experimental Setup

3.5.1 Baselines

We compare our models against (1) text-only; (2) image-only; and (3) other state-of-the-art multimodal approaches.[6]

Text-only We fine-tune BERT for POI type classification by adding a classification layer with softmax activation function on top of the [CLS] token which is the best performing model in Sánchez Villegas et al. (2020).

Image-only We fine-tune three pre-trained models that are popular in various computer vision classification tasks: (1) ResNet101 (He et al., 2016); (2) EfficientNet (Tan and Le,

[6]We include a majority class baseline (i.e. assigning all instances in the test set the most frequent label in the train set).

2019); and (3) Xception (Chollet, 2017). Each model is fine-tuned on POI type classification by adding an output softmax layer.

Text and Image For combining text and image information, we experiment with different standard fusion strategies: (1) we project the image representation f^v, to the same dimensionality as $f^t \in \mathbb{R}^{d_t}$ using a linear layer and then we concatenate the vectors (**Concat**); (2) we project the textual and visual features to the same space and then we apply self-attention to learn weights for each modality (**Attention**); (3) we also adapt the guided attention introduced by Anderson et al. (2018) for learning attention weights at the object-level (and other salient regions) rather than equally sized grid-regions (**Guided Attention**); (4) we compare against **LXMERT**, a transformer-based model that has been pre-trained on text and image pairs for learning cross-modality interactions (Tan and Bansal, 2019). All models are fine-tuned by adding a classification layer with a softmax activation function for POI type prediction. Finally, we evaluate a simple ensemble strategy by using LXMERT for classifying tweets that are originally accompanied by an image and BERT for classifying text-only tweets (**Ensemble**).

3.5.2 Text Processing

We use the same tokenization settings as in Sánchez Villegas et al. (2020). For each tweet, we lowercase text and replace URLs and @-mentions of users with placeholder tokens.

3.5.3 Image Processing

Each image is resized to (224×224) pixels representing a value for the red, green and blue color in the range of $[0, 255]$. The pixel values of all images are normalized. For LXMERT and Guided Attention fusion, we extract *object-level* features using Faster-RCNN (Ren et al., 2016) pretrained on Visual Genome (Krishna et al., 2017) following Anderson et al. (2018). We keep 36 objects for each image as in Tan and Bansal (2019).

3.5.4 Implementation Details

We select the hyperparameters for all models using early stopping by monitoring the validation loss using the Adam optimizer (Kingma and Ba, 2014). Because the data is imbalanced, we estimate the class weights using the 'balanced' heuristic (King and Zeng, 2001). All experiments are performed using a Nvidia V100 GPU.

Text-only We fine-tune BERT 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) with a maximal sequence length of 50 tokens. We fine-tune BERT for 2 epochs and learning rate $\eta = 2e^{-5}$ with $\eta \in \{2e^{-5}, 3e^{-5}, 5e^{-5}\}$.

Image-only For ResNet101, we fine-tune for 5 epochs with learning rate $\eta = 1e^{-4}$ and dropout $\delta = 0.2$ (δ in $[0, 0.5]$ using random search) before passing the image representation through the classification layer. EfficientNet is fine-tuned for 7 epochs with $\eta = 1e^{-5}$ and $\delta = 0.5$. Xception is fine-tuned for 6 epochs with $\eta = 1e^{-5}$ and $\delta = 0.5$.

Text and Image Concat-BERT+Xception, Concat-BERT+ResNet and Guided Attention-BERT+Xception are fine-tuned for 2 epochs with $\eta = 1e^{-5}$ and $\delta = 0.25$; Concat-BERT + EfficientNet for 4 epochs with $\eta = 1e^{-5}$ and $\delta = 0.25$; Attention-BERT+Xception for 3 epochs with $\eta = 1e^{-5}$ and $\delta = 0.25$; MM-XAtt for 3 epochs with $\eta = 1e^{-5}$ and $\delta = 0.15$; MM-Gate and MM-Gated-XAtt for 2 epochs with $\eta = 1e^{-5}$ and $\delta = 0.05$; $\eta \in \{2e^{-5}, 3e^{-5}, 5e^{-5}\}$, δ from $[0, 0.5]$ (random search) before passing through the classification layer. We fine-tune LXMERT for 4 epochs with $\eta = 1e^{-5}$ where $\eta \in \{1e^{-3}, 1e^{-4}, 1e^{-5}\}$ and dropout $\delta = 0.25$ (δ in $[0, 0.5]$, random search) before passing through the classification layer.

3.5.5 Evaluation

We evaluate the performance of all models using macro F1, precision, and recall. Results are obtained over three runs using different random seeds reporting the average and the standard deviation.

Model	F1	P	R
Majority	5.30	3.36	12.50
BERT (Sánchez Villegas et al., 2020)	43.67 (0.01)	**48.44** (0.02)	41.33 (0.01)
ResNet	21.11 (1.81)	23.23 (2.09)	29.90 (3.31)
EfficientNet	24.72 (0.76)	28.05 (0.28)	35.48 (0.23)
Xception	23.64 (0.44)	25.62 (0.50)	34.12 (0.49)
Concat-BERT+ResNet	43.28 (0.37)	42.72 (0.51)	47.59 (0.45)
Concat-BERT+EfficientNet	41.56 (0.71)	41.54 (0.88)	43.97 (0.79)
Concat-BERT+Xception	44.00 (0.52)	43.34 (0.70)	48.35 (0.75)
Attention-BERT+Xception	42.89 (0.44)	42.74 (0.19)	46.78 (1.28)
Guided Attention-BERT+Xception	41.53 (0.57)	41.10 (0.55)	45.36 (0.48)
LXMERT	40.17 (0.62)	40.26 (0.24)	42.25 (2.38)
Ensemble-BERT+LXMERT	43.82 (0.47)	43.50 (0.20)	44.67 (0.66)
MM-Gate	44.64 (0.65)	43.67 (0.49)	48.50 (0.18)
MM-XAtt	27.31 (1.58)	37.06 (2.66)	29.71 (0.60)
MM-Gated-XAtt (Ours)	**47.21**† (1.70)	46.83 (1.45)	**50.69** (2.21)

Table 3.3: Macro F1-Score, precision (P) and recall (R) for POI type prediction (± std. dev.) Best results are in bold. † indicates statistically significant improvement (t-test, $p < 0.05$) over BERT (Sánchez Villegas et al., 2020).

3.6 Results

The results of POI type prediction are presented in Table 3.3. We first examine the impact of each modality by analyzing the performance of the unimodal models, then we investigate the effect of multimodal methods for POI type prediction, and finally we examine the performance of our proposed model MM-Gated-XAtt by analyzing each component independently.

We observe that the text-only model (BERT) achieves 43.67 F1 which is substantially higher than the performance of image-only models (e.g. the best performing EfficientNet model obtains 24.72 F1). This suggests that text encapsulates more relevant information for this task than images on their own, similar to other studies in multimodal computational social science (Wang et al., 2020; Ma et al., 2021).

Models that simply concatenate text and image vectors have close performance to BERT (44.0 for Concat-BERT+Xception) or lower (41.56 for Concat-BERT+EfficientNet). This suggests that assigning equal importance to text and image information can deteriorate performance. It also shows that modeling cross-modal interactions is necessary to boost performance of POI type classification models.

Text-Image Only	
Model	**F1**
LXMERT	47.72 (0.98)
MM-Gate	45.87 (1.48)
MM-XAtt	48.93 (2.08)
MM-Gated-XAtt (Ours)	**57.64** (3.64)

Table 3.4: Macro F1-Score for POI type prediction on tweets that are originally accompanied by an image. Best results are in bold.

Surprisingly, we observe that the pre-trained multimodal LXMERT fails to improve over BERT (40.17 F1) while its performance is lower than simpler concatenative fusion models. We speculate that this is because LXMERT is pretrained on data where both, text and image modalities share common semantic relationships which is the case in standard vision-language tasks including image captioning and visual question answering (Zhou et al., 2020b; Lu et al., 2019a). On the other hand, text-image relationships in social media data for inferring the type of location from which a message was sent are more diverse, highlighting the particular challenges for modeling text and images together (Hessel and Lee, 2020).

Our proposed MM-Gated-XAtt model achieves 47.21 F1 which significantly (t-test, $p < 0.05$) improves over BERT, the best performing model in Sánchez Villegas et al. (2020) and consistently outperforms all other image-only and multimodal approaches. This confirms our main hypothesis that modeling text with image jointly to learn the interactions between modalities benefit performance in POI type prediction. We also observe that using only the gating mechanism (MM-Gate) outperforms (44.64 F1) all other models except for MM-Gated-XAtt. This highlights the importance of controlling the information flow for the two modalities. Using cross-attention on its own (MM-XAtt), on the other hand, fails to improve over other multimodal approaches, implying that learning cross-modal interactions is not sufficient on its own. This supports our hypothesis that language and visual information in posts sent from specific locations may be or may not be related, and that managing the flow of information from each modality improves the classifier's performance.

Finally, we investigate using less noisy text-image pairs in alignment with related computational social science studies involving text and images (Moon et al., 2018a; Cai et al., 2019; Chinnappa et al., 2019). We train and test LXMERT, MM-Gate, MM-XAtt, and MM-Gated-XAtt on tweets that are originally accompanied by an image (see Section 3.3), excluding all text-only tweets. The results are shown in Table 3.4. In general, performance is higher for all models using less noisy data. Our proposed model MM-Gated-XAtt consistently achieves

Text-Image Only –> All	
Model	**F1**
MM-Gate	40.67 (0.45)
MM-XAtt	31.00 (0.89)
MM-Gated-XAtt (Ours)	**42.45** (2.94)

Table 3.5: Macro F1-Score for POI type prediction. Models are trained on tweets that are originally accompanied by an image. Results are on all tweets. Best results are in bold.

the best performance (57.64 F1). In addition, we observe that LXMERT and MM-XAtt produce similar results (47.72 and 48.93 F1 respectively) suggesting that cross-attention can be applied directly to text-image pairs in low-noise settings without hurting the model performance. The benefit of controlling the flow of information through a gating mechanism, on the other hand, strongly improves model robustness.

3.6.1 Training on Text-Image Pairs Only

To compare the effect of the 'average' image (see Section 3.3) on the performance of the models, we train MM-Gate, MM-XAtt, and MM-Gated-XAtt on tweets that are originally accompanied by an image excluding all text-only tweets; and we test on all tweets as in our original setting (text-only tweets are paired with the 'average' image). The results are shown in Table 3.5. MM-Gated-XAtt is consistently the best performing model, followed by MM-Gate. However, their performance is inferior than when models are trained on all tweets using the 'average' image as in the original setting. This suggests that the gate operation not only regulates the flow of information for each modality but also learns how to use the noisy modality to improve classification prediction. This result is similar to findings by (Arevalo et al., 2020).

3.7 Analysis

3.7.1 Modality Contribution

To determine the influence of each modality in MM-Gated-XAtt when assigning a particular label to a tweet, we compute the average percentage of activations for the textual and visual

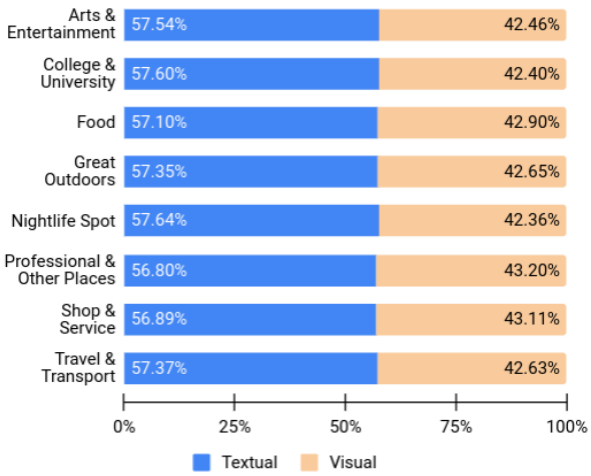


Figure 3.3: Average percentage of MM-Gated-XAtt activations for the textual and visual modalities for each POI category on the test set.

modalities for each POI category on the test set. The outcome of this analysis is depicted in Figure 3.3. As anticipated, the textual modality has a greater influence on the model prediction, which is consistent with our findings in Section 3.6. The category where the visual modality has greater impact on the predicted label is *Professional & Other Places* (43.20%) followed by *Shop & Service* (43.11%).

To examine how the visual information impacts the POI type prediction task, Figure 3.4 shows examples of posts where the contribution of the image is large while the text-only model (BERT) misclassified the POI category. We observe that the text content of Post (a) misled BERT towards *Food*, probably due to the term 'powder'. On the other hand, MM-Gated-XAtt can filter irrelevant information from the text, and prioritize relevant content from the image in order to assign the correct POI category for Post (a) (*Great Outdoors*). Likewise, Post (b) was correctly classified by MM-Gated-XAtt as *Shop & Service* and misclassified by BERT as *Arts & Entertainment*. For this post 40% of the contribution corresponds to the image and 60% to text. This shows how image information can help to address the ambiguity in short texts (Moon et al., 2018b), improving POI type prediction.

3.7.2 Cross-attention (XAtt)

Figure 3.4 shows examples of the XAtt visualization. We note that the model focuses on relevant nouns and pronouns (e.g. 'track', 'it'), which are common informative words in vision-and-language tasks (Tan et al., 2019). Moreover, our model focuses on relevant words such

Figure 3.4: POI type predictions of MM-Gated-XAtt (Ours) and BERT (Sánchez Villegas et al., 2020) showing the contribution of each modality (%) and the XAtt visualization. Correct predictions are in bold.

as 'track' for classifying Post (a) as *Great Outdoors*. Lastly, we observe that the XAtt often captures a general image information, with emphasis on specific sections for the predicted POI category such as the pine trees for *Great Outdoors* and the display racks for *Shop & Service*.

3.7.3 Error Analysis

To shed light on the limitations of our multimodal MM-Gated-XAtt model for predicting POI types, we performed an analysis of misclassifications. In general, we observe that the model struggles with identifying POI categories where people might perform similar activities in each of them such as *Food*, *Nightlife Spot*, and *Shop & Service* similar to findings by Ye et al. (2011).

Figure 3.5 (a) and (b) show examples of tweets misclassified as *Food* by the MM-Gated-XAtt model. Post (a) belongs to the category *Nightlife Spot* and Post (b) belongs to the *Shop & Service* category. In both cases, the text and image content is related to the *Food* category,

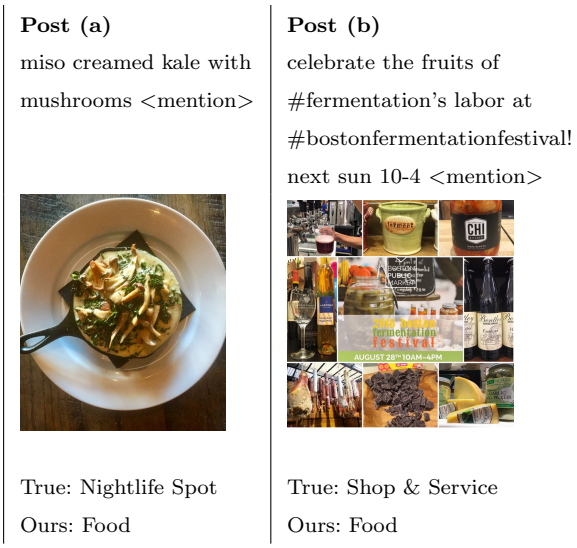


Figure 3.5: Example of misclassifications made by our MM-Gated-XAtt model.

misleading the classifier towards this POI type. Posting about food is a common practice in hospitality establishments such as restaurants and bars (Zhu et al., 2019), where customers are more likely to share content such as photos of dishes and beverages, intentionally designed to show that are associated with the particular context and lifestyle that a specific place represents (Homburg et al., 2015; Brunner et al., 2016; Apaolaza et al., 2021). Similarly, Post (b) shows an example of a tweet that promotes a POI by communicating specific characteristics of the place (Kruk et al., 2019; Aydin, 2020). To correctly classify the category of POIs, the model might need access to deeper contextual information about the locations (e.g. finer subcategories of a type of place and how POI types are related to one another).

3.8 Conclusion and Future Work

This paper presents the first study on multimodal POI type classification using text and images from social media posts motivated by studies in geosemiotics, visual semiotics and cultural geography. We enrich a publicly available data set with images and we propose a multimodal model that uses: (1) a gate mechanism to control the information flow from each modality; (2) a cross-attention mechanism to align and capture the interactions between modalities. Our model achieves state-of-the-art performance for POI type prediction significantly outperforming the previous text-only model and competitive pretrained multimodal models.

In future work, we plan to perform more granular prediction of POI types and user

information to provide additional context to the models. Our models could also be used for modeling other tasks where text and images naturally occur in social media such as analyzing political ads (Sánchez Villegas et al., 2021), parody (Maronikolakis et al., 2020b) and complaints (Preoţiuc-Pietro et al., 2019b; Jin and Aletras, 2020, 2021).

Acknowledgments

We would like to thank Mali Jin, Panayiotis Karachristou and all reviewers for their valuable feedback. DSV is 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. NA is supported by a Leverhulme Trust Research Project Grant.

Ethical Statement

Our work complies with Twitter data policy for research,[7] and has received approval from the Ethics Committee of our institution (Ref. No 039665).

In this work, the analyses are focused on extracting general trends for different types of POIs rather than individual user profiles. However, it is important to acknowledge the potential misuse of inferred location information, particularly by entities such as health insurance providers or employers. Secondly, in this work, we collected publicly available information only, from users that have their GPS activated. However, an inherent risk in models predicting POI types from social media posts lies in their potential application to users without activated GPS who may not wish for their location or the type of place they are posting from to be inferred. Introducing an opt-in feature within social media platforms would empower users to willingly engage in research or commercial experiments, granting them control over the disclosure of their location-related and other user information. Finally, educating users about the potential implications of their social media content, including the inference of location-related information, is important for informed and ethical participation.

[7]See: https://developer.twitter.com/en/developer-terms/agreement-and-policy

Chapter 4

Publication III: Analyzing Online Political Advertisements

Analyzing Online Political Advertisements

Danae Sánchez Villegas[α], *Saeid Mokaram*[β], *Nikolaos Aletras*[α]

[α] Computer Science Department, University of Sheffield, UK. [β] Emotech

Abstract

Online political advertising is a central aspect of modern election campaigning for influencing public opinion. Computational analysis of political ads is of utmost importance in political science to understand the characteristics of digital campaigning. It is also important in computational linguistics to study features of political discourse and communication on a large scale. In this work, we present the first computational study on online political ads with the aim to (1) infer the political ideology of an ad sponsor; and (2) identify whether the sponsor is an official political party or a third-party organization. We develop two new large datasets for the two tasks consisting of ads from the U.S.. Evaluation results show that our approach that combines textual and visual information from pre-trained neural models outperforms a state-of-the-art method for generic commercial ad classification. Finally, we provide an in-depth analysis of the limitations of our best-performing models and linguistic analysis to study the characteristics of political ads discourse.[1]

[1]Data is available here: https://archive.org/details/pol_ads

4.1 Introduction

Online advertising is an integral part of modern digital election campaigning (Fulgoni et al., 2016; Fowler et al., 2020a). The increased spending on online political ads (e.g. the 2020 U.S. election campaign spending hit an all-time record[2]) poses a significant challenge to the democratic oversight of digital campaigning,[3] with serious implications about transparency and accountability, for example how voters are targeted and by whom (Kriess and Barrett, 2020).

Political advertising is defined as *'any controlled message communicated through any channel designed to promote the political interests of individuals, parties, groups, government, or other organizations'* (Kaid and Holtz-Bacha, 2006). It is guided by ideology and morals (Scammell and Langer, 2006; Kumar and Pathak, 2012), and often expresses more negativity (Haselmayer, 2019; Iyengar and Prior, 1999; Lau et al., 1999) compared to the aesthetic nature of commercial advertising. Table 4.1 shows examples of online political ads across different political parties and sponsor types.

While the closely related online *commercial* advertising domain has recently been explored in natural language processing (NLP) for predicting the category (e.g. politics, cars, electronics) and sentiment of an ad (Hussain et al., 2017; Kalra et al., 2020), online *political* advertising has yet to be explored. Large-scale studies of online political advertising have so far focused on understanding targeting strategies rather than developing predictive models for analyzing its content (Edelson et al., 2019; Medina Serrano et al., 2020).

Automatically analyzing political ads is important in political science for researching the characteristics of online campaigns (e.g. voter targeting, sponsors, non-party campaigns, privacy, and misinformation) on a large scale (Scammell and Langer, 2006; Johansson and Holtz-Bacha, 2019). Moreover, identifying ads sponsored by third-party organizations is critical to ensuring transparency and accountability in elections (Liu et al., 2013; Speicher et al., 2018; Fowler et al., 2020b; Edelson et al., 2019). For example, third-party advertising had an increased presence in the U.S. House and Senate races in 2018 considerably more than in 2012 where almost half of the third-party sponsored ads were funded by dark-money sources (Fowler et al., 2020b). Moreover, ad publishers and social media users are not

[2] https://www.cnbc.com/2020/10/01/election-2020-campaign-spending-set-to-hit-record-11-billion.html

[3] https://www.electoral-reform.org.uk/latest-news-and-research/publications/democracy-in-the-dark-digital-campaigning-in-the-2019-general-election-and-beyond/

Political Ideology	Ad Sponsor Type	Sample Ad
Liberal	Political Party	
Conservative	Political Party	
N/A	Third-Party	

Table 4.1: Examples of online political ads by sponsor political ideology and type.

obliged to provide any disclaimer about the sponsor of the ad in several countries. Thus, computational methods, including machine learning models, can be applied for large-scale analysis of online political ads especially for ads where sponsor information is not available. Finally, political ads analysis can help linguists to study features of political discourse and communication (Kenzhekanova, 2015; Skorupa and Dubovičienė, 2015).

In this paper, we present a systematic study of online political ads (consisting of text and images) in the U.S. to uncover linguistic and visual cues across political ideologies and sponsor types using computational methods for the first time. Our contributions are as follows:

1. A new classification task for predicting the political ideology (conservative or liberal) of an ad (Section 4.3). We collect 5,548 distinct political ads in English from 242 different advertisers in the U.S., and label them according to the dominant political ideology of the respective sponsor's party affiliation (*Liberal* or *Conservative*);

2. A new classification task to automatically classify ads that were sponsored by official political parties and third-party organizations, such as businesses and non-profit organizations (Section 4.3). For this task, we extract 15,116 advertisements in English from 665 distinct advertisers in the U.S., and label them as *Political Party* (i.e. officially registered) and *Third-Party* (i.e. other organizations) following Fowler et al. (2020b);

3. Experiments with text-based and multimodal (text and images) models (Section 4.4) for political ideology prediction and sponsor type classification reaching up to 75.76 and 87.36 macro F1 in each task respectively (Section 4.6);

4. Analysis of textual and visual features of online political ads (Section 4.7) and error analysis to understand model limitations.

4.2 Related Work

4.2.1 Political Communication and Advertising

Previous work on analyzing political advertising has covered television and online ads (Kaid and Postelnicu, 2005; Reschke and Anand, 2012; West, 2017; Fowler et al., 2020b). Ridout et al. (2010) analyze a series of YouTube videos posted during the 2008 presidential campaign to understand its influence on election results as well as the actors and formats compared to traditional television ads. Anstead et al. (2018) study how online platforms such as Facebook are being used for political communication and identify challenges for understanding the role of these platforms in political elections, highlighting the lack of transparency (Caplan and Boyd, 2016). Fowler et al. (2020b) explore differences between television and online ads, and demonstrate that there is a greater number of candidates advertising online than on television.

4.2.2 Political Ideology Prediction

Inferring the political ideology of various types of text including news articles, political speeches and social media has been vastly studied in NLP (Lin et al., 2008; Gerrish and Blei, 2011; Sim et al., 2013; Iyyer et al., 2014; Preoţiuc-Pietro et al., 2017; Kulkarni et al., 2018; Stefanov et al., 2020). Bhatia and P (2018) exploit topic-specific sentiment analysis for ideology detection (i.e. conservative, liberal) in speeches from the U.S. Congress. Kulkarni et al. (2018) propose a multi-view model that incorporates textual and network information to predict the ideology of news articles. Johnson and Goldwasser (2018) investigate the relationship between political ideology and language to represent morality by analyzing political slogans in tweets posted by politicians. Maronikolakis et al. (2020b) present a study of political parody on Twitter focusing on the linguistic differences between tweets shared by real and parody accounts. Baly et al. (2019) estimate the trustworthiness and political ideology (left/right bias) of news sources as a multi-task problem. Stefanov et al. (2020) develop methods to predict the overall political leaning (left, center or right) of online media and popular Twitter users.

Political ideology and communicative intents have also been studied in computer vision. Political images have been analyzed to infer the persuasive intents using various features

such as facial display types, body poses, and scene context (Joo et al., 2014; Huang and Kovashka, 2016; Joo and Steinert-Threlkeld, 2018; Bai et al., 2021; Chen et al., 2020). Joo et al. (2015) introduce a method that infers the perceived characteristics of politicians using face images and show that those characteristics can be used in elections forecasting. Xi et al. (2020) analyze the political ideology of Facebook photographs shared by members of the U.S. Congress. Chen et al. (2020) examine the role of gender stereotypical cues from photographs posted in social media by political candidates and their relationship to voter support.

4.2.3 Computational Analysis of Online Ads

Hussain et al. (2017) propose the task of ad understanding using vision and language. The aim is to predict the topical category, sentiment and rhetoric of an ad (i.e. what the message is about). The latter task has been approached as a visual question-answering task by ranking human generated statements that explain the intent of the ad in computer vision (Ye and Kovashka, 2018; Ahuja et al., 2018). More recently in NLP, Kalra et al. (2020) propose a BERT-based (Devlin et al., 2019) model for this task using the text and visual descriptions of the ad (Johnson et al., 2016). Thomas and Kovashka (2018) study the persuasive cues of faces across ad categories (e.g. beauty, clothing). Zhang et al. (2018) explore the relationship between the text of an ad and the visual content to analyze the semantics across modalities. Ye et al. (2018) integrates audio and visual modalities to predict the climax of an advertisement (i.e. stress levels) using sentiment annotations.

4.3 Tasks & Data

We aim to analyze the political ideology of ads consisting of image and text, and the type of the ad sponsor for the first time. To this end, we present two new binary classification tasks motivated by related studies in political communication (Grigsby, 2008; Fowler et al., 2020b):

- **Task 1: *Conservative/Liberal*** The aim is to label an ad according to the political party that sponsored the ad either as *Conservative* (i.e. assuming that the dominant ideology of the Republican Party is conservatism), or *Liberal* (i.e. assuming that the dominant ideology of the Democratic Party is social liberalism) (Grigsby, 2008);
- **Task 2: *Political Party/Third-Party*** The goal is to classify an ad according

to the type of the organization that sponsored the ad. We distinguish between ads sponsored by official political parties and non-political organizations, such as businesses and non-profit groups, following Fowler et al. (2020b).

To the best of our knowledge, no datasets are available for modeling these two tasks. Therefore, we develop two new publicly available datasets consisting of political ads and ideology/sponsor type labels from the U.S.. We opted to use data only from the U.S. because its Federal Election Commission[4] (FEC) provides publicly available information of political ads sponsors such as official political parties (e.g. Democratic, Republican) via their FEC ID; and third-party organizations can be identified via their Employer Identification Number[5] (EIN) suitable for our study.

4.3.1 Collecting Online Political Ads

We use the public Google transparency report platform[6] to collect political ads. This platform contains information on verified political advertisers (i.e. sponsors) and provides links to actual political ads from Google Ad Services.

We collect all U.S. available data from the Google platform consisting of ads published from May 31, 2018 up to October 11, 2020 (note that there is no data prior to 2018). This corresponds to a total of 168,146 image ads. Each ad is associated with a URL that links to its summary metadata consisting of a URL to the original image file and sponsor information, i.e. name and FEC ID, state elections registration or EIN ID.[7]

We scrape all available image files resulting into a total of 158,599 ads which corresponds to 94.32% of all ads in the Google database. The rest of the ads were either not available due to violations to Google's Advertising Policy, the summary metadata was missing, or the file URL was not included in the metadata.

[4] https://www.fec.gov/

[5] https://www.irs.gov/businesses/small-businesses-self-employed/do-you-need-an-ein

[6] https://transparencyreport.google.com/political-ads/region/US

[7] All ad sponsors must apply for eligibility verification in order to publish political ads on Google platforms - https://support.google.com/displayvideo/answer/9014141

Sample Ad	
Image Text	FIGHTING FOR WORKING FAMILIES, FOR GOOD JOBS, AND FAIR PAY. PAID FOR BY DEFAZIO FOR CONGRESS
Densecap	the man is wearing glasses, a man holding a red tie, the background is blue

Table 4.2: Example of text, and visual information extracted from a sample Ad.

4.3.2 Extracting Text and Visual Information

Before, we label the ads with ideology and sponsor type, we extract two types of information from the images: (1) the text contained in each ad (Image Text; IT) using the Google Vision API;[8] and (2) the descriptive caption or denscap (D) of the image using the DenseCap API,[9] following the method proposed by Kalra et al. (2020) for commercial ad classification. This way, we obtain both the actual text appearing on the ad and the textual descriptions of the ad such as entities in the images, their characteristics and relationships. Table 4.2 shows an example of an ad consisting of an image, text information and the densecap.

We use the textual and visual information to eliminate all duplicate images by comparing the URL of the image, its text and densecap. Finally, we filter out all ads that contain non-English text (i.e. IT).[10] This results in 15,116 unique ads from 665 unique ad sponsors.

4.3.3 Labeling Ads with Political Ideology

Our aim is to label political ads as *Conservative* or *Liberal* (see Task 1 description). First, we retrieve all the ad sponsors and their corresponding ads that are available in the Google Ads database. Official political committees associated with the Democratic or Republican parties are identified by their FEC ID (included in the sponsor's information in the Google database). However, the name of the political party associated with a sponsor is not available in the Google database. Thus, we query the FEC database to obtain the affiliation for all committees of the Democratic and Republican parties (e.g. Donald J. Trump for President, Inc.). Then, we compare this information with the Google database (FEC ID and exact

[8] https://cloud.google.com/vision/docs/ocr
[9] https://deepai.org/machine-learning-model/densecap
[10] https://pypi.org/project/langdetect/

T1: Liberal/Conservative				
	Train	**Dev**	**Test**	**Total**
C	2,576 (58%)	369 (69%)	453 (75%)	3,398 (61%)
L	1,835 (42%)	165 (31%)	150 (25%)	2,150 (39%)
All	4,411 (79.5%)	534 (9.6%)	603 (10.9%)	5,548 (100%)
Start	05-31-18	02-01-20	07-04-20	-
End	01-30-20	06-30-20	10-10-20	-
T2: Political Party/Third-Party				
	Train	**Dev**	**Test**	**Total**
PP	4,663 (39%)	324 (21%)	561 (37%)	5,548 (37%)
TP	7,427 (61%)	1,188 (79%)	953 (63%)	9,568 (63%)
All	12,090 (80%)	1,512 (10%)	1,514 (10%)	15,116 (100%)
Start	05-31-18	04-14-20	07-20-20	-
End	04-13-18	07-19-20	10-11-20	-

Table 4.3: Data set statistics for Task 1: *Conservative* (C)/ *Liberal* (L), and Task 2: *Political Party* (PP)/ *Third-Party* (TP).

name), to assign the corresponding affiliation to the sponsors. For example an ad sponsored by the 'Donald J. Trump for President, Inc.' official committee is labeled as *Republican* and subsequently as *Conservative* (in a similar way we label ads for the *Liberal* class).

In total, we collect 242 unique sponsors corresponding to 5,548 ads. *Liberal* ads represent the 39% of the total ads and the rest are *Conservative* (61%).

4.3.4 Labeling Ads with Sponsor Type

We first label all ads from sponsors that have an associated FEC ID in the Google database as *Political Party*. These sponsors correspond to official political committees affiliated with the Democratic or Republican parties (e.g. Biden for President).

Third-party sponsors of political ads consist of groups not officially associated to any political party such as not-for-profit organizations (e.g. NRDC Action Fund) and businesses (Fowler et al., 2020b). This type of sponsors are identified with their EIN ID (included in the Google database). Thus, we label all ads linked to an EIN ID as *Third-Party*. We collected a total of 15,116 ads where 37% corresponds to *Political Party* and 63% corresponds to *Third-Party*.

Avg. Tokens (Train/Dev/Test)			
Task	**IT**	**D**	**IT+D**
T1	17.1/16.5/17.1	38.3/39.9/36.9	55.4/56.4/54.0
T2	16.2/17.6/19.2	36.7/38.9/37.2	52.9/56.5/56.4

Table 4.4: Average number of tokens in image text (IT), densecaps (D) and both (IT+D) for sponsor ad ideology (T1) and type (T2) prediction.

4.3.5 Data Splits

We split both datasets chronologically into train (80%), development (10%), and test (10%) sets. Table 4.3 shows the dataset statistics and splits for each task.

4.3.6 Data Preprocessing

Text We normalize the text from the image (IT) and the densecap (D) by lower-casing, and replacing all URLs and person names with a placeholder token. To identify the person names we use the Stanford NER Tagger (Finkel et al., 2005). Also, we replace tokens that appear in less than five ads with an 'unknown' token. We tokenize the text using the NLTK tokenizer (Bird et al., 2009). Table 4.4 shows the average number of tokens in IT and D for each data split.

Image Each image is resized to (300×300) pixels represented by red, green and blue color values. Each color channel is an integer in the range [0, 255]. The pixel values of all images are dived by 255 to normalize them in the range [0, 1].

4.4 Predictive Models

We experiment with textual, visual and multimodal models for political ad classification.

4.4.1 Linear Baselines

As baseline models, we use logistic regression with bag of n-grams and L2 regularization using (1) the image text (LR_{IT}); (2) densecap (LR_{D}); and (3) their concatenation (LR_{IT+D}) for representing each ad.

4.4.2 BERT

We also test three models proposed by Kalra et al. (2020) for generic ad classification demonstrating state-of-the-art performance. The models are based on Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019) using a combination of the image text and the densecap. We follow a similar approach and fine-tune BERT for predicting the corresponding class in each task by adding an output dense layer for binary classification that receives the 'classification' [CLS] token as input. We use three types of inputs for each ad: (1) image text (BERT_{IT}); (2) densecap (BERT_{D}); and (3) their concatenation (BERT_{IT+D}).

4.4.3 EfficientNet

EfficientNet (Tan and Le, 2019) is a family of Convolutional Neural Network (CNN) (LeCun et al., 1995) models which has achieved state-of-the-art accuracy on ImageNet (Deng et al., 2009). In particular, we use EfficientNet-B3 and fine-tune it on political ad classification by adding an output dense layer for each binary classification task.

4.4.4 BERT+EffN

We finally test two multimodal models by combining: (1) BERT_{IT} and EfficientNet (BERT_{IT}+EffN); and (2) BERT_{IT+D} and EfficientNet (BERT_{IT+D}+EffN). We concatenate the text representation obtained by BERT and the visual information from EfficientNet into a $768 + 1536$ dimensional vector from BERT and EfficientNet respectively. This vector is then passed to an output layer for binary classification. We fine-tune the entire architecture for each task.

4.5 Experimental Setup

We select the hyperparameters for all neural models using early stopping by monitoring the validation binary cross-entropy loss, and we estimate the class weights using the 'balanced' heuristic (King and Zeng, 2001) for each task, as both datasets are imbalanced. BERT and EfficientNet models use ADAM optimizer (Kingma and Ba, 2014), and experiments use 1 GPU (Nvidia V100).

LR For LR we use bag of n-grams with $n = (1, 3)$, $n \in \{(1,1),(1,2),(1,3)\}$ weighted by TF.IDF and L2 regularization. The average training time is 30 seconds.

BERT We fine-tune BERT 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 implementation (12-layer 768-dimensional) trained on English Wikipedia (Wolf et al., 2019a). The maximal sequence length is 512 tokens. We fine-tune BERT for 2 epochs and learning rate $\eta = 2e^{-5}$ for ideology prediction; and $\eta = 1e^{-5}$ for advertiser type prediction with $\eta \in \{1e^{-5}, 2e^{-5}, 3e^{-5}, 4e^{-5}\}$. The average training time is 8.1 minutes.

EfficientNet We use EfficientNet-B3 with Noisy-Student weights (Xie et al., 2020). For ideology prediction, we first freeze the layers of the EfficientNet (Tan and Le, 2019) model and train it for 11 epochs with learning rate $\eta = 1e^{-3}$ to learn the parameters of the output layer. We then unfreeze and train the whole network for another 30 epochs with $\eta = 1e^{-4}$, as it has been shown that unfreezing the CNN during the latter stages of training improves the performance of the network (Faghri et al., 2017). For predicting the type of sponsor, we train for 45 epochs and $\eta = 1e^{-2}$ keeping the EfficientNet layers frozen. Unfreezing the base model did not result into lower validation loss. We use dropout rate of 0.2 before passing the output of EfficientNet to the classification layer. The average training time is 37.8 minutes.

BERT+EffN For ideology prediction, we freeze all the layers of the pre-trained models (BERT and EfficientNet) apart from the classification layer and train for 27 epochs with $\eta = 1e^{-3}$. We then fine-tune BERT for 30 epochs with $\eta = 1e^{-5}$. For sponsor type prediction, we freeze all EfficientNet layers and fine-tune BERT for 30 epochs with $\eta = 2e^{-6}$. We train

T1: Conservative/Liberal			
Model	**P**	**R**	**F1**
Majority	50.00 (0.00)	37.56 (0.00)	42.90 (0.00)
LR_D	55.76 (0.85)	54.91 (0.89)	54.85 (1.12)
LR_{IT}	78.38 (0.70)	71.99 (0.56)	72.65 (0.73)
LR_{IT+D}	72.57 (1.03)	71.52 (0.62)	71.99 (0.79)
Kalra et al. (2020)			
$BERT_D$	59.40 (0.78)	57.77 (0.98)	57.64 (1.52)
$BERT_{IT}$	72.88 (0.24)	73.46 (0.16)	73.16 (0.20)
$BERT_{IT+D}$	78.62 (3.14)	74.08 (2.81)	75.49 (3.01)
EfficientNet	69.02 (3.48)	67.87 (1.23)	68.15 (1.89)
Ours			
$BERT_{IT}$+EffN	74.99 (1.23)	72.01 (2.27)	73.02 (2.07)
$BERT_{IT+D}$+EffN	**80.24 (0.06)**	**74.59 (1.70)**	**75.76 (2.19)**

Table 4.5: Macro Precision (P), Macro Recall (R), and Macro F1-Score (F1) for political ideology prediction (± std. dev. for 3 runs). Best results are in bold.

in stages to ensure that the parameters of each part of the model (textual and visual) are properly updated (Kiela et al., 2019). The average training time is 56.65 minutes.

4.6 Results

This section presents the experimental results for the two predictive tasks, political ideology and sponsor type prediction (Section 4.3) using the methods described in Section 4.4. We evaluate our models using macro precision, recall and F1 score since the data in both tasks is imbalanced. Note that for all models we report the average and standard deviation over three runs using different random seeds. We also report the majority class baseline for each task.

4.6.1 Predictive Performance

Task 1: Conservative/Liberal Table 4.5 shows the results for the political ideology prediction. We first observe that $BERT_{IT}$ (73.16%) which uses as input information the image text outperforms $BERT_D$ (57.64%) and EfficientNet (68.15%) in macro F1. This suggests that the text shown on a political ad is the dominant medium for conveying its main message, corroborating findings in related research on commercial ads (Dey et al., 2021; Kalra et al., 2020).

T2: Political Party/Third-Party			
Model	**P**	**R**	**F1**
Majority	50.00 (0.00)	31.47 (0.00)	38.62 (0.00)
LR_D	53.60 (0.72)	53.40 (0.65)	53.11 (0.58)
LR_{IT}	84.02 (0.14)	85.04 (0.31)	84.47 (0.18)
LR_{IT+D}	86.46 (0.13)	86.63 (0.09)	86.54 (0.05)
Kalra et al. (2020)			
$BERT_D$	56.50 (0.89)	56.31 (0.78)	53.45 (1.26)
$BERT_{IT}$	85.57 (0.86)	86.42 (2.01)	85.86 (1.23)
$BERT_{IT+D}$	87.00 (0.89)	86.81 (0.83)	86.90 (0.86)
EfficientNet	53.27 (2.86)	53.93 (2.40)	51.53 (5.46)
Ours			
$BERT_{IT}$+EffN	**87.02 (2.74)**	85.81 (0.20)	86.29 (1.11)
$BERT_{IT+D}$+EffN	86.78 (0.03)	**88.18 (1.10)**	**87.36 (0.39)**

Table 4.6: Macro Precision (P), Macro Recall (R), and Macro F1-Score (F1) for sponsor type prediction (± std. dev. for 3 runs). Best results are in bold.

Moreover, combining image text and densecap ($BERT_{IT+D}$), leads to higher performance, than using only image text ($BERT_{IT}$), i.e. 75.49% and 73.16% F1 respectively. This indicates that the combination of textual with visual information (in the form of image descriptions) improves the model performance. Finally, using all visual information sources, i.e. densecaps and image representation from EfficientNet ($BERT_{IT+D}$+EffN), further improves performance achieving the highest macro F1 (75.76%) across models, followed by $BERT_{IT+D}$ (75.49%).

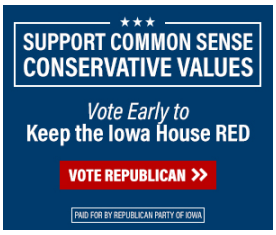
Task 2: Political-Party/Third-Party Table 4.6 shows the results for the sponsor type prediction. The best overall performance is obtained by $BERT_{IT+D}$+EffN (87.36%) which combines both image and textual information. $BERT_{IT+D}$ (86.90%) and LR_{IT+D} (86.54%) follow very closely. By inspecting our data, we identified the presence of noise in image text, particularly sentences are interrupted by logos and other aesthetic elements. This negatively affects the performance of BERT because such models are usually pre-trained on 'cleaner' generic corpora Kumar et al. (2020). On the other hand, LR models trained from scratch can adapt to the noisy text (see Section 4.6.2 for error analysis).

Overall, our results in both tasks suggest that text is a stronger modality for inferring the political ideology and sponsor type of political ads compared to visual information extracted from the images. However, integrating visual information in the form of text descriptions (densecaps) or representations obtained by pre-trained image classification models, enhances model performance.

(a) True: *Lib* - Pred: *Cons*

(b) True: *Cons* - Pred: *Lib*

(c) True: *PP* - Pred: *TP*



(d) True: *TP* - Pred: *PP*

Figure 4.1: Examples of ads with their true and predicted labels Lib (Liberal), Cons (Conservative), PP (Political Party), TP (Third-Party).

4.6.2 Error Analysis

We further perform an error analysis to examine the behavior of our best performing models (BERT$_{IT+D}$+EffN and BERT$_{IT+D}$) and identify potential limitations.

The ad shown in Figure 4.1 (a) was misclassified as *Conservative* by BERT$_{IT+D}$ and BERT$_{IT+D}$+EffN. This particular ad requires common knowledge of social issues (e.g. inadequate health support) that are often discussed in political campaigns to inform voters about a party's views on the issue (Scammell and Langer, 2006). This makes the classification task difficult for the models since it requires contextual knowledge. Incorporating external relevant knowledge to the models (e.g. political speeches, interviews or public meetings) might improve performance (Lin et al., 2018).

The ad depicted in Figure 4.1 (b) was misclassified by BERT$_{IT+D}$ and BERT$_{IT+D}$+EffN as *Conservative*. After analyzing the densecap descriptions, we found that this information tends to be noisy. For this particular example, it contains descriptions such as 'a man is holding a horse', 'the sign is blue', 'a blue and white stripe shirt', and 'a man wearing a hat'. In fact, BERT$_{IT}$, which only takes the image text into account, classified this ad correctly as *Conservative*. Improving the quality of the image descriptions (e.g. pre-training on advertising or political images, capturing specific attributes such as 'military hat') might be beneficial for these models.

Figure 4.1 (c) shows an example of a *Political Party* ad misclassified by BERT$_{IT+D}$+EffN as *Third-Party*. The ad contains the following text:

WE CAN'T LET <person> WIN! VOTE EARLY

The message has a confrontational and divisive tone that is common in *Third Party* ads (Edelson et al., 2019), but is typically used as a political tactic for negative campaigning (Skaperdas and Grofman, 1995; Gandhi et al., 2016; Haselmayer, 2019).

Finally, Figure 4.1 (d) shows an example of a *Third-Party* ad misclassified as *Political Party* by BERT$_{IT+D}$+EffN. The text content promotes voter participation (e.g. *Vote*), a characteristic of *Political Party* advertising (see Table 4.8). However, one of the aims of the *Third-Party* advertising is precisely to encourage voting and activism (Dommett and Temple, 2018).

There is a considerable difference between the models using visual information only (LR$_D$, BERT$_D$, EfficientNet), and those that also use the ad text as input (IT, IT+D). Our intuition is that models get confused by the appearance of shapes, colors and other aesthetic features that are domain specific and appear frequently in political advertisements (Sartwell, 2011). For instance, several ads that belong to the *Third-Party* category, include buttons linking to websites (see Fig, 4.1 (c), (d)). However, *Political Party* ads, also make use of these type of buttons to link users to donation or informative websites (Edelson et al., 2019).

4.7 Linguistic Analysis

We perform an analysis based on our new data set to study the linguistic characteristics of political ads. We first analyze the specific features of each class for both tasks. For this purpose, we use a method introduced by Schwartz et al. (2013) to analyze uni-gram features from image text (see Section 4.4) using univariate Pearson correlation. Features are normalized to sum up to unit for each ad. For each feature, we compute correlations independently between its distribution across ads and its label (*Conservative*/*Liberal*, or *Political Party*/*Third Party*).

4.7.1 Conservative vs. Liberal

Table 4.7 presents the top unigrams correlated with *Liberal* and *Conservative* ads. We first notice that the top words in the *Conservative* category are closely related to its ideology such as 'conservative' and 'republican'. Other prominent terms in these categories are words related to current political issues, such as immigration (e.g. 'border') and taxation (e.g. 'taxes'). In fact, these are examples of emotionally evocative terms (e.g. anger about taxes) that are

Liberal		Conservative	
Feature	**r**	**Feature**	**r**
necessary	0.197	senate	0.271
end	0.196	republican	0.196
prohibited	0.190	!	0.176
approx	0.186	conservative	0.127
contrib	0.181	national	0.116
void	0.177	committee	0.112
values	0.173	petition	0.109
prz	0.161	border	0.102
subj	0.156	taxes	0.099
make	0.156	radical	0.098
win	0.144	sign	0.096
place	0.140	stop	0.094
beer	0.139	states	0.093

Table 4.7: Feature correlations with *Conservative/Liberal* Ads, sorted by Pearson correlation (r). All correlations are significant at $p < .01$, two-tailed t-test.

frequently used in political campaigns to influence voters (Brader, 2005).

Top terms of *Liberal* ads include 'necessary', 'end','values', and 'win'. For example, the following ads belong to the *Liberal* class:

> *I'm supporting <person> because he has the same* ***values*** *that I do and he's an honest person.*

> *<person> FOR CONGRESS To* ***End*** *Gun Violence*

These are examples of ads containing a combination of moral and controversial topics (e.g. gun regulation) which are typical characteristics of political advertising (Kumar and Pathak, 2012).

4.7.2 Political Party vs. Third-Party

Table 4.8 shows the top unigram features correlated with the sponsor type of an ad (*Political Party/Third-Party*). We observe that some top terms in the *Political Party* class also belong to the top terms of the political ideology task (see Table 4.7) such as 'committee', 'republican'

Political Party		Third-Party	
Feature	**r**	**Feature**	**r**
congress	0.365	state	0.193
vote	0.308	learn	0.181
senate	0.292	champion	0.175
!	0.269	senator	0.166
president	0.248	thank	0.153
committee	0.236	action	0.147
candidate	0.223	congressman	0.130
republican	0.208	urge	0.129
authorized	0.208	protect	0.128
donate	0.202	access	0.119
join	0.199	award	0.117
<url>	0.187	american	0.116
$	0.180	?	0.113

Table 4.8: Feature correlations with Political Party/*Third-Party* Ads, sorted by Pearson correlation (r). All correlations are significant at $p < .01$, two-tailed t-test.

and 'senate'. Messages calling for vote and donation support ('vote', 'donate', '$') are also prevalent in *Political Party* ads (Fulgoni et al., 2016), as in the next example (See Figure 4.1 (b)):

Making sure our veterans get the care they've earned ***VOTE*** *FOR <person>*

On the other hand, top features from the *Third-Party* category (e.g. 'action', 'protect') share common characteristics with the rhetoric used by media outlets focused on promoting specific political messaging (Edelson et al., 2019; Dommett and Temple, 2018). Many of these ads direct people to websites to read about a particular topic. For example:

Is <person> HIDING ANTI-GUN VIEWS? ***Learn*** *More*

This ad belongs to the *Third-Party* class and points the viewer to an external website for reading further details.

T2: Political Party / Third Party	**Acc**	**F1**
Majority	60.7 (0.0)	37.8 (0.0)
Model	**87.6** (1.2)	**87.1** (1.3)
Human	74.7 (2.6)	63.1 (1.8)

Table 4.9: Accuracy and Macro F1-Score (F1) for sponsor type prediction (± std. dev. for 3 runs) including human performance on a sample of ads from the test set. Best results are in bold.

4.8 Human Evaluation

We conducted an analysis of human performance in the sponsor type classification task, i.e., categorizing an ad as either *Political Party* or *Third Party* based on the sponsor type. We carefully plan our experimental setup to ensure the reliability of our evaluation process. We sample ads from the test set in a manner that reflects the diversity of advertisers, resulting in 285 ads, constituting 18.8% of the test set.

We ask two annotators to assign the label *Political Party* or *Third Party* to an ad. The inter-annotator agreement, measured by Cohen's Kappa (Cohen, 1960), is 0.56, indicating a moderate level of agreement. In Table 4.9, we present the average accuracy and macro F1 scores across participants. Additionally, we include a majority label baseline, where the most common label, *Third Party*, is chosen, and the performance of our best performing model, BERT$_{IT+D}$+EffN.

We observe that the human macro F1 performance, at 63.1, surpasses the majority baseline of 37.8 but is lower than the model's macro F1 performance of 87.1. This discrepancy suggests that our model is able to identify subtle patterns or features within the data that humans may overlook or find challenging to discern. Furthermore, unlike humans, who may be biased by personal beliefs or opinions, our model maintains a focus on objectively analyzing the data, contributing to more impartial decision-making. These findings highlight the relevance of this task in enhancing the transparency of political campaigns, specifically, in ensuring the explicit disclosure of the type of sponsor of the ads.

4.9 Conclusion

We have presented the first study in NLP for analyzing political ads motivated by prior studies in political communication. We have introduced two new publicly available datasets containing political ads from the U.S. in English labeled by (1) the ideology of the sponsor (*Conservative*/*Liberal*); and (2) the sponsor type (*Political Party*/*Third Party*). We have defined both tasks as advertisement-level binary classification and evaluated a variety of approaches, including textual, visual and multimodal models reaching up to 75.76 and 87.36 macro F1 in each task respectively. Our results suggest that text is a stronger modality for inferring the political ideology and sponsor type of a political advertisement compared to image-based features. However, the inclusion of visual information in the form of text descriptions or image-encoder features, improves the performance of the models.

In the future, we plan to incorporate other modalities such as speech, and video, and explore other methods of acquiring and integrating multimodal information. In addition, we aim to extend our work for analyzing political advertising discourse across different regions, languages and platforms. Finally, we defer the extension of our linguistic analysis to examine the topics and intent underlying political advertisements as future work. Specifically, we aim to delve into aspects such as engagement, fundraising, or voting, drawing inspiration from prior research in the domain of political communication (Stromer-Galley et al., 2021).

Acknowledgments

We would like to thank Kate Dommett, Alexandra Boutopoulou, Mali Jin, Katerina Margatina, George Chrysostomou, Peter Vickers, Emily Lau, and all reviewers for their valuable feedback. DSV is 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. NA is supported by a Leverhulme Trust Research Project Grant.

Ethics Statement

Our work complies with the Terms of Service of the Google Political Ads Dataset.[11] We provide, for reproducibility purposes, the list of ad IDs and corresponding labels used for each task, as well as the data splits (train, development, test). All data used in this paper is in English. The ads information can be retrieved from Google according to their policy. For the human evaluation task, we received ethics approval from our Institution. Reference Number: 57526.

[11] https://console.cloud.google.com/marketplace/product/transparency-report/google-political-ads?pli=1

Chapter 5

Publication IV: A Multimodal Analysis of Influencer Content on Twitter

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, 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]

5.1 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 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 terms[3](Wojdynski, 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 or both (Sánchez Villegas et al., 2021). Figure 5.1 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.

[1]**Data is available here:** https://github.com/danaesavi/micd-influencer-content-twitter

[2]https://icas.global/advertising-self-regulation/influencer-guidelines/

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

Commercial: Quick & easy fried rice w/shrimp - perfect for busy weeknights and meets all of requirements for a fast, tasty meal that's sure to satisfy the whole family! #ZENSlifestyle

Non-commercial: Love coffee cake and donuts? Put two of your favorites together and make these mouthwatering Coffee Cake Donuts for breakfast this weekend! #donuts #coffeecake

Figure 5.1: *Commercial* and *non-commercial* tweets in our dataset. The distinction between *commercial* and *non-commercial* posts is frequently uncertain.

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; Wojdynski 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

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 5.1: A comparison of existing datasets for influencer content analysis

detecting *commercial* content, and provide insights into when each modality is beneficial.

5.2 Related Work

5.2.1 Computational Studies on Influencers

Previous work has analyzed the characteristics of influencers on social media platforms such as Twitter (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.

5.2.2 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.

Limitations of existing resources Table 5.1 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]

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

5.3 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.

5.3.1 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 (Alassani and Göretz, 2019; Weber et al., 2021). To retrieve influencers, we query for accounts that contain domain-specific keywords in 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.

5.3.2 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).

[5] https://influencermarketinghub.com/influencer-marketing-benchmark-report-2021/
[6] Influencer accounts were manually validated to ensure bots are not included.
[7] https://developer.twitter.com/en/products/twitter-api

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 affiliate marketing (e.g., *#aff*, *discount code*). A complete list of keywords can be found in Appx. 5.9. 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. 5.5.1) from the posts after labeling them. As a result, our models can identify *commercial* content without the use of such terms (see Sec. 5.4).

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.

5.3.3 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]

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

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

Table 5.2: Number of influencer accounts by domain

5.3.4 Human Data Annotation

To ensure a high quality data set for evaluation, we use human annotators for labeling all tweets in both 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 0.73 Krippendorff's alpha (Krippendorff, 2018) that corresponds to the upper part of the *substantial* agreement band (Artstein and Poesio, 2008). 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 5.3 shows the distribution of *commercial* and *non-commercial* tweets by split.

The training data is labeled using automatic weak labels, involving the removal of keywords employed for data labeling from tweets. This strategy is implemented to encourage the models to capture stylistic distinctions in both text and images between *commercial* and *non-commercial* content. While this approach may introduce bias since only tweets without commercial keywords are used for training, it is noteworthy that the test sets rely solely on

[9]We received approval from the Ethics Committee of our institution. Annotation guidelines can be found in Appx. 5.10.

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

human annotations rather than weak labels. This results in the creation of test sets that are not only more challenging but also closely aligned with real-world scenarios, specifically instances where influencers may or may not disclose their promotion of products or services.

5.3.5 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 5.2 shows the number of influencer accounts per domain. In average, each domain contains 22 accounts, and 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.

5.3.6 Comparison with Related Datasets

Table 5.1 compares our dataset, MICD, to related datasets for influencer content analysis (see Sec. 5.2). 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. 5.3.2).

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.

5.4 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*).

5.4.1 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 examples[10] (two examples from each class) before each prompt (few-shot). We run this three times 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) **ResNet**152 (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**).

[10]Appx. 5.12 includes the template we use for these prompts.

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.

5.4.2 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. 5.5.1). ViLT and LXMERT are multimodally pre-trained on visual-language tasks such as image-text matching and 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: $\text{Cross-Att}(L, I) = \text{softmax}(\frac{[W^Q L][W^K I]^T}{\sqrt{d_k}})[W^V I]$, 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 (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]

5.5 Experimental Setup

5.5.1 Data Processing

Text For each tweet, we lowercase and tokenize text using DLATK (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. 5.3.2) are removed from all *commercial* tweets.

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).

5.5.2 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-averaged[12] F1, precision, and recall to manage imbalanced classes. Results are obtained over three runs using different random seeds reporting average and standard deviation.

[11] Figure 5.4 shows a diagram of the model.

[12] Macro-averaged results are included in Appx. 5.11.

5.5.3 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., 2019b). For GPT-3 (Brown et al., 2020), we use the *text-davinci-003* model via the OpenAI[13] Library. Prompt templates are included in Appx. 5.12.

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. 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., 2019b), 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.

[13] https://platform.openai.com/docs/

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.

5.6 Results

Table 5.4 presents the performance on *commercial* and *non-commercial* influencer content prediction of all predictive models on our new multimodal influencer content dataset (MICD).

5.6.1 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 previous 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.

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)	$\mathbf{77.50}^{\ddagger}_{0.6}$	$\mathbf{78.46}^{\dagger}_{0.5}$	$\mathbf{77.61}^{\ddagger}_{0.6}$

Table 5.4: 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.

5.6.2 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

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)	$\mathbf{77.50}_{0.6}$	$\mathbf{78.46}_{0.5}$	$\mathbf{77.61}_{0.6}$

Table 5.5: 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.

answering (Zhou et al., 2020b; Lu et al., 2019a) 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 Preoţiuc-Pietro, 2019; Hessel and Lee, 2020). 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.

5.6.3 Ablation Study

To analyze the contribution of each component of our ViT-BERTweet-Att in identifying *commercial* posts, Table 5.5 shows the performance of ViT, BERTweet, and ViT-BERTweet-Att with and without the Cross-Att layer (see Section 5.4). 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

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)	$\mathbf{88.69}_{0.2}$	$\mathbf{88.69}_{0.2}$	$\mathbf{88.93}_{0.5}$

Table 5.6: 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.

relevant regions in images for identifying *commercial* posts.

5.6.4 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 since not all posts contain both modalities. Thus, we further evaluate our models on our text-only test set (see Section 5.3.3). Table 5.6 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.

5.6.5 Cross-domain Experiments

Table 5.7 presents the predictive performance (macro F1 score) of models trained on tweets (text-image pairs) from one domain and tested on all tweets from other domains using ViT-BERTweet-Att model. We observe that predictive performance is related to the proximity among domains. For example, tweets trained on 'Fitness' and tested on 'Lifestyle' obtain high performance (74.48 F1), as well as trained on 'Travel' and tested on 'Food' (71.72 F1). In

Test / **Train**	FT	FD	LS	TCH	TR	BT
FT	-	66.72	74.48	70.89	54.94	65.19
FD	68.94	-	70.67	72.55	56.99	70.68
LS	72.10	69.78	-	73.79	58.46	69.75
TCH	52.64	45.48	54.68	-	57.76	49.49
TR	71.54	71.72	81.99	76.97	-	66.83
BT	69.66	74.58	73.92	58.14	69.19	-

Table 5.7: Macro F1-Score performance of models trained with tweets from one domain and tested on other domains: 'Fitness' (FT), 'Food' (FD), 'Lifestyle' (LS), 'Tech' (TCH), 'Travel' (TR), 'Beauty' (BT).

general, we observe that 'Tech' obtains the lowest performance across domains, which might be because there is a small number of examples in the dataset compared to other domains. Moreover, the vocabulary of posts shared by tech influencers is particularly specialized to the products they promote such as technical words to describe product specifications. On the other hand, lifestyle influencers share content around many different topics including makeup, fitness, and cooking (Thelwall, 2021). This is reflected in the results, with 'Lifestyle' performing well across all domains.

5.7 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 5.1. Similarly, for text-only posts, we observe that 69% of BERTweet missclassifications

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 5.2 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 approach[14] to the test sets (see Section 5.3.2). 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 misslabeled 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 5.1 (*commercial*) and Figure 5.2 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

[14]Using the text before removing commercial keywords.

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 5.2: Examples of classifications of BERTweet and ViT-BERTweet-Att.

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 5.2 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).

5.8 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 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 5.7.

Ethics Statement

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

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 5.3.2). 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.

[15] See: https://developer.twitter.com/en/developer-terms/agreement-and-policy

5.9 Appendix 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 5.8.

5.10 Appendix 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:

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
Endoresements	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 5.8: Commercial keywords. @USER refers to an @-mention of a brand account.

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		☑	☑	☐	LINK	

Figure 5.3: Example of Annotation

- 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

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

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).

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

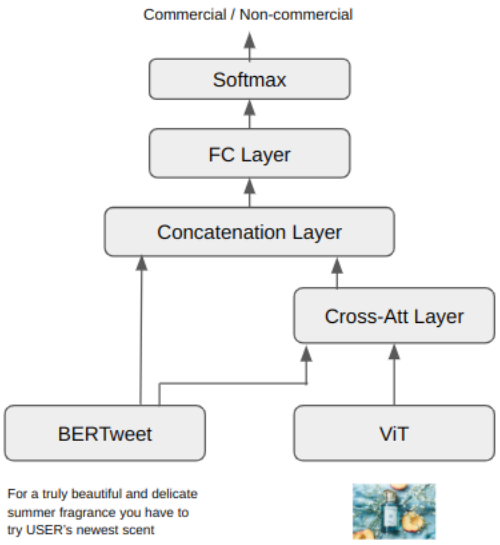


Figure 5.4: ViT-BERTweet-Att model for detecting *commercial* content. FC: fully-connected layer.

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 backgrounds 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.

5.11 Appendix C: Predictive Performance

Table 5.10 and Table 5.11 present the macro-averaged results of *commercial* content prediction.

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 5.10: 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 5.11: 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.

5.12 Appendix D: Prompt Templates

5.12.1 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.

5.12.2 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 5.4 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> //

<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*).

Chapter 6

Publication V: Improving Multimodal Classification of Social Media Posts by Leveraging Image-Text Auxiliary Tasks

Improving Multimodal Classification of Social Media Posts by Leveraging Image-Text Auxiliary Tasks

Danae Sánchez Villegas[α], *Daniel Preoţiuc-Pietro*[β], *Nikolaos Aletras*[α]

[α] Computer Science Department, University of Sheffield, UK

[β] Bloomberg

Abstract

Effectively leveraging multimodal information from social media posts is essential to various downstream tasks such as sentiment analysis, sarcasm detection or hate speech classification. Jointly modeling text and images is challenging because cross-modal semantics might be hidden or the relation between image and text is weak. However, prior work on multimodal classification of social media posts has not yet addressed these challenges. In this work, we present an extensive study on the effectiveness of using two auxiliary losses jointly with the main task during fine-tuning multimodal models. First, Image-Text Contrastive (ITC) is designed to minimize the distance between image-text representations within a post, thereby effectively bridging the gap between posts where

the image plays an important role in conveying the post's meaning. Second, Image-Text Matching (ITM) enhances the model's ability to understand the semantic relationship between images and text, thus improving its capacity to handle ambiguous or loosely related posts. We combine these objectives with five multimodal models, demonstrating consistent improvements of up to 2.6 F1 score across five diverse social media datasets. Our comprehensive analysis shows the specific scenarios where each auxiliary task is most effective.

6.1 Introduction

Multimodal content including text and images is prevalent in social media platforms (Vempala and Preoţiuc-Pietro, 2019). The content of both text and images has been widely used to improve upon single modality approaches in various downstream tasks such as sentiment analysis (Niu et al., 2016; Ju et al., 2021; Tian et al., 2023b), hate speech detection (Botelho et al., 2021; Hossain et al., 2022; Cao et al., 2022; Ocampo et al., 2023) and sarcasm detection (Cai et al., 2019; Xu et al., 2020; Liang et al., 2022; Tian et al., 2023a).

Multimodal classification methods for social media tasks often combine text and image representations obtained from pre-trained models. These are usually pre-trained on standard vision-language data such as image captions where strong image-text connections are assumed, i.e., captions that explicitly describe a corresponding image (Hessel and Lee, 2020; Xu and Li, 2022). Modeling text-image pairs from social media posts presents additional challenges. A notable difficulty lies in effectively capturing latent cross-modal semantics that may not be apparent. Figure 6.1 (left) shows an example where the text refers specifically to the mood of the person in the photo (i.e., "unhappy feeling" *when @USER gets more followers...*). Moreover, cases where the visuals are weakly related to the text are also prevalent (Xu et al., 2022). For instance, Figure 6.1 (right) shows an image of a hen accompanied by the text *My baby approves*. It is difficult to draw a direct relationship between the two without any additional context.

Multimodal models for social media classification can be divided into: (1) *single-stream* models where image and text features are concatenated together and fed into the same module such as Unicoder (Li et al., 2020a), VisualBERT (Li et al., 2019a), ViLT (Kim et al., 2021c) and ALPRO (Li et al., 2022a); and (2) *dual-stream* approaches where images and text are processed separately, e.g., ViLBert (Lu et al., 2019b), LXMERT (Tan and Bansal, 2019),

METER (Dou et al., 2022) and BLIP-2 (Li et al., 2023). Consequently, these models might still suffer from the aforementioned issues.

In this work, we examine the use of two tasks – Image-Text Contrastive (ITC) and Image-Text Matching (ITM) – as auxiliary losses during fine-tuning for improving social media post classification. By using the ITC contrastive loss (He et al., 2020; Li et al., 2021; Yu et al., 2022), we anticipate that when the image contributes to the post's meaning, as illustrated in Fig. 6.1 (left), the model will place them closer in the representation space. Conversely, ITM leverages binary classification loss for image-text alignment (Chen et al.; Tan and Bansal, 2019; Wang et al., 2021). We expect that this will improve the model's ability to handle posts where associations may not be explicitly stated as shown in Fig. 6.1 (right). Although ITC and ITM have been used as pre-training objectives using generic images and their corresponding captions (Radford et al., 2021; Wang et al., 2021; Chen et al., 2022), their potential for enhancing fine-tuning in social media classification has yet to be explored.

Our main contributions are as follows: (1) we present an extensive study on comparing multimodal models jointly fine-tuned with ITC and ITM covering both *single-* and *dual-stream* approaches; (2) we show that models using ITC and ITM as auxiliary losses consistently improve their performance across five diverse multimodal social media datasets; (3) we offer a comprehensive analysis revealing the effectiveness of individual auxiliary tasks and their combination across various image-text relationship types in posts.

6.2 Multimodal Auxiliary Tasks

Image-Text Contrastive (ITC) Modeling text-image pairs in social media posts involves capturing hidden cross-modal semantics (Vempala and Preoţiuc-Pietro, 2019; Kruk et al., 2019). For instance, in Figure 6.1 (left) the visible mood of the person on the photo is related to the text of the post. Instead of directly matching images with textual descriptions (e.g., *a man wearing a helmet*), we aim to encourage the model to capture the dependencies between the image and text within the posts.

For this purpose, we use the ITC objective (He et al., 2020; Li et al., 2021; Yu et al., 2022) which pushes towards a feature space in which image and text representations of a post are brought closer together, while image and text representations that appear in different posts are pushed further apart. Let L_n and I_n be the n-th (normalized) representation of text and

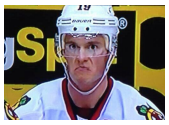
Post	 When @USER gets more followers than you in 12 hours	 My baby approves
Img-Txt Relation	The image adds to the meaning	The image does not add to the meaning
Caption	A close up of a hockey player wearing a helmet	A gray and white chicken standing in the dirt

Figure 6.1: Image-text relations in social media posts from Vempala and Preoţiuc-Pietro (2019) and corresponding image captions generated with InstructBLIP.

accompanying image of a post in a training batch. While the cosine similarity of the pair L_n and I_n is minimized, the cosine similarity of all other random pairs (e.g., L_n and I_m; I_m is an image from a different post in the current batch) is maximized. Given N posts within a training batch, ITC loss is defined as follows:

$$l_{ITC} = \frac{1}{2}(l_1 + l_2) \tag{6.1}$$

$$l_1 = -\frac{1}{N}\Sigma_{n=1}^{N} log \frac{exp(LI^T/e^\tau)}{\Sigma_{j=1}^{N} exp(LI^T/e^\tau)} \tag{6.2}$$

$$l_2 = -\frac{1}{N}\Sigma_{n=1}^{N} log \frac{exp(IL^T/e^\tau)}{\Sigma_{j=1}^{N} exp(IL^T/e^\tau)} \tag{6.3}$$

τ is a learnable temperature parameter to scale the logits (Jia et al., 2021).

Image-Text Matching (ITM) In social media posts, unrelated or weakly related text-image pairs are common (Hessel and Lee, 2020; Xu et al., 2022) such as the post depicted in Fig. 6.1 (right). To address this, we use the ITM objective (Chen et al.; Tan and Bansal, 2019; Wang et al., 2021) during fine-tuning to understand the semantic correspondence between images and text. ITM involves a binary classification loss that penalizes the model when a given text and image do not appear together in a post. Let I_n and L_n be the image and text representation of the n-th post in a training batch, we randomly replace I_n with an image of another post from the current batch with a probability of 0.5 following (Wang et al., 2021;

Kim et al., 2021c). If I_n is replaced, then the image and text do not match, otherwise I_n and L_n match. Thus, the ITM loss corresponds to the cross-entropy loss for penalizing incorrect predictions, $l_{ITM} = -\Sigma_{i=1}^{2} t_i log(p_i)$ where t_i is the gold label (matched or mismatched) and p_i is the softmax probability for each label.

Joint Fine-tuning Objectives The joint fine-tuning loss function includes the cross-entropy classification loss (l_{CE}) and the two auxiliary training objectives defined as: $l_{C+M} = \lambda_1 l_{CE} + \lambda_2 l_{ITC} + \lambda_3 l_{ITM}$, where $\lambda_1, \lambda_2, \lambda_3$ are hyperparameters to control the influence of each loss.

6.3 Experimental Setup

6.3.1 Datasets

We experiment with five diverse multimodal public datasets in English: (1) **TIR** – text-image relationship categorization (Vempala and Preoţiuc-Pietro, 2019); (2) **MVSA** – multi-view sentiment analysis (Niu et al., 2016); (3) **MHP** – multimodal hate speech detection (Gomez et al., 2020; Botelho et al., 2021); (4) **MSD** – multimodal sarcasm detection (Cai et al., 2019): and (5) **MICD** – multimodal commercial influencer content detection (Sánchez Villegas et al., 2023). Table 6.1 presents dataset statistics.

6.3.2 Data Splits

We use the same data splits for MVSA, MHP, MSD and MICD as in the original papers. For TIR, instead of a 10-fold cross-validation, we randomly split the data in 80%, 10%, and 10% for training, validation, and testing for consistency with the other tasks.

6.3.3 Data Processing

Text For each tweet, we lowercase and tokenize text using the NLTK Twitter tokenizer (Bird and Loper, 2004). We also replace URLs and user @-mentions with placeholder tokens.

Dataset	Classification Task	#	Train	Val	Test	All
TIR	Text-Image Relation Classification	4	3,575	447	449	4,471
MVSA	Sentiment Analysis	3	3,611	451	451	4,511
MHP	Hate Speech Classification	4	3,998	500	502	5,000
MSD	Sarcasm Detection	2	19,816	2,410	2,409	24,635
MICD	Influencer Commercial Content Detection	2	11,377	1,572	1,435	14,384

Table 6.1: Description and statistics of each dataset. # refers to number of classes.

Emojis are replaced with their corresponding text string, e.g thumbs_up following Nguyen et al. (2020).

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).

6.3.4 Single Modality Methods

Text-only We fine-tune **BERT** (Devlin et al., 2019) and **Bernice** (DeLucia et al., 2022), a BERT based model pre-trained on a corpus of multilingual tweets. We also experiment with few-shot (FS) prompting using **Flan-T5** (Chung et al., 2022) and **GPT-3** (Brown et al., 2020). For each dataset, we construct a few-shot prompt and include two randomly selected training examples for each class.[1]

Image-only We fine-tune **ResNet**152 (He et al., 2016) and **ViT** (Dosovitskiy et al., 2020), both pre-trained on ImageNet (Russakovsky et al., 2015). We experiment with few-shot prompting using **IDEFICS** (Laurençon et al., 2023) and zero-shot prompting using **InstructBLIP** (Dai et al., 2023).

[1]Appx. 6.8 shows the prompt templates.

6.3.5 Multimodal Models

Ber-ViT We use Bernice and ViT to obtain representations of the text (L) and image (I). **Ber-ViT-Conc** appends the text and image vectors from the corresponding L and I [CLS] tokens to obtain the multimodal representation h^{LI}; **Ber-ViT-Att** computes cross-attention between L and I. h^{LI} is obtained by appending the [CLS] token from L and the [CLS] token from the attention layer. We fine-tune each model by adding a classification layer.

MMBT (Kiela et al., 2019). Image embeddings obtained from Resnet152 are concatenated with token embeddings and passed to a BERT-like transformer. The [CLS] token is used as the multimodal representation (h^{LI}) for classification.

LXMERT (Tan and Bansal, 2019) consists of three encoders and their corresponding outputs for vision I, language L, and a multimodal vector h^{LI}.

ViLT We fine-tune ViLT (Dosovitskiy et al., 2020) and extract the multimodal h^{LI} that corresponds to the first token from the last hidden state.

ITC and ITM Inputs The ITC task inputs are the text and image vectors of each model. The ITM auxiliary task input is the respective multimodal representation h^{LI}.

6.3.6 Evaluation

Results are obtained over three runs using different random seeds reporting average and standard deviation. We use weighted F1 for model evaluation following standard practice on the TIR, MHP and MICD datasets to manage class imbalance.[2]

[2]Implementation details are included in Appx. 6.7.

Model	TIR	MVSA	MHP	MSD	MICD	▲
Majority Class	16.0 (0.0)	59.8 (0.0)	53.4 (0.0)	45.2 (0.0)	48.0 (0.0)	-
Text-only Models						
BERT	37.2 (1.3)	70.1 (0.8)	73.3 (1.3)	83.9 (0.2)	74.3 (0.6)	-
Bernice	38.9 (1.1)	71.6 (0.6)	73.6 (0.6)	84.5 (0.8)	74.5 (2.2)	-
Flan-T5*	3.8 (0.0)	58.9 (0.0)	46.5 (1.3)	59.6 (2.2)	48.7 (1.6)	-
GPT-3*	16.3 (6.1)	55.9 (0.1)	58.2 (4.6)	69.6 (2.7)	69.6 (1.5)	-
Image-only Models						
ResNet152	48.2 (0.0)	63.8 (0.1)	51.8 (5.8)	46.9 (0.1)	59.6 (0.5)	-
ViT	51.4 (1.3)	68.2 (0.6)	57.2 (1.2)	71.5 (0.1)	60.8 (1.3)	-
IDEFICS*	12.4 (3.6)	34.7 (6.1)	34.9 (2.7)	58.9 (2.4)	35.6 (0.0)	-
InstructBLIP*	3.9 (0.0)	47.2 (0.0)	11.0 (0.0)	22.7 (0.0)	35.6 (0.0)	-
Multimodal Models						
Ber-ViT-Conc	43.6 (1.2)	70.4 (0.0)	76.6 (0.6)	88.8 (0.0)	75.5 (1.9)	-
+ ITC	$44.9_{1.3}$ (0.7)	$72.0^{\dagger}_{1.6}$ (0.2)	$77.3_{0.7}$(1.1)	$\underline{\mathbf{89.7}}^{\dagger}_{0.9}$ (0.0)	$\underline{77.2}_{1.7}$ (0.4)	1.2
+ ITM	$44.1_{0.5}$ (0.2)	$\underline{73.6}^{\dagger}_{3.2}$ (0.9)	$\underline{77.8}_{1.2}$ (0.6)	$89.2^{\dagger}_{0.4}$ (0.1)	$76.1_{0.6}$ (0.8)	1.2
+ ITC + ITM	$\underline{45.8}_{2.2}$ (0.8)	$73.4^{\dagger}_{3.0}$ (0.4)	$77.7^{\dagger}_{1.1}$ (0.6)	$\underline{\mathbf{89.7}}^{\dagger}_{0.9}$ (0.2)	$76.3_{0.7}$ (0.5)	1.6
Ber-ViT-Att	53.7 (1.0)	72.1 (0.7)	76.8 (0.5)	88.8(0.3))	75.6 (0.8)	-
+ ITC	$54.8_{1.1}$ (0.8)	$72.8_{0.7}$ (0.2)	$77.5_{0.7}$ (0.6)	$89.5^{\dagger}_{0.7}$ (0.2)	$\underline{\mathbf{77.8}}^{\dagger}_{2.2}$ (0.1)	0.8
+ ITM	$\underline{\mathbf{55.9}}^{\dagger}_{2.2}$ (0.8)	$73.5^{\dagger}_{1.4}$ (0.2)	$77.4_{0.6}$ (0.6)	$89.4_{0.6}$ (0.5)	$76.6_{1.0}$(0.5)	1.2
+ ITC + ITM	$54.6_{0.9}$ (0.7)	$\underline{\mathbf{74.6}}^{\dagger}_{2.5}$ (0.3)	$\underline{\mathbf{78.0}}^{\dagger}_{1.2}$ (0.1)	$\underline{\mathbf{89.7}}^{\dagger}_{0.9}$ (0.3)	$76.3_{0.7}$ (0.2)	1.7
MMBT	53.2 (1.2)	72.4 (0.4)	74.5 (0.5)	83.2 (0.0)	73.6 (0.4)	-
+ ITC	$\underline{53.7}_{0.5}$ (1.1)	$73.2_{0.8}$ (1.0)	$75.7_{1.2}$ (1.7)	$\underline{84.4}^{\dagger}_{1.2}$ (0.3)	$74.1_{0.5}$ (0.8)	1.1
+ ITM	$\underline{53.7}_{0.5}$ (0.7)	$73.4_{1.0}$ (0.8)	$75.4_{0.9}$ (1.3)	$84.3^{\dagger}_{1.1}$ (0.3)	$\underline{74.8}^{\dagger}_{1.2}$ (0.6)	0.9
+ ITC + ITM	$53.6_{0.4}$ (0.2)	$\underline{73.5}^{\dagger}_{1.1}$ (0.0)	$\underline{75.7}_{1.2}$ (0.2)	$83.4_{0.2}$ (0.2)	$73.8_{0.2}$ (0.5)	0.6
LXMERT	51.3 (0.5)	68.2 (1.1)	70.7 (0.8)	81.9 (0.5)	69.9 (1.0)	-
+ ITC	$51.9_{0.6}$ (0.3)	$\underline{70.4}^{\dagger}_{2.2}$ (0.5)	$\underline{72.1}^{\dagger}_{1.4}$ (0.2)	$\underline{82.7}_{0.8}$ (0.1)	$70.8_{1.0}$ (0.5)	1.2
+ ITM	$51.8_{0.5}$ (0.4)	$69.5_{1.3}$ (0.2)	$71.8_{1.1}$ (0.8)	$82.3_{0.4}$ (0.5)	$\underline{70.9}_{1.1}$ (0.2)	0.9
+ ITC + ITM	$\underline{52.3}_{1.0}$ (1.4)	$69.3_{1.1}$ (0.9)	$71.9_{1.2}$ (1.7)	$82.1_{0.2}$ (0.4)	$70.3_{0.5}$ (0.3)	0.8
ViLT	53.1 (1.1)	70.5 (1.3)	71.8 (0.0)	83.0 (0.8)	67.8 (1.6)	-
+ ITC	$\underline{55.7}^{\dagger}_{2.6}$ (0.2)	$\underline{72.9}_{2.4}$ (1.0)	$72.5^{\dagger}_{0.7}$ (0.4)	$83.4_{0.4}$ (0.4)	$68.3_{0.5}$ (0.2)	1.3
+ ITM	$\underline{55.7}^{\dagger}_{2.6}$ (0.3)	$72.1_{1.6}$ (2.3)	$72.0_{0.2}$ (0.5)	$\underline{83.5}_{0.5}$ (0.2)	$68.7_{0.8}$ (1.1)	1.1
+ ITC + ITM	$55.3^{\dagger}_{2.2}$ (0.3)	$\underline{72.9}_{2.4}$ (1.3)	$\underline{73.4}_{1.6}$ (1.4)	$83.2_{0.2}$ (0.4)	$\underline{70.0}_{2.1}$ (1.3)	1.7

Table 6.2: Results in weighted F1 for all datasets. Best results for each base multimodal model are underlined and best results for each dataset are in bold. † indicates statistically significant improvement (t-test, $p < 0.05$) over the corresponding base model. Subscripts denote the relative increment over each base model and standard deviations are included in parenthesis. ▲ refers to the average relative improvement over each base model across datasets.* denotes prompting.

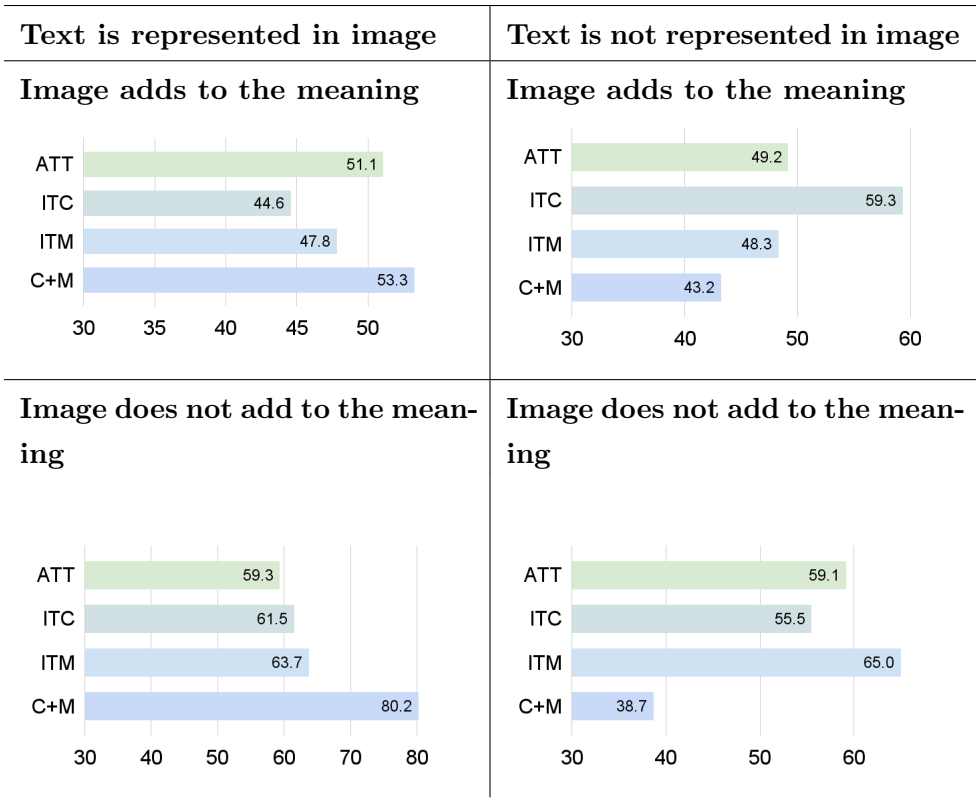


Figure 6.2: Accuracy per label using Ber-ViT-Att (ATT) across different image-text relation types based on image contribution to the post's meaning and text representation on the image. C+M refers to ITC+ITM.

6.4 Results

Image-text auxiliary tasks improve multimodal classification. Table 6.2 shows that multimodal models surpass single-modality approaches across datasets. We consistently find performance gains when using either ITC, ITM, or both auxiliary losses during fine-tuning, with improvements up to 2.6 F1 over each base model. Therefore, we can improve performance without costly pre-training on social media text-image tasks. These findings are especially valuable in multimodal computational social science studies, where grasping the interplay between text and images is vital (Hessel and Lee, 2020; Xu et al., 2022)

Dual-stream methods are effective in leveraging information from the auxiliary tasks. Across MVSA, MHP and MSD datasets, the Ber-ViT-Att+ITC+ITM model achieves the best performance (74.6, 78.0, and 89.7 F1 respectively). Generally, we observe that both

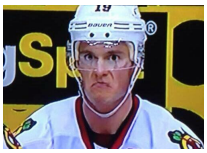
Text is represented in image	**Text is not represented in image**
Image adds to the meaning  *New Years Resolution.* ATT:✗ — ITC:✓ — ITM:✓ — C+M:✓	**Image adds to the meaning**  *When @USER gets more followers than you in 12 hours* ATT:✗ — ITC:✓ — ITM:✗ — C+M:✗
Image does not add to the meaning  *Babyface and Whitney Houston* ATT:✗ — ITC:✗ — ITM:✗ — C+M:✓	**Image does not add to the meaning**  *My baby approves* ATT:✗ — ITC:✗ — ITM:✓ — C+M:✗

Figure 6.3: Bert-ViT-Att (ATT) predictions on randomly selected examples with varying image-text relations.

ITC and ITM contribute to the performance improvements of Ber-ViT-Att. Overall, Ber-ViT-Att+ITC and Ber-ViT-Att+ITM models average improvements over the base model across datasets are 0.8 and 1.2 respectively, while Ber-ViT-Att+ITC+ITM improvement is 1.7. The performance gap between *dual-* and *single-stream* models is narrower in TIR. ViLT+ITM achieves 55.7 F1 while Ber-ViT-Att+ITM obtains 55.9. This is likely due to the importance of visual information for this task (i.e., predicting the semiotic relationship between images and text), which is better aligned with ViLT as a visual-based model.

6.5 Analysis

We analyze Ber-ViT-Att's predictions on TIR to understand when each auxiliary task benefits different image-text relations as categorized by Vempala and Preoţiuc-Pietro (2019) based on image contribution and text representation (Fig. 6.2 and 6.3).

When the text is represented in the image using both auxiliary tasks (models denoted with C+M), the model achieves the best performance, especially when the visual content is not semantically relevant to the post. We observe that 80.2% of the tweets are correctly classified achieving a substantial improvement over the Ber-ViT-Att baseline where only 59.3% of the posts are correctly classified.

When text is not represented on the image, we find that including ITC performs best when the visual content is relevant, with 59.3% of the tweets correctly classified compared to 49.2% using Ber-ViT-Att. Finally, in cases where the image does not enhance the semantic meaning, Ber-ViT-Att+ITM exhibits the highest performance, correctly classifying 65% of the posts. This validates our hypothesis that incorporating ITM helps models to effectively identify posts with weaker image-text relationships.

6.6 Conclusion

We presented an extensive study on the effectiveness of using two auxiliary tasks, Image-Text Contrastive (ITC) and Image-Text Matching (ITM) when fine-tuning multimodal models for social media posts classification. This approach addresses the challenges of hidden cross-modal semantics and weak image-text relationships in social media content. Future work includes evaluation on different social media platforms and languages.

Limitations

First, the datasets used in our experiments are solely in English. This choice allows for consistency and comparability across the datasets, but it does not test the generalizability of our findings to other languages. In future work, we plan to extend our research to a multilingual setting to address this limitation. The effectiveness of the models incorporating

auxiliary tasks depends on the underlying base model, however, our approach can easily be adapted to new models. Finally, the inclusion of auxiliary tasks in our models introduces an increase in training time. For instance, the training time for Ber-ViT-Att on the TIR dataset is approximately 1.5 hours on an Nvidia A100 GPU. However, when incorporating the auxiliary tasks (Ber-ViT-Att+ITC+ITM), the training time extends to around 2.5 hours, a 66% relative increase in training time.

Dataset	Classification Task	#	Train	Val	Test	All
POI	Point-of-interest Type Prediction	8	157,029	19,559	19,647	196,235
POLID	Political Ads Ideology Conservative or Liberal	2	4,411	534	603	5,548
POLADV	Political Ads Sponsor: Political Party or Third Party	2	12,090	1,512	1,514	15,116
MICD	Influencer Commercial Content Detection	2	11,377	1,572	1,435	14,384

Table 6.3: Description and statistics of each dataset: POI, POLID, POLADV and MICD. # refers to number of classes.

Additional Experiments

To comprehensively assess the effectiveness of our top-performing models — Ber-ViT-Att (late-fusion) and ViLT (early-fusion) — we conduct fine-tuning experiments on multimodal datasets for point-of-interest type prediction – **POI** (Sánchez Villegas et al., 2020; Sánchez Villegas and Aletras, 2021), predicting the ideology (conservative or liberal) of political ad sponsors – **POLID** (Sánchez Villegas et al., 2021), and predicting whether an ad was posted by an official political party or a third-party sponsor – **POLADV** (Sánchez Villegas et al., 2021). Additionally, we include the results on **MICD** (Sánchez Villegas et al., 2023), a dataset for identifying commercial influencer content. The dataset statistics for each dataset are presented in Table 6.3. We follow the same experimental setup as in Section 6.3. The results of these experiments are included in Table 6.4. To offer a comprehensive reference, we include the results of unimodal models — text-only models (BERT and Bernice) and image-only models (ResNet and ViT).

Multimodal models performance is superior compared to text-only and image-only models across datasets. When comparing the performance of the multimodal models, we observe that Ber-ViT-Att obtains higher performance than ViLT for all datasets except for POLID. Moreover, we observe the highest improvements in performance for all datasets except for POLID when including ITC auxiliary loss, with improvements in weighted F1 of up to 2.2 on the MICD dataset using Ber-ViT-Att. Particularly for the POLID dataset, the best performing model is observed when integrating both ITC and ITM losses on ViLT, resulting in a 6.1 weighted F1 improvement over vanilla ViLT. These findings highlight the importance

Model	POI	POLID	POLADV	MICD	▲
Text-only Models					
BERT	47.2 (0.4)	71.8 (0.6)	87.5 (0.2)	74.3 (0.6)	-
Bernice	44.7 (0.4)	73.7 (0.5)	87.8 (0.1)	74.5 (2.2)	-
Image-only Models					
ResNet	43.3 (0.1)	61.6 (1.0)	75.6 (1.6)	59.6 (0.5)	-
ViT	48.8 (0.3)	71.9 (0.1)	76.1 (0.7)	60.8 (1.3)	-
Multimodal Models					
Ber-ViT-Att	53.2 (0.8)	74.7 (1.2)	88.5 (0.2)	75.6 (0.8)	-
+ ITC	$\mathbf{54.4}_{1.2}$ (0.5)	$75.5_{0.8}$ (0.8)	$\mathbf{89.2}_{0.7}$ (0.4)	$\mathbf{77.8}^{\dagger}_{2.2}$ (0.1)	1.2
+ ITM	$53.5_{0.4}$ (0.2)	$75.6_{0.9}$ (1.0)	$88.6_{0.1}$ (0.1)	$76.6_{1.0}$ (0.5)	0.6
+ ITC + ITM	$53.9_{0.7}$ (0.4)	$76.5_{1.8}$ (0.4)	$89.0_{0.5}$ (0.3)	$76.3_{0.7}$ (0.2)	0.9
ViLT	48.9 (1.1)	76.4 (0.7)	84.0	67.8 (1.6)	-
+ ITC	$49.9_{1.0}$ (1.1)	$81.4^{\dagger}_{5.0}$ (1.4)	$85.8^{\dagger}_{1.8}$ (0.3)	$68.3_{0.5}$ (0.2)	2.1
+ ITM	$50.8_{1.9}$ (1.7)	$79.8^{\dagger}_{3.4}$ (1.1)	$85.3^{\dagger}_{1.3}$ (0.2)	$68.7_{0.8}$ (1.1)	1.9
+ ITC + ITM	$51.0_{1.1}$ (1.6)	$\mathbf{82.5}^{\dagger}_{6.1}$ (0.7)	$84.9_{0.9}$ (0.7)	$70.0_{2.1}$ (1.3)	2.6

Table 6.4: Results in weighted F1 for place type prediction (POI), Political Ads Analysis (POLID and POLADV), and Influencer Content Analysis (MICD) datasets. Best results for each dataset are in bold. Subscripts denote the relative increment over each base model and standard deviations are included in parenthesis. † indicates statistically significant improvement (t-test, $p < 0.05$) over the corresponding base model. ▲ refers to the average relative improvement over each base model across datasets.

of tailoring multimodal methods to accommodate the unique characteristics of social media posts. Specifically, addressing the challenges posed by hidden cross-modal semantics and nuanced image-text relationships is crucial for optimizing performance in this context.

6.7 Appendix A: Implementation details

6.7.1 Hyperparameters

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 are performed using an Nvidia A100 GPU with a batch size of 8 for TIR and MHP and 16 for MVSA and MSD datasets. For prompting implementation details see Appx. 6.8.

6.7.2 Unimodal Models

Image-only For ResNet152 (He et al., 2016), we fine-tune for 1, 5, 8, 6, 1, 3, 1 and 4 epochs for TIR, MVSA, MHP, MSD, MICD, POI, POLID and POLADV datasets respectively, 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 for TIR, MSD, MICD and POLADV, 10 epochs for MVSA and MHP datasets, and 4 epochs for POI and POLID datasets 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 Transformers We fine-tune BERT and Bernice 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 the Hugging Face library (12-layer, 768-dimensional) (Wolf et al., 2019b), and the base model for Bernice (DeLucia et al., 2022) with a maximal sequence length of 128. We fine-tune BERT for 3, 9, 5, 2 and 1, 1, 3, 1 epochs for TIR, MVSA, MHP, MSD, MICD, POI, POLID, and POLADV with learning rate $\eta = 1e^{-5}$ and dropout $\delta = 0.05$; and Bernice for 3, 4, 7, 3, 3, 4, 2 and 1 epochs for TIR, MVSA, MHP, MSD, MICD, POI, POLID, and POLADV datasets, $\eta = 1e^{-5}$ and $\delta = 0.05$. For all models $\eta \in \{2e^{-5}, 1e^{-4}, 1e^{-5}\}$ and $\delta \in [0, 0.5]$, random search.

6.7.3 Multimodal Predictive Models

We train MMBT (Kiela et al., 2019), ViLT (Kim et al., 2021c), LXMERT (Tan and Bansal, 2019) and Bernice-ViT models with λ_1, λ_2, λ_3; λ_2 and $\lambda_3 \in [0, 1.5]$ (as explained in Section 6.2), and number of fine-tuning epochs (E) for each model as shown in Tables 6.5 and 6.6. For ViLT models we keep the vision layers frozen and we use a learning rate of $\eta = 1e^{-4}$, dropout $\delta = 0.05$ and weight decay of 0.0002. For all other multimodal models we use a learning rate of $\eta = 1e^{-5}$, dropout $\delta = 0.05$ and weight decay of 0.00025.

6.8 Appendix B: Few-shot Prompting

For each dataset, we construct a prompt to include two randomly selected training examples for each class (GPT-3, FLAN-T5, IDEFICS) as follows:

- TIR (GPT-3 & FLAN-T5)

  Label the next text as 'image adds and text is represented', 'image adds and text is not represented', 'image does not add and text is represented', 'image does not add and text is not represented'. Text: <TWEET-TRAIN> // <LABEL-TRAIN> $\times 8$
  Label the next text as 'image adds and text is represented', 'image adds and text is not represented', 'image does not add and text is represented', 'image does not add and text is not represented'. Text: <TWEET> //

- TIR (IDEFICS)

  User: <IMAGE-TRAIN> Label the image as 'image adds and text is represented', 'image adds and text is not represented', 'image does not add and text is represented', 'image does not add and text is not represented'. Assistant:<LABEL-TRAIN> $\times 8$
  User: <IMAGE-TEST> Label the image as 'image adds and text is represented', 'image adds and text is not represented', 'image does not add and text is represented', 'image does not add and text is not represented'. Assistant:

- TIR (InstructBLIP)

  - Prompt: *Label the image as 'image adds and text is represented', 'image adds and text is not represented', 'image does not add and text is represented', 'image does not add and text is not represented'*

  - Image: *<IMAGE-TEST>*

- MVSA (GPT-3 & FLAN-T5)‘

  Label the next text as ‘positive’ or ‘negative’ or ‘neutral’. Text: <TWEET-TRAIN> // <LABEL-TRAIN> ×6

  Label the next text as ‘positive’ or ‘negative’ or ‘neutral’. Text: <TWEET> //

- MVSA (IDEFICS)

  User: <IMAGE-TRAIN> Is the sentiment of the image ‘positive’ or ‘negative’ or ‘neutral’?. Assistant:<LABEL-TRAIN> ×6

  User: <IMAGE-TEST> Is the sentiment of the image ‘positive’ or ‘negative’ or ‘neutral’?. Assistant:

- MVSA (InstructBLIP)

  - Prompt: *Is the sentiment of the image ‘positive’ or ‘negative’ or ‘neutral’?*
  - Image: *<IMAGE-TEST>*

- MHP

  Label the next text as ‘hateful’, ‘counterspeech’, ‘reclaimed’ or ‘none’. Text: <TWEET-TRAIN> // <LABEL-TRAIN> ×8

  Label the next text as ‘hateful’, ‘counterspeech’, ‘reclaimed’ or ‘none’. Text: <TWEET> //

- MHP (IDEFICS)

  User: <IMAGE-TRAIN> Is the image ‘hateful’, ‘counterspeech’, ‘reclaimed’ or ‘none’?. Assistant:<LABEL-TRAIN> ×8

  User: <IMAGE-TEST> Is the image ‘hateful’, ‘counterspeech’, ‘reclaimed’ or ‘none’?. Assistant:

- MHP (InstructBLIP)

  - Prompt: *Is the image ‘hateful’, ‘counterspeech’, ‘reclaimed’ or ‘none’?*
  - Image: *<IMAGE-TEST>*

- MSD (GPT-3 & FLAN-T5)

Label the next text as 'sarcastic' or 'not sarcastic'. Text: <TWEET-TRAIN> // <LABEL-TRAIN> ×4

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

- MSD (IDEFICS)

  ' User: <IMAGE-TRAIN> Is the image 'sarcastic' or 'not sarcastic'?. Assistant:<LABEL-TRAIN> ×4

  User: <IMAGE-TEST> Is the image 'sarcastic' or 'not sarcastic'?. Assistant:

- MSD (InstructBLIP)

  - Prompt: *Is the image 'sarcastic' or 'not sarcastic'?*
  - Image: *<IMAGE-TEST>*

- MICD (GPT-3 & FLAN-T5)

  Label the next text as 'commercial' or 'not commercial'. Text: <TWEET-TRAIN> // <LABEL-TRAIN> ×4

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

- MICD (IDEFICS)

  User: <IMAGE-TRAIN> Is the image 'commercial' or 'non-commercial'?. Assistant:<LABEL-TRAIN> ×4

  User: <IMAGE-TEST> Is the image 'commercial' or 'non-commercial'?. Assistant:

- MICD (InstructBLIP)

  - Prompt: *Is the image 'commercial' or 'non-commercial'?.*
  - Image: *<IMAGE-TEST>*

<Label-TRAIN> corresponds to the true label of the <TWEET-TRAIN> training example, <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.

6.8.1 Implementation Details

FLAN-T5 & IDEFICS We use one GPU T4 to obtain the inference results from Flan-T5 (Chung et al., 2022) and IDEFICS (Laurençon et al., 2023) models. For Flan-T5 we use the large version from the HuggingFace library (780M parameters) (Wolf et al., 2020). For IDEFICS, we use the 9B parameters instruct version of the model (*idefics-9b-instruct*) via Hugging Face library.

InstructBLIP We use one A100 GPU to obtain inference results from InstructBLIP (Dai et al., 2023). We use the 7B-parameters version (*instructblip-vicuna-7b*) from the HuggingFace library.

GPT-3 For GPT-3 (Brown et al., 2020), we use the *text-davinci-003* model via the OpenAI[3] Library ($82.61 USD total).

[3] https://platform.openai.com/docs/api-reference

Dataset	TIR		MVSA		MHP		MSD		MICD	
	λ_1, λ_2, λ_3	E	λ_1, λ_2, λ_3	E	λ_1, λ_2, λ_3	E	λ_1, λ_2, λ_3	E	λ_1, λ_2, λ_3	E
Ber-ViT-Conc	-	3	-	7	-	7	-	1	-	2
Ber-ViT-Conc+ITC	0.9, 0.1, 0	3	0.9, 0.1, 0	5	0.9, 0.1, 0	7	0.9, 0.1,0	6	0.9,0.1,0	2
Ber-ViT-Conc+ITM	0.9, 0, 0.1	4	0.9, 0, 0.1	6	0.9, 0, 0.1	9	0.9, 0, 0.1	3	0.9,0,0.1	1
Ber-ViT-Conc+ITC+ITM	0.8, 0.1, 0.1	6	0.8, 0.1, 0.1	4	0.8, 0.1, 0.1	6	0.8, 0.1, 0.1	3	0.8,0.1,0.1	2
Ber-ViT-Att	-	2	-	8	-	7	-	1	-	3
Ber-ViT-Att+ITC	0.9, 0.1,0	2	0.9, 0.1, 0	8	0.9,0.1,0	7	0.9, 0.1, 0	3	0.9,0.1,0	2
Ber-ViT-Att+ITM	0.92, 0, 0.08	3	0.9, 0, 0.1	6	0.9,0,0.1	6	0.9, 0, 0.1	3	0.9,0,0.1	1
Ber-ViT-Att+ITC+ITM	0.8, 0.1, 0.1	4	0.8, 0.1, 0.1	15	0.8,0.1,0.1	13	0.8, 0.1, 0.1	5	0.8,0.1,0.1	2
MMBT	-	2	-	9	-	5	-	1	-	1
MMBT+ITC	0.9, 0.1, 0	4	0.9, 0.1, 0	5	0.9, 0.1, 0	9	0.9,0.1,0	3	0.9,0.1,0	2
MMBT+ITM	0.9, 0, 0.1	4	0.7, 0 ,0.2	6	0.9, 0, 0.1	9	0.82, 0, 0.08	4	0.9,0,0.1	2
MMBT+ITC+ITM	0.84, 0.08, 0.08	3	0.85, 0.1, 0.05	11	0.8, 0.1, 0.1	10	0.85,0.1,0.05	3	0.6,0.2,0.2	4
LXMERT	-	2	-	5	-	5	-	2	-	3
LXMERT+ITC	0.9,0.1,0	2	0.9,0.1,0	8	0.9, 0.1, 0	5	0.9,0.1,0	2	0.9,0.1,0	2
LXMERT+ITM	0.85,0,0.05	1	0.9,0,0.1	6	0.8, 0, 0.1	12	0.85,0,0.05	2	0.9,0,0.1	3
LXMERT+ITC+ITM	0.9, 0.08, 0.02	2	0.83,0.02,0.15	7	0.8, 0.1, 0.1	11	0.85, 0.1, 0.05	2	0.8,0.1,0.1	3
ViLT	-	6	-	5	-	4	-	1	-	4
ViLT+ITC	0.9, 0.1, 0	6	0.9, 0.1, 0	11	0.9, 0.1, 0	4	0.9, 0.1, 0	1	0.95,0.05,0	2
ViLT+ITM	0.85, 0, 0.05	5	0.9,0,0.1	3	0.9, 0, 0.1	7	0.9, 0, 0.1	2	0.92,0,0.08	2
ViLT+ITC+ITM	0.8, 0.1, 0.1	2	0.8, 0.1, 0.1	13	0.8, 0.1, 0.1	9	0.8, 0.1, 0.1	2	0.87,0.05,0.08	1

Table 6.5: Hyperaprameter values for λ_1, λ_2, λ_3 as explained in Section 6.2, and number of fine-tuning epochs (E) for each model.

Dataset	POI		POLID		POLADV	
	λ_1, λ_2, λ_3	E	λ_1, λ_2, λ_3	E	λ_1, λ_2, λ_3	E
Ber-ViT-Att	-	4	-	4	-	2
Ber-ViT-Att+ITC	0.9, 0.1,0	5	0.9, 0.1, 0	4	0.9,0.1,0	3
Ber-ViT-Att+ITM	0.9, 0, 0.1	3	0.9, 0, 0.1	3	0.9,0,0.1	3
Ber-ViT-Att+ITC+ITM	0.8, 0.1, 0.1	3	0.8, 0.1, 0.1	5	0.8,0.1,0.1	4
ViLT	-	2	-	7	-	2
ViLT+ITC	0.9, 0.1, 0	1	0.9, 0.1, 0	4	0.9, 0.1, 0	7
ViLT+ITM	0.9, 0, 0.1	4	0.9,0,0.1	6	0.9, 0, 0.1	3
ViLT+ITC+ITM	0.8, 0.1, 0.1	3	0.8, 0.1, 0.1	3	0.8, 0.1, 0.1	3

Table 6.6: Hyperaprameter values for λ_1, λ_2, λ_3 as explained in Section 6.2, and number of fine-tuning epochs (E) for each model.

Chapter 7

Conclusions

This thesis presented novel work on three under-explored multimodal classification tasks: POI type prediction, online political advertisements analysis and influencer content analysis. Moreover, we proposed fine-tuning methods for tackling the challenges of modeling text and image content in social media. This chapter summarizes the tasks, findings and contributions, and suggests future research directions.

7.1 Summary of Thesis

Publication I: *Point-of-Interest Type Inference from Social Media Text* included in Chapter 2 presented the first study on predicting the POI type a social media message was posted from. We developed a large-scale dataset containing tweets mapped to their POI category, performed an analysis to find attributes specific to place type, then trained predictive models to infer the POI category using only tweet content and posting time.

Publication II: *Point-of-Interest Type Prediction using Text and Images* is found in Chapter 3. In this work, we enriched the dataset introduced in Publication I with images. Furthermore, we proposed a multimodal model that employs: (1) a gate mechanism to control the flow of information from each modality; and (2) a cross-attention mechanism to align and capture interactions between modalities. Our approach outperforms the text-only model in Publication I and competitive pre-trained multimodal models.

Publication III: *Analyzing Online Political Advertisements* corresponds to Chapter 4, where we presented the first study in NLP for analyzing the language of political ads. We introduced two new datasets containing political ads from the U.S. in English labeled by (1) the sponsor's ideology (*Conservative/Liberal*); and (2) the sponsor type (*Political Party/Third Party*). Both tasks were defined as advertisement-level binary classification and we evaluated a variety of approaches, including textual, visual and multimodal models. Our findings imply that text is a stronger modality for inferring the political ideology and sponsor type of a political advertisement compared to image-based features. However, the inclusion of visual information in the form of text descriptions or image-encoder features, improves the performance of the models.

Publication IV: *A Multimodal Analysis of Influencer Content on Twitter* is included in Chapter 5, where 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 findings demonstrate that our proposed cross-attention strategy for combining text and visuals, outperforms state-of-the-art multimodal models. Our new dataset will enable additional research into automatically detecting influencer advertising as well as studies in computational linguistics for large-scale analysis of commercial language characteristics.

Publlication V: *Improving Multimodal Classification of Social Media Posts by Leveraging Image-Text Auxiliary Tasks* is presented in Chapter 6. We proposed two auxiliary losses to be used when fine-tuning pre-trained multimodal models for social media classification. Image-Text Contrastive (ITC) encourages the model to capture the underlying dependencies in multimodal posts while Image-Text Matching (ITM) enables visual and language alignment. Our findings suggest that including these objectives improves prediction performance consistently. Moreover, our approach can be adapted to any multimodal architecture. This work contributes to advancing multimodal learning in the context of social media and provides insights for improving classification performance on text-image tasks.

7.2 Research Questions Discussion

In this section we discuss how we addressed the research questions proposed in Section 1.1.

Q1: What are the various methodologies available for extracting visual information from social media posts, and how can these methodologies be effectively used to enhance classification models? In Publications II, III, IV, and V, we delved into extracting features from pre-trained models like ResNet (He et al., 2016), leveraging convolution to aggregate information across locations. Additionally, we explored object-level features, isolating relevant objects and regions from images and merging them with text via the attention mechanism. While effective, this method requires an extra processing step and is influenced by the quality of the object detection model, such as Faster-RCNN (Ren et al., 2016). Publications IV and V further explore the use of features from pre-trained models using self-attention including ViT (Dosovitskiy et al., 2020). Finally, in Publication III, we investigated the use of text descriptions of image content. While promising, this method's performance is contingent on the quality of image descriptions, often introducing noise due to pre-training on diverse, sometimes less curated, data.

Q2: Can pre-trained multimodal models be directly applied to classify social media posts, or how can these models be adapted to account for the unique characteristics of social media posts? We explored various predictive models in Publications II to V. We examined *single-stream* models, such as VisualBERT and ViLT, as well as *dual-stream* approaches, like LXMERT, where images and text are processed separately. Furthermore, we introduced models specifically tailored to the characteristics of social media posts. MM-Gated-XATT, proposed in Publication II, effectively manages the intricate relationship between images and text through gated multimodal fusion and cross-attention. It addresses the challenge of incomplete data, such as missing images in posts, by incorporating an *average* image as a placeholder for such cases. In Publication IV, we introduced ViT-BERTweet-Att, which eliminates the need for multimodal gated fusion. This is achieved by leveraging image features from ViT, which employs self-attention. The model concatenates the multimodal representation with the original text content, recognizing the stronger signal in text. Notably, this model removes the requirement for a placeholder image, allowing it to seamlessly handle both text-only and image-text posts. Finally, in Publication V, we proposed leveraging two auxiliary tasks, Image-Text Contrastive (ITC) and Image-Text Matching (ITM), as

auxiliary losses during fine-tuning to enhance social media post classification. Our extensive study shows consistent performance improvement across five diverse multimodal social media datasets for models fine-tuned with ITC and ITM, covering both single- and dual-stream approaches. This approach addresses the challenges of hidden cross-modal semantics and weak image-text relationships in social media content.

Q3: To what extent does multimodal commercial content exist in social media beyond traditional forms of paid product advertising? Moreover, how transparent are these types of advertising to social media users? We analyze political advertising in Publication III and influencer commercial content in Publication IV. In Publication III, we analyze political advertising, an important aspect of digital election campaigning. Additionally, we introduce the task of sponsor type prediction, aiming to differentiate between ads from official political committees and those from third-party advertisers, crucial for transparency in elections. Third-party advertising has seen increased presence in previous elections, often funded by undisclosed dark-money sources. Next, in Publication IV, we delve into influencer marketing, where distinguishing between commercial and non-commercial content is challenging. Our focus is on automatically detecting commercial content in influencer posts to enhance transparency regarding promoted products, addressing concerns related to misleading advertising and undisclosed sponsorships at large scale.

7.3 Impact of Thesis Work

The tasks introduced in this thesis open new research directions, including inferring the location type from the content of a social media post, analyzing political advertising for sponsor type and ideology, and detecting the presence of commercial content in social media posts. These research directions hold particular relevance for computational social science, and offer valuable data and methods for social scientists to conduct large-scale studies.

In recent years, there has been a notable increase in interest in multimodal models for analyzing social media (Lu et al., 2018; Moon et al., 2018a; Xu and Li, 2022; Cheema et al., 2022; Yu et al., 2023). The datasets introduced in this thesis together with the proposed methods, provide diverse applications in future research. They can be employed for analyzing the characteristics of multimodal social media posts and advancing multimodal research more broadly. For instance, in Ilias et al. (2023), the authors base their model for detecting dementia

on our two-level multimodal model from Publication II and insights from Publication III. Moreover, our proposed datasets and methods can be used for refining research in each proposed task, and for contributing to the pre-training of large language models (Li et al., 2023; Laurençon et al., 2023). Our methods, enriched by insights into tailoring models to account for the characteristics of text-image relationships in social media posts, stand as valuable baseline models for upcoming research work. The next section suggests future research directions.

7.4 Future Directions

Publications I and II presented a study on POI type prediction. In the future, this work could be extended by focusing on inference at a more granular level of location types. Currently, the study categorized places into eight general types, but there is potential for more specific sub-categories. For example, instead of predicting just 'Food' the model could predict 'Italian restaurant', 'Chinese restaurant' or 'fast-food restaurant'. This would require more fine-grained data labeling. Additionally, it might require the implementation of hierarchical classification models to tackle this task. Moreover, to accommodate diverse aspects within a single tweet, one approach could involve redefining the problem as a multi-label classification scenario. For instance, enabling multiple labels per tweet, such as 'Food' and 'Shop & Service' as opposed to the current constraint of a single-label classification. Incorporating user and network information could also enhance the models' predictive capabilities. User behavior, preferences, and demographics might influence their interactions with different types of locations. Moreover, social network information could offer insights into the dynamics of POI engagement and popularity. Exploring ways to effectively incorporate this contextual information into the existing models could lead to more accurate predictions.

Publication III introduced work on online political ads consisting of images and text wordings from the image. Expanding the research on political ads to include other modalities such as speech and video would offer a more comprehensive analysis of the content used in political campaigns. Moreover, the extension of the work to different regions, languages, and platforms would allow for cross-cultural comparisons and insights into how political discourse varies across contexts. Analyzing political ads across multiple languages could present challenges in terms of natural language processing and translation, but it could lead to valuable insights into political communication on a global scale.

Publication IV presented a study of influencer content in Twitter (now known as 'X') using images and text in English. For modeling influencer content in multilingual settings across platforms, the research could expand its focus beyond Twitter and English. Different social media platforms may have distinct user behaviors and content formats, so adapting the influencer content analysis to platforms like Instagram, YouTube and TikTok in various languages, would provide a broader understanding of influencer marketing. Multilingual modeling may involve developing language-specific models or multilingual models capable of processing and understanding content in multiple languages.

Publication V introduced a novel approach to enhance pre-trained models by jointly fine-tuning them with two multimodal auxiliary tasks. This approach was evaluated on four popular multimodal tasks using Twitter posts in English. Assessing the models' performance across multiple platforms and languages will help generalize the findings and identify any platform or language-specific effects. Additional tasks could include emotion recognition, fake news detection, or named entity recognition, which could provide valuable insights into the dynamics of social media interactions.

Bibliography

Karuna Ahuja, Karan Sikka, Anirban Roy, and Ajay Divakaran. 2018. Understanding visual ads by aligning symbols and objects using co-attention. *arXiv preprint arXiv:1807.01448*.

Aisha Saadi Al-Subhi. 2022. Metadiscourse in online advertising: Exploring linguistic and visual metadiscourse in social media advertisements. *Journal of Pragmatics*, 187:24–40.

Najma Al Zydjaly. 2014. 8. geosemiotics: Discourses in place. In *Interactions, Images and Texts*, pages 63–76. De Gruyter Mouton.

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.

Ahmed N Alazzawi, Alia I Abdelmoty, and Christopher B Jones. 2012. What can I do there? Towards the automatic discovery of place-related services and activities. *International Journal of Geographical Information Science*, 26(2):345–364.

Nikolaos Aletras and Benjamin Paul Chamberlain. 2018. Predicting Twitter user socioeconomic attributes with network and language information. In *Proceedings of the 29th Conference on Hypertext and Social Media*, pages 20–24.

Malihe Alikhani, Sreyasi Nag Chowdhury, Gerard de Melo, and Matthew Stone. 2019. CITE: A corpus of image-text discourse relations. 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 570–575, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

Nick Anstead, João Carlos Magalhães, Richard Stupart, and Damian Tambini. 2018. Political advertising on facebook: The case of the 2017 united kingdom general election. In *American Political Science Association, Annual Meeting, Boston*.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Vanessa Apaolaza, Mario R. Paredes, Patrick Hartmann, and Clare D'Souza. 2021. How does restaurant's symbolic design affect photo-posting on instagram? the moderating role of community commitment and coolness. *Journal of Hospitality Marketing & Management*, 30(1):21–37.

John Arevalo, Thamar Solorio, Manuel Montes-y Gomez, and Fabio A González. 2020. Gated multimodal networks. *Neural Computing and Applications*, pages 1–20.

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational linguistics*, 34(4):555–596.

Gökhan Aydin. 2020. Social media engagement and organic post effectiveness: A roadmap for increasing the effectiveness of social media use in hospitality industry. *Journal of Hospitality Marketing & Management*, 29(1):1–21.

Chongyang Bai, Haipeng Chen, Srijan Kumar, Jure Leskovec, and VS Subrahmanian. 2021. M2p2: Multimodal persuasion prediction using adaptive fusion. *IEEE Transactions on Multimedia*.

Ramy Baly, Georgi Karadzhov, Abdelrhman Saleh, James Glass, and Preslav Nakov. 2019. Multi-task ordinal regression for jointly predicting the trustworthiness and the leading political ideology of news media. 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 2109–2116, Minneapolis, Minnesota. Association for Computational Linguistics.

Francesco Barbieri, Luis Espinosa Anke, and Jose Camacho-Collados. 2022. XLM-T: Multilingual language models in Twitter for sentiment analysis and beyond. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 258–266, Marseille, France. European Language Resources Association.

Bethan Benwell and Elizabeth Stokoe. 2006. *Discourse and identity*. Edinburgh University Press.

Sumit Bhatia and Deepak P. 2018. Topic-specific sentiment analysis can help identify political ideology. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 79–84, Brussels, Belgium. Association for Computational Linguistics.

Giscard Biamby, Grace Luo, Trevor Darrell, and Anna Rohrbach. 2022. Twitter-COMMs: Detecting climate, COVID, and military multimodal misinformation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1530–1549, Seattle, United States. Association for Computational Linguistics.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Steven Bird and Edward Loper. 2004. NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain. Association for Computational Linguistics.

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*.

Austin Botelho, Scott Hale, and Bertie Vidgen. 2021. Deciphering implicit hate: Evaluating automated detection algorithms for multimodal hate. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1896–1907, Online. Association for Computational Linguistics.

Ted Brader. 2005. Striking a responsive chord: How political ads motivate and persuade voters by appealing to emotions. *American Journal of Political Science*, 49(2):388–405.

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.

Christian Boris Brunner, Sebastian Ullrich, Patrik Jungen, and Franz-Rudolf Esch. 2016. Impact of symbolic product design on brand evaluations. *Journal of product & brand management*.

Yitao Cai, Huiyu Cai, and Xiaojun Wan. 2019. Multi-modal sarcasm detection in Twitter with hierarchical fusion model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2506–2515, Florence, Italy. Association for Computational Linguistics.

Rui Cao, Roy Ka-Wei Lee, Wen-Haw Chong, and Jing Jiang. 2022. Prompting for multimodal hateful meme classification. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 321–332, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Robyn Caplan and Danah Boyd. 2016. Who controls the public sphere in an era of algorithms. *Mediation, Automation, Power*, pages 1–19.

Nathanael Chambers, Victor Bowen, Ethan Genco, Xisen Tian, Eric Young, Ganesh Harihara, and Eugene Yang. 2015. Identifying political sentiment between nation states with social media. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 65–75, Lisbon, Portugal. Association for Computational Linguistics.

Gullal Singh Cheema, Sherzod Hakimov, Abdul Sittar, Eric Müller-Budack, Christian Otto, and Ralph Ewerth. 2022. MM-claims: A dataset for multimodal claim detection in social media. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 962–979, Seattle, United States. Association for Computational Linguistics.

Danni Chen, Kunwoo Park, and Jungseock Joo. 2020. Understanding gender stereotypes and electoral success from visual self-presentations of politicians in social media. In *Joint Workshop on Aesthetic and Technical Quality Assessment of Multimedia and Media Analytics for Societal Trends*, pages 21–25.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Learning universal image-text representations.

Zhongzhi Chen, Guang Liu, Bo-Wen Zhang, Fulong Ye, Qinghong Yang, and Ledell Wu. 2022. Altclip: Altering the language encoder in clip for extended language capabilities. *arXiv preprint arXiv:2211.06679*.

Zhiyuan Cheng, James Caverlee, and Kyumin Lee. 2010. You Are Where You Tweet: A Content-Based Approach to Geo-Locating Twitter Users. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM '10, pages 759–768.

Aleena Chia. 2012. Welcome to me-mart: The politics of user-generated content in personal blogs. *American Behavioral Scientist*, 56(4):421–438.

Dhivya Chinnappa, Srikala Murugan, and Eduardo Blanco. 2019. Extracting possessions from social media: Images complement language. 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 663–672, Hong Kong, China. Association for Computational Linguistics.

François Chollet. 2017. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258.

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.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning.

Giovanni De Gregorio and Catalina Goanta. 2020. The influencer republic: Monetizing political speech on social media. *Available at SSRN*.

Christine De Kock and Andreas Vlachos. 2022. Leveraging Wikipedia article evolution for promotional tone detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5601–5613, Dublin, Ireland. Association for Computational Linguistics.

Alexandra DeLucia, Shijie Wu, Aaron Mueller, Carlos Aguirre, Philip Resnik, and Mark Dredze. 2022. Bernice: A multilingual pre-trained encoder for Twitter. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6191–6205, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.

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.

Jacob Devlin, Hao Cheng, Hao Fang, Saurabh Gupta, Li Deng, Xiaodong He, Geoffrey Zweig, and Margaret Mitchell. 2015. Language models for image captioning: The quirks and what works. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 100–105, Beijing, China. Association for Computational Linguistics.

Arka Ujjal Dey, Suman K Ghosh, Ernest Valveny, and Gaurav Harit. 2021. Beyond visual semantics: Exploring the role of scene text in image understanding. *Pattern Recognition Letters*, 149:164–171.

Katharine Dommett and Luke Temple. 2018. Digital campaigning: The rise of facebook and satellite campaigns. *Parliamentary Affairs*, 71(suppl_1):189–202.

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*.

Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Pengchuan Zhang, Lu Yuan, Nanyun Peng, et al. 2022. An empirical study of training end-to-end vision-and-language transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18166–18176.

Mark Dredze, Miles Osborne, and Prabhanjan Kambadur. 2016. Geolocation for Twitter: Timing matters. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1064–1069, San Diego, California. Association for Computational Linguistics.

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.

Laura Edelson, Shikhar Sakhuja, Ratan Dey, and Damon McCoy. 2019. An analysis of united states online political advertising transparency. *arXiv preprint arXiv:1902.04385*.

Jacob Eisenstein, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. 2010. A latent variable model for geographic lexical variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1277–1287, Cambridge, MA. Association for Computational Linguistics.

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.

Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2017. Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612*.

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.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 363–370, Ann Arbor, Michigan. Association for Computational Linguistics.

Erika Franklin Fowler, Michael M Franz, Gregory J Martin, Zachary Peskowitz, and Travis N Ridout. 2020a. Political advertising online and offline. *American Political Science Review*, pages 1–20.

Erika Franklin Fowler, Michael M Franz, and Travis N Ridout. 2020b. The blue wave: Assessing political advertising trends and democratic advantages in 2018. *PS: Political Science & Politics*, 53(1):57–63.

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.

Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. 2016. Multimodal compact bilinear pooling for visual question answering and visual grounding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 457–468, Austin, Texas. Association for Computational Linguistics.

Gian M Fulgoni, Andrew Lipsman, and Carol Davidsen. 2016. The power of political advertising: Lessons for practitioners: How data analytics, social media, and creative strategies shape us presidential election campaigns. *Journal of Advertising Research*, 56(3):239–244.

Amit Gandhi, Daniela Iorio, and Carly Urban. 2016. Negative advertising and political competition. *The Journal of Law, Economics, and Organization*, 32(3):433–477.

Huiji Gao, Jiliang Tang, Xia Hu, and Huan Liu. 2015. Content-aware point of interest recommendation on location-based social networks. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI, pages 1721–1727.

Alexandra Georgakopoulou and Tereza Spilioti. 2015. *The Routledge handbook of language and digital communication*. Routledge.

Sean M Gerrish and David M Blei. 2011. Predicting legislative roll calls from text. In *Proceedings of the 28th International Conference on Machine Learning, ICML 2011*.

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.

Raul Gomez, Jaume Gibert, Lluis Gomez, and Dimosthenis Karatzas. 2020. Exploring hate speech detection in multimodal publications. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1470–1478.

Ellen Grigsby. 2008. *Analyzing Politics: An Introduction to Political Science*. Cengage Learning.

Jana Gross and Florian V Wangenheim. 2018. The big four of influencer marketing. a typology of influencers. *Marketing Review St. Gallen*, 2:30–38.

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206.

Bo Han, Paul Cook, and Timothy Baldwin. 2012. Geolocation prediction in social media data by finding location indicative words. In *Proceedings of COLING 2012*, pages 1045–1062, Mumbai, India. The COLING 2012 Organizing Committee.

Jinda Han, Qinglin Chen, Xilun Jin, Weikai Xu, Wanxian Yang, Suhansanu Kumar, Li Zhao, Hari Sundaram, and Ranjitha Kumar. 2021. Fitnet: Identifying fashion influencers on twitter. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW1).

Martin Haselmayer. 2019. Negative campaigning and its consequences: a review and a look ahead. *French Politics*, pages 1–18.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738.

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.

Arthur E Hoerl and Robert W Kennard. 1970. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

Christian Homburg, Martin Schwemmle, and Christina Kuehnl. 2015. New product design: Concept, measurement, and consequences. *Journal of marketing*, 79(3):41–56.

Eftekhar Hossain, Omar Sharif, and Mohammed Moshiul Hoque. 2022. MUTE: A multimodal dataset for detecting hateful memes. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 32–39, Online. Association for Computational Linguistics.

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.

X. Huang and A. Kovashka. 2016. Inferring visual persuasion via body language, setting, and deep features. In *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 778–784.

Pere-Lluís Huguet Cabot, Verna Dankers, David Abadi, Agneta Fischer, and Ekaterina Shutova. 2020. The Pragmatics behind Politics: Modelling Metaphor, Framing and Emotion in Political Discourse. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4479–4488, Online. Association for Computational Linguistics.

Z. Hussain, M. Zhang, X. Zhang, K. Ye, C. Thomas, Z. Agha, N. Ong, and A. Kovashka. 2017. Automatic understanding of image and video advertisements. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1100–1110.

Loukas Ilias, Dimitris Askounis, and John Psarras. 2023. Detecting dementia from speech and transcripts using transformers. *Computer Speech & Language*, 79:101485.

Shanto Iyengar and Markus Prior. 1999. Political advertising: what effect on commercial advertisers? *Retrieved from http://web.stanford.edu/s̃iyengar/research/papers/advertising.html*.

Mohit Iyyer, Peter Enns, Jordan Boyd-Graber, and Philip Resnik. 2014. Political ideology detection using recursive neural networks. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1113–1122, Baltimore, Maryland. Association for Computational Linguistics.

Yosra Jarrar, Ayodeji Olalekan Awobamise, and Adebola Adewunmi Aderibigbe. 2020. Effectiveness of influencer marketing vs social media sponsored advertising. *Utopia y Praxis Latinoamericana*, 25(12):40–54.

Akshay Java, Xiaodan Song, Tim Finin, and Belle Tseng. 2007. Why we Twitter: understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis*, pages 56–65.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR.

Mali Jin and Nikolaos Aletras. 2020. Complaint identification in social media with transformer networks. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1765–1771, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Mali Jin and Nikolaos Aletras. 2021. Modeling the severity of complaints in social media. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2264–2274, Online. Association for Computational Linguistics.

Bengt Johansson and Christina Holtz-Bacha. 2019. From analogue to digital negativity: Attacks and counterattacks, satire, and absurdism on election posters offline and online. In *Visual Political Communication*, pages 99–118. Springer.

J. Johnson, A. Karpathy, and L. Fei-Fei. 2016. Densecap: Fully convolutional localization networks for dense captioning. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4565–4574.

Kristen Johnson and Dan Goldwasser. 2018. Classification of moral foundations in microblog political discourse. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 720–730, Melbourne, Australia. Association for Computational Linguistics.

J. Joo, W. Li, F. F. Steen, and S. Zhu. 2014. Visual persuasion: Inferring communicative intents of images. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 216–223.

J. Joo, F. F. Steen, and S. Zhu. 2015. Automated facial trait judgment and election outcome prediction: Social dimensions of face. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 3712–3720.

Jungseock Joo and Zachary C Steinert-Threlkeld. 2018. Image as data: Automated visual content analysis for political science. *arXiv preprint arXiv:1810.01544*.

Xincheng Ju, Dong Zhang, Rong Xiao, Junhui Li, Shoushan Li, Min Zhang, and Guodong Zhou. 2021. Joint multi-modal aspect-sentiment analysis with auxiliary cross-modal relation detection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4395–4405, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Lynda L Kaid and Christina Holtz-Bacha. 2006. Television advertising and democratic systems around the world. *The Sage handbook of political advertising*, pages 445–457.

Lynda Lee Kaid and Monica Postelnicu. 2005. Political advertising in the 2004 election: Comparison of traditional television and internet messages. *American Behavioral Scientist*, 49(2):265–278.

Kanika Kalra, Bhargav Kurma, Silpa Vadakkeeveetil Sreelatha, Manasi Patwardhan, and Shirish Karande. 2020. Understanding advertisements with BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7542–7547, Online. Association for Computational Linguistics.

Yuhao Kang, Qingyuan Jia, Song Gao, Xiaohuan Zeng, Yueyao Wang, Stephan Angsuesser, Yu Liu, Xinyue Ye, and Teng Fei. 2019. Extracting human emotions at different places based on facial expressions and spatial clustering analysis. *Transactions in GIS*, 23(3):450–480.

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.

Kuralay Kenzhekankyzy Kenzhekanova. 2015. Linguistic features of political discourse. *Mediterranean Journal of Social Sciences*, 6(6 S2):192.

Sarthak Khanal, Maria Traskowsky, and Doina Caragea. 2022. Identification of fine-grained location mentions in crisis tweets. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7164–7173, Marseille, France. European Language Resources Association.

Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, and Davide Testuggine. 2019. Supervised multimodal bitransformers for classifying images and text. *arXiv preprint arXiv:1909.02950*.

Seungbae Kim, Xiusi 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*.

Efthymios Kouloumpis, Theresa Wilson, and Johanna D Moore. 2011. Twitter sentiment analysis: The good the bad and the omg! In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, ICWSM, pages 538–541.

Gunther R Kress, Theo Van Leeuwen, et al. 1996. *Reading images: The grammar of visual design*. Psychology Press.

Daniel Kriess and Bridget Barrett. 2020. Democratic tradeoffs: Platforms and political advertising. *Ohio State Technology Law Journal*, 16(2):494–518.

Klaus Krippendorff. 2018. *Content analysis: An introduction to its methodology*. Sage publications.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73.

Julia Kruk, Jonah Lubin, Karan Sikka, Xiao Lin, Dan Jurafsky, and Ajay Divakaran. 2019. Integrating text and image: Determining multimodal document intent in Instagram posts. 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 4622–4632, Hong Kong, China. Association for Computational Linguistics.

Anna Kruspe, Matthias Häberle, Iona Kuhn, and Xiao Xiang Zhu. 2020. Cross-language sentiment analysis of European Twitter messages during the COVID-19 pandemic. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, Online. Association for Computational Linguistics.

Vivek Kulkarni, Junting Ye, Steve Skiena, and William Yang Wang. 2018. Multi-view models for political ideology detection of news articles. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3518–3527, Brussels, Belgium. Association for Computational Linguistics.

Alok Kumar and Pramod Pathak. 2012. Political advertising in india: a perspective. *Management Insight*, 8(1):15–29.

Ankit Kumar, Piyush Makhija, and Anuj Gupta. 2020. Noisy text data: Achilles' heel of BERT. In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, pages 16–21, Online. Association for Computational Linguistics.

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.

Richard R Lau, Lee Sigelman, Caroline Heldman, and Paul Babbitt. 1999. The effects of negative political advertisements: A meta-analytic assessment. *American Political Science Review*, pages 851–875.

Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander M Rush, Douwe Kiela, et al. 2023. Obelisc: An open web-scale filtered dataset of interleaved image-text documents. *arXiv preprint arXiv:2306.16527*.

Yann LeCun, Yoshua Bengio, et al. 1995. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995.

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).

Dongxu Li, Junnan Li, Hongdong Li, Juan Carlos Niebles, and Steven CH Hoi. 2022a. Align and prompt: Video-and-language pre-training with entity prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4953–4963.

Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. 2020a. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11336–11344.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022b. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019a. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Quanzhi Li, Qiong Zhang, and Luo Si. 2019b. Rumor detection by exploiting user credibility information, attention and multi-task learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1173–1179, Florence, Italy. Association for Computational Linguistics.

Yang Li, Gang Li, Luheng He, Jingjie Zheng, Hong Li, and Zhiwei Guan. 2020b. Widget captioning: Generating natural language description for mobile user interface elements. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5495–5510, Online. Association for Computational Linguistics.

Yue Li and Carolina Scarton. 2020. Revisiting rumour stance classification: Dealing with imbalanced data. In *Proceedings of the 3rd International Workshop on Rumours and Deception in Social Media (RDSM)*, pages 38–44, Barcelona, Spain (Online). Association for Computational Linguistics.

Bin Liang, Chenwei Lou, Xiang Li, Min Yang, Lin Gui, Yulan He, Wenjie Pei, and Ruifeng Xu. 2022. Multi-modal sarcasm detection via cross-modal graph convolutional network. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1767–1777, Dublin, Ireland. Association for Computational Linguistics.

Wei-Hao Lin, Eric Xing, and Alexander Hauptmann. 2008. A joint topic and perspective model for ideological discourse. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 17–32. Springer.

Y. Lin, J. Hoover, G. Portillo-Wightman, C. Park, M. Dehghani, and H. Ji. 2018. Acquiring background knowledge to improve moral value prediction. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 552–559.

Bin Liu, Anmol Sheth, Udi Weinsberg, Jaideep Chandrashekar, and Ramesh Govindan. 2013. Adreveal: improving transparency into online targeted advertising. In *Proceedings of the Twelfth ACM Workshop on Hot Topics in Networks*, pages 1–7.

Haibin Liu, Bo Luo, and Dongwon Lee. 2012. Location type classification using tweet content. In *2012 11th International Conference on Machine Learning and Applications*, volume 1, pages 232–237. IEEE.

Kang Liu, Ling Yin, Feng Lu, and Naixia Mou. 2020a. Visualizing and exploring poi configurations of urban regions on poi-type semantic space. *Cities*, 99:102610.

Tongcun Liu, Jianxin Liao, Zhigen Wu, Yulong Wang, and Jingyu Wang. 2020b. Exploiting geographical-temporal awareness attention for next point-of-interest recommendation. *Neurocomputing*, 400:227–237.

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.

Di Lu, Leonardo Neves, Vitor Carvalho, Ning Zhang, and Heng Ji. 2018. Visual attention model for name tagging in multimodal social media. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1990–1999, Melbourne, Australia. Association for Computational Linguistics.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019a. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019b. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32.

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.

Jing Ma, Wei Gao, and Kam-Fai Wong. 2018. Rumor detection on Twitter with tree-structured recursive neural networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1980–1989, Melbourne, Australia. Association for Computational Linguistics.

Eleonora Mancini, Federico Ruggeri, Andrea Galassi, and Paolo Torroni. 2022. Multimodal argument mining: A case study in political debates. In *Proceedings of the 9th Workshop on Argument Mining*, pages 158–170, Online and in Gyeongju, Republic of Korea. International Conference on Computational Linguistics.

Antonios Maronikolakis, Danae Sánchez Villegas, Daniel Preotiuc-Pietro, and Nikolaos Aletras. 2020a. Analyzing political parody in social media. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4373–4384.

Antonis Maronikolakis, Danae Sánchez Villegas, Daniel Preotiuc-Pietro, and Nikolaos Aletras. 2020b. Analyzing political parody in social media. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4373–4384, 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.

Grant McKenzie, Krzysztof Janowicz, Song Gao, Jiue-An Yang, and Yingjie Hu. 2015. POI pulse: A multi-granular, semantic signature–based information observatory for the interactive visualization of big geosocial data. *Cartographica: The International Journal for Geographic Information and Geovisualization*, 50(2):71–85.

Matthew K McKitrick, Nadine Schuurman, and Valorie A Crooks. 2023. Collecting, analyzing, and visualizing location-based social media data: review of methods in gis-social media analysis. *GeoJournal*, 88(1):1035–1057.

Juan Carlos Medina Serrano, Orestis Papakyriakopoulos, and Simon Hegelich. 2020. Exploring political ad libraries for online advertising transparency: Lessons from germany and the 2019 european elections. In *International Conference on Social Media and Society*, pages 111–121.

Julia Mendelsohn, Ceren Budak, and David Jurgens. 2021. Modeling framing in immigration discourse on social media. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2219–2263, Online. Association for Computational Linguistics.

Piyush Mishra. 2020. Geolocation of tweets with a BiLSTM regression model. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 283–289, Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).

Seungwhan Moon, Leonardo Neves, and Vitor Carvalho. 2018a. Multimodal named entity disambiguation for noisy social media posts. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2000–2008, Melbourne, Australia. Association for Computational Linguistics.

Seungwhan Moon, Leonardo Neves, and Vitor Carvalho. 2018b. Multimodal named entity recognition for short social media posts. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language*

Technologies, Volume 1 (Long Papers), pages 852–860, New Orleans, Louisiana. Association for Computational Linguistics.

Preslav Nakov, Alan Ritter, Sara Rosenthal, Fabrizio Sebastiani, and Veselin Stoyanov. 2016. SemEval-2016 task 4: Sentiment analysis in Twitter. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1–18, San Diego, California. Association for Computational Linguistics.

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. BERTweet: 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.

Thien Hai Nguyen and Kiyoaki Shirai. 2015. Topic modeling based sentiment analysis on social media for stock market prediction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1354–1364, Beijing, China. Association for Computational Linguistics.

Teng Niu, Shiai Zhu, Lei Pang, and Abdulmotaleb El Saddik. 2016. Sentiment analysis on multi-view social data. In *MultiMedia Modeling: 22nd International Conference, MMM 2016, Miami, FL, USA, January 4-6, 2016, Proceedings, Part II 22*, pages 15–27. Springer.

Nicolas Ocampo, Elena Cabrio, and Serena Villata. 2023. Playing the part of the sharp bully: Generating adversarial examples for implicit hate speech detection. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2758–2772, Toronto, Canada. 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.

Shraman Pramanick, Dimitar Dimitrov, Rituparna Mukherjee, Shivam Sharma, Md. Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021. Detecting harmful memes and their targets. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2783–2796, Online. Association for Computational Linguistics.

Daniel Preoţiuc-Pietro and Trevor Cohn. 2013. Mining user behaviours: a study of check-in patterns in location based social networks. In *Proceedings of the 5th annual ACM Web Science Conference*, Web Science, pages 306–315.

Daniel Preoţiuc-Pietro, Mihaela Gaman, and Nikolaos Aletras. 2019a. Automatically identifying complaints in social media. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5019.

Daniel Preoţiuc-Pietro, Mihaela Gaman, and Nikolaos Aletras. 2019b. Automatically identifying complaints in social media. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5019, Florence, Italy. Association for Computational Linguistics.

Daniel Preoţiuc-Pietro, Ye Liu, Daniel Hopkins, and Lyle Ungar. 2017. Beyond binary labels: Political ideology prediction of Twitter users. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 729–740, Vancouver, Canada. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Afshin Rahimi, Duy Vu, Trevor Cohn, and Timothy Baldwin. 2015. Exploiting text and network context for geolocation of social media users. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1362–1367, Denver, Colorado. Association for Computational Linguistics.

Arijit Ray, Karan Sikka, Ajay Divakaran, Stefan Lee, and Giedrius Burachas. 2019. Sunny and dark outside?! improving answer consistency in VQA through entailed question generation. 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 5860–5865, Hong Kong, China. 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.

Kevin Reschke and Pranav Anand. 2012. POLITICAL-ADS: An annotated corpus for modeling event-level evaluativity. In *Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis*, pages 84–88, Jeju, Korea. Association for Computational Linguistics.

Travis N Ridout, Erika Franklin Fowler, and John Branstetter. 2010. Political advertising in the 21st century: The rise of the youtube ad. In *APSA 2010 Annual Meeting Paper*.

Stephen Roller, Michael Speriosu, Sarat Rallapalli, Benjamin Wing, and Jason Baldridge. 2012. Supervised text-based geolocation using language models on an adaptive grid. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1500–1510, Jeju Island, Korea. Association for Computational Linguistics.

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.

Roney Santos, Gabriela Pedro, Sidney Leal, Oto Vale, Thiago Pardo, Kalina Bontcheva, and Carolina Scarton. 2020. Measuring the impact of readability features in fake news

detection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1404–1413, Marseille, France. European Language Resources Association.

Crispin Sartwell. 2011. *Political aesthetics*. Cornell University Press.

Margaret Scammell and Ana I Langer. 2006. Political advertising: why is it so boring? *Media, culture & society*, 28(5):763–784.

H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Stephanie M Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin EP Seligman, et al. 2013. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8(9):e73791.

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.

Ron Scollon and Suzie Wong Scollon. 2003. *Discourses in place: Language in the material world*. Routledge.

Qingyi Si, Fandong Meng, Mingyu Zheng, Zheng Lin, Yuanxin Liu, Peng Fu, Yanan Cao, Weiping Wang, and Jie Zhou. 2022. Language prior is not the only shortcut: A benchmark for shortcut learning in VQA. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3698–3712, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yanchuan Sim, Brice D. L. Acree, Justin H. Gross, and Noah A. Smith. 2013. Measuring ideological proportions in political speeches. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 91–101, Seattle, Washington, USA. Association for Computational Linguistics.

Stergios Skaperdas and Bernard Grofman. 1995. Modeling negative campaigning. *American Political Science Review*, 89(1):49–61.

Pavel Skorupa and Tatjana Dubovičienė. 2015. Linguistic characteristics of commercial and social advertising slogans. *Coactivity: Philology, Educology/Santalka: Filologija, Edukologija*, 23(2):108–118.

Till Speicher, Muhammad Ali, Giridhari Venkatadri, Filipe Ribeiro, George Arvanitakis, Fabrício Benevenuto, Krishna Gummadi, Patrick Loiseau, and Alan Mislove. 2018. Potential for discrimination in online targeted advertising. In *FAT 2018-Conference on Fairness, Accountability, and Transparency*, volume 81, pages 1–15.

Peter Stefanov, Kareem Darwish, Atanas Atanasov, and Preslav Nakov. 2020. Predicting the topical stance and political leaning of media using tweets. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 527–537, Online. Association for Computational Linguistics.

Jennifer Stromer-Galley, Patrícia Rossini, Jeff Hemsley, Sarah E Bolden, and Brian McKernan. 2021. Political messaging over time: A comparison of us presidential candidate facebook posts and tweets in 2016 and 2020. *Social Media+ Society*, 7(4):20563051211063465.

Lin Sun, Jiquan Wang, Yindu Su, Fangsheng Weng, Yuxuan Sun, Zengwei Zheng, and Yuanyi Chen. 2020. RIVA: A pre-trained tweet multimodal model based on text-image relation for multimodal NER. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1852–1862, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Danae Sánchez Villegas, Catalina Goanta, and Nikolaos Aletras. 2023. A multimodal analysis of influencer content on twitter. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 225–240, Nusa Dua, Bali. 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.

Hao Tan, Licheng Yu, and Mohit Bansal. 2019. Learning to navigate unseen environments: Back translation with environmental dropout. 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 2610–2621, Minneapolis, Minnesota. Association for Computational Linguistics.

Mingxing Tan and Quoc Le. 2019. EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6105–6114. PMLR.

Vlad Tanasescu, Christopher B Jones, Gualtiero Colombo, Martin J Chorley, Stuart M Allen, and Roger M Whitaker. 2013. The personality of venues: Places and the five-factors ('big five') model of personality. In *Fourth IEEE International Conference on Computing for Geospatial Research and Application*, pages 76–81.

Mike Thelwall. 2021. Lifestyle information from youtube influencers: some consumption patterns. *Journal of Documentation*.

Christopher Thomas and Adriana Kovashka. 2018. Persuasive faces: Generating faces in advertisements. *arXiv preprint arXiv:1807.09882*.

James Thorne and Andreas Vlachos. 2018. Automated fact checking: Task formulations, methods and future directions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3346–3359, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Lin Tian, Xiuzhen Zhang, and Jey Han Lau. 2022. DUCK: Rumour detection on social media by modelling user and comment propagation networks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4939–4949, Seattle, United States. Association for Computational Linguistics.

Yuan Tian, Nan Xu, Ruike Zhang, and Wenji Mao. 2023a. Dynamic routing transformer network for multimodal sarcasm detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2468–2480, Toronto, Canada. Association for Computational Linguistics.

Yuanhe Tian, Weidong Chen, Bo Hu, Yan Song, and Fei Xia. 2023b. End-to-end aspect-based sentiment analysis with Combinatory Categorial Grammar. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13597–13609, Toronto, Canada. Association for Computational Linguistics.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Annual Meeting of the Association for*

Computational Linguistics, pages 6558–6569, Florence, Italy. Association for Computational Linguistics.

Adam Tsakalidis, Nikolaos Aletras, Alexandra I Cristea, and Maria Liakata. 2018. Nowcasting the stance of social media users in a sudden vote: The case of the Greek Referendum. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 367–376.

Yi-Fu Tuan. 1977. *Space and place: The perspective of experience*. U of Minnesota Press.

Yi-Fu Tuan. 1991. Language and the making of place: A narrative-descriptive approach. *Annals of the Association of American geographers*, 81(4):684–696.

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 Preoţiuc-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.

Jianfeng Wang, Xiaowei Hu, Zhe Gan, Zhengyuan Yang, Xiyang Dai, Zicheng Liu, Yumao Lu, and Lijuan Wang. 2021. Ufo: A unified transformer for vision-language representation learning. *arXiv preprint arXiv:2111.10023*.

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.

Demi van Weerdenburg, Simon Scheider, Benjamin Adams, Bas Spierings, and Egbert van der Zee. 2019. Where to go and what to do: Extracting leisure activity potentials from web data on urban space. *Computers, Environment and Urban Systems*, 73:143–156.

Darrell M West. 2017. *Air wars: television advertising and social media in election campaigns, 1952-2016*. CQ Press.

Bartosz W Wojdynski. 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 Wojdynski, 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, and Jamie Brew. 2019a. Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019b. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Nan Xi, Di Ma, Marcus Liou, Zachary C Steinert-Threlkeld, Jason Anastasopoulos, and Jungseock Joo. 2020. Understanding the political ideology of legislators from social media images. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 726–737.

Q. Xie, M. T. Luong, E. Hovy, and Q. V. Le. 2020. Self-training with noisy student improves imagenet classification. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695.

Chunpu Xu and Jing Li. 2022. Borrowing human senses: Comment-aware self-training for social media multimodal classification. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5644–5656, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Chunpu Xu, Hanzhuo Tan, Jing Li, and Piji Li. 2022. Understanding social media cross-modality discourse in linguistic space. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2459–2471, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Nan Xu, Zhixiong Zeng, and Wenji Mao. 2020. Reasoning with multimodal sarcastic tweets via modeling cross-modality contrast and semantic association. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3777–3786, Online. Association for Computational Linguistics.

Hyeongjun Yang, Heesoo Won, Youbin Ahn, and Kyong-Ho Lee. 2023. CLICK: Contrastive learning for injecting contextual knowledge to conversational recommender system. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1875–1885, Dubrovnik, Croatia. Association for Computational Linguistics.

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.

Ussama Yaqub, Nitesh Sharma, Rachit Pabreja, Soon Ae Chun, Vijayalakshmi Atluri, and Jaideep Vaidya. 2020. Location-based sentiment analyses and visualization of twitter election data. *Digital Government: Research and Practice*, 1(2):1–19.

Keren Ye, Kyle Buettner, and Adriana Kovashka. 2018. Story understanding in video advertisements. *arXiv preprint arXiv:1807.11122*.

Keren Ye and Adriana Kovashka. 2018. Advise: Symbolism and external knowledge for decoding advertisements. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 837–855.

Mao Ye, Dong Shou, Wang-Chien Lee, Peifeng Yin, and Krzysztof Janowicz. 2011. On the semantic annotation of places in location-based social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 520–528.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. 2022. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*.

Jianfei Yu, Ziyan Li, Jieming Wang, and Rui Xia. 2023. Grounded multimodal named entity recognition on social media. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9141–9154, Toronto, Canada. Association for Computational Linguistics.

Quan Yuan, Gao Cong, Zongyang Ma, Aixin Sun, and Nadia Magnenat Thalmann. 2013. Time-aware point-of-interest recommendation. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 363–372.

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.

Xingshan Zeng, Jing Li, Lu Wang, Zhiming Mao, and Kam-Fai Wong. 2020. Dynamic online conversation recommendation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3331–3341, Online. Association for Computational Linguistics.

Fan Zhang, Jinyan Zu, Mingyuan Hu, Di Zhu, Yuhao Kang, Song Gao, Yi Zhang, and Zhou Huang. 2020. Uncovering inconspicuous places using social media check-ins and street view images. *Computers, Environment and Urban Systems*, 81:101478.

Mingda Zhang, Rebecca Hwa, and Adriana Kovashka. 2018. Equal but not the same: Understanding the implicit relationship between persuasive images and text. *arXiv preprint arXiv:1807.08205*.

Siyuan Zhang and Hong Cheng. 2018. Exploiting context graph attention for poi recommendation in location-based social networks. In *International Conference on Database Systems for Advanced Applications*, pages 83–99. Springer.

Ye Zhi, Haifeng Li, Dashan Wang, Min Deng, Shaowen Wang, Jing Gao, Zhengyu Duan, and Yu Liu. 2016. Latent spatio-temporal activity structures: a new approach to inferring intra-urban functional regions via social media check-in data. *Geo-spatial Information Science*, 19(2):94–105.

Chaoran Zhou, Hang Yang, Jianping Zhao, and Xin Zhang. 2020a. Poi classification method based on feature extension and deep learning. *Journal of Advanced Computational Intelligence and Intelligent Informatics*, 24(7):944–952.

Kaimin Zhou, Chang Shu, Binyang Li, and Jey Han Lau. 2019. Early rumour detection. 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 1614–1623, Minneapolis, Minnesota. Association for Computational Linguistics.

Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. 2020b. Unified Vision-Language Pre-Training for Image Captioning and VQA. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):13041–13049.

Jiang Zhu, Lan Jiang, Wenyu Dou, and Liang Liang. 2019. Post, eat, change: the effects of posting food photos on consumers' dining experiences and brand evaluation. *Journal of Interactive Marketing*, 46:101–112.