文本意圖的多模態分析：以 Instagram 為例

An Analysis of Multimodal Document Intent in Instagram Posts

陳盈瑜 Ying-Yu Chen
國立臺灣大學語言學研究所
Graduate Institute of Linguistics
National Taiwan University
r06142009@ntu.edu.tw

謝舒凱 Shu-Kai Hsieh
國立臺灣大學語言學研究所
Graduate Institute of Linguistics
National Taiwan University
shukaihsieh@ntu.edu.tw

摘要

時至今日，社群媒體(如 Instagram)趨向結合圖片以及文字表徵，建構出一種新的「多模態」溝通方式。利用計算方法分析多模態關係已成為一個熱門的主題，然而，尚未有研究針對台灣的百大網紅發文中的多模態圖文配對(Image-caption Pair)來分析文本意圖和圖文關係。利用文字和圖片的多模態表徵，本研究沿用 Kruk et al. (2019)的圖文關係分類方法(contextual relationship/semiotic relationship/author’s intent)，對此三種分類提出新的圖文表徵方式(Sentence-BERT 及 image embedding)，並利用計算模型(Random Forest, Decision Tree Classifier)精準分類三種圖文關係，研究結果顯示正確率高達86.23%。

Abstract

Present-day, a majority of representation style on social media (i.e., Instagram) tends to combine visual and textual content in the same message as a consequence of building up a modern way of communication. Message in multimodality is essential in almost any type of social interaction especially in the context of social multimedia content online. Hence, effective computational approaches for understanding documents with multiple modalities are needed to identify the relationship between them. This study extends recent advances in authors intent classification by putting forward an approach using Image-caption Pairs

(ICPs). Several Machine Learning algorithm like Decision Tree Classifier (DTC's), Random Forest (RF) and encoders like Sentence-BERT and picture embedding are undertaken in the tasks in order to classify the relationships between multiple modalities, which are 1) contextual relationship 2) semiotic relationship and 3) authors intent. This study points to two possible results. First, despite the prior studies consider incorporating the two synergistic modalities in a combined model will improve the accuracy in the relationship classification task, this study found out the simple fusion strategy that linearly projects encoded vectors from both modalities in the same embedding space may not strongly enhance the performance of that in a single modality. The results suggest that the incorporating of text and image needs more effort to complement each other. Second, we show that these text-image relationships can be classified with high accuracy (86.23%) by using only text modality. In sum, this study may be essential in demonstrating a computational approach to access multimodal documents as well as providing a better understanding of classifying the relationships between modalities.

關鍵詞：多模態文本分析，自然語言處理，決策樹，隨機森林

Keywords: multimodal documents understanding, contextual relationship, semiotic relationship, authors intent, Natural Language Processing, Decision Tree Classifier, Random Forest, Sentence-BERT, image embedding

1. Introduction

Up to date, a majority of representation style on social media tends to combine visual and textual content in the same message, the growing of multimodal documents is thus at a staggering rate. The developing multimodal document builds up a modern way of communication. The fact is that social multimedia, such as Instagram, inevitably hosting the way for information conveying. When reading a post on a multimodal social media, a simple question is raised: how do people analyze the relationship between image and text? Regarding the meaning of a word, we found out that it would help to know the difference between denotation and connotation, which can be something suggested or implied by a word or constructions of words. As in semiotics, the terms denotation and connotation may be seen as different ways comprising a particular semantic domain, making up of the two obligatory relata of the sign function - (1) the expression and content, and (2) of a portion of the world in correspondence of the content, that is, the referent (Wason & Jones, 1963).

To access multimodal data in a computational way, an essential inclination of current studies utilized an amount of data with annotation for solving different scales of computer science problems in data mining and multimedia (Jin et al., 2010). Besides, a substantial body of related research documents the tendency of assuming images accompanied by basic text labels or captions such as interpreting author intent (Kruk et al., 2019). With the growing volume of multimodal social media, the detection of the relationships between text and image has become more critical. Research engaging fully with the multimodal aspects of the texts and images on social media is still in its infancy. Some researchers (i.e., Castro et al. (2019), Kruk et al. (2019), O'Halloran et al. (2019), Phan et al. (2019)) considered image and text as both the primary content, viewing that incorporating multimodal features can result in Meaning Multiplication, hence improving the automatic classification. Unfortunately, there is little general consensus on how does Meaning Multiplication (Lemke, 1998) comes with online multimodal documents, and few empirical studies have been done on this issue. In addition, more people consider that ignoring images means ignoring a large portion of potential meaning (Summaries & Panel, n.d.). Castro et al. (2019) further indicated that multimodal information can reduce the relative error rate compared to the use of individual modalities, thus motivating the purposes in this study - to clarify whether there is Meaning Multiplication in the multimodal documents and to distinguish the alignment between the image and text modalities.

As a consequence, two major sets of research questions are addressed in this study. The first purpose is to describe whether the combined modality “stronger” - means which has better performance in the multimodal classification task - than individual modalities or not, and in what way and how? This study uses the text-only feature, image-only feature, and combined model from both modalities of Image-caption Pairs (abbreviated as ICPs) to develop an automatic classification system, aiming to classify the presence of the *contextual relationship* between the literal meanings (referred to what is said) of the image and caption, the *semiotic relationship* between what is signified by the image modality and text modality, and the *author’s intent* hiding behind the ICPs. The second purpose is to access the relationship between text and image, and their combination given Meaning Multiplication of multimodal documents. Given the complex nature of multimodal documents, many researchers are motivated to develop a framework to classify different relationships in the multimodal documents (Chancellor et al., 2017; Illendula & Sheth, 2019; Kruk et al., 2019; Zeppelzauer & Schopfhauser, 2016). Among these recent advances, Kruk et al. (2019) proposed a framework for a sounder theoretical bases to solve this problem. In this study, we adapt the taxonomy designed by Kruk by modifying existing taxonomies (Bateman, 2014; Marsh & White, 2003) to explore the relationship of the posts of Key Opinion Leaders (KOLs) in Instagram, and establish a new dataset with 936 posts from a variety of KOLs. We further put forward a multimodal approach that can classify the relationships between multiple modalities.

2. Literature Review

2.1 Taxonomies

Since this study focuses on exploring the ways that images and text interact, we adapt three taxonomies extracted by Kruk et al. (2019), which are possible to identify their relationships applicable to all subject areas and document types according to the closeness of the conceptual relationship. In the three proposed taxonomies introduced by Kruk et al. (2019), two (contextual and semiotic) are taken advantages to capture different aspects of the relationship between image and caption, and one to capture speaker intent (Kruk et al., 2019) while the three taxonomies address the underlying concept of Marsh and White (2003) – framing the image only as subordinate to the text. A survey was conducted to identify those three relationships in the Instagram contexts.

First, it is investigated that the **contextual taxonomy** claimed to report the relationship between the literal meaning of the image and text (Kruk et al., 2019; Marsh & White, 2003). Kruk et al. (2019) considered three categories of Marsh and White (2003) taxonomy which generally captured the closeness of the relationship between image and text essential; hence, they further generalized them to three top-level categories to make them symmetric for the Instagram domain: *minimal*, *close*, and *transcendent*. Second, to answer questions concerning the more complex forms of meaning multiplication, capturing the relationship between what is signified by the respective modalities turned out to be the priority of the **semiotic relationship**. Kruk et al. (2019) categorized the semiotic relationship between ICPs as *divergent*, *parallel*, and *additive* by taking advantage of the earlier 3-way distinction (Bateman, 2014; Kloepfer, 1976) and the two-way (parallel vs. non-parallel) classification (Zhang et al., 2018). Third, with the advantages that prior work has drawn on **author's intent**, like Goffman's proposal of self-presentation (Goffman et al., 1978; Mahoney et al., 2016), eight illocutionary intents had been developed (Kruk et al., 2019). Details of the eight intents are quoted below:

Advocative: advocate for a figure, idea, movement, etc.
Promotive: promote events, products, organizations, etc.
Exhibitionist: create a self-image for the user using selfies, pictures of belongings, etc.
Expressive: express emotion, attachment, or admiration at an external entity or group.
Informative: relay information regarding a subject or event using factual language.
Entertainment: entertain using art, humor, memes, etc.
Provocative/Discrimination: directly attach an individual or group.
Provocative/Controversial: be shocking.

2.2 Multimodal Document Understanding

The literature is full of discussions surrounding the definitions of modality, and scholars have debated its nature for decades. “A modality is a communication channel, for instance, related to the human senses or the form of expression (Bongers & van der Veer, 2007)”. Modality, which most people associate sensory modalities with, is to represent our primary channels of communication and sensation, such as vision or touch (Baltrušaitis et al., 2018). The objects we see, the sounds we hear, the odors we smell are different modalities that we received in the surrounding world. For example, text and images are sometimes considered from a different modality, that is, different “modes” of communication. Text-image relations and their related work hence fall within the general area of multimodality - the investigation of diverse modes of expressions and their combinations. Additionally, the research regarding multiple modalities of document understanding is hence called multimodality document understanding.

As discussed in the last paragraph, such a combination of diverse modalities sometimes results in Meaning Multiplication, a metaphor first promoted by the socio-functional semiotician Jay Lemke (Lemke, 1998). Prior work (Bateman, 2014) states that “under the right condition, the value of a combination of different modes of meaning can be worth more than the information (whatever that might be) that we get from the modes when used alone. In other words, text ‘multiplied by’ images is more than text simply occurring with or alongside images. (...) Somehow the meanings of one and the meanings of the other resonate so as to produce more than the sum of the parts”. As for NLP perspective, Morency and

Baltrušaitis (2017) provided the view that, with the goals of recognizing language and vision projects such as image and video captioning, Multimodal Machine Learning is, inevitably, a vibrant multi-disciplinary research field for Artificial Intelligence, for example, integrating and modeling multiple communicative modalities, like linguistics, acoustic, and visual messages. In this study, we focus primarily on two modalities: the natural language that can be written, and visual signal which is represented with images. So far, seminal work on defining modality was carried out.

In order to interpret and reason about multimodal messages, it is necessary to develop a computational model that can not only deal with the heterogeneity of the data and the contingency often found between modalities but understand the dependencies across modalities. Additionally, in this study, it also requires the knowledge of the multimodal language, and thus, a multimodal framework was created and briefly introduced below. Most researchers working on exploring the relationship between text and image and extracting meaning often assigning a subordinate role to either text or images, i.e., image captioning, visual question answering, which claimed that text is a subordinate modality to image. Thus, our work builds on the framework of Marsh and White (2003) who offers a taxonomy of the relationship between image and text which Kruk et al. (2019) draw on to create a new one. With respect to this, Baltrušaitis et al. (2018) brought out five unique challenges regarding the research field of Multimodal Machine Learning as follows.

Representation: The first challenge state how to represent and summarize multimodal data to highlight the complementarity and synchrony between modalities. To construct multimodal data representations, the researcher may face several difficulties like combining the data from heterogeneous origins and dealing with different levels of data noise and missing data. In place of author intent on Instagram, almost all researchers explore classification tasks on Instagram by taking advantage of word embedding by Word2vec (Le & Mikolov, 2014). With reference to image representation, some Singla et al. (2018) exploit ResNet50 (He et al., 2016) to convert the images into vectors based on the different research purpose and essence. However, word-embedding seemed to be out of state now. With the target of modeling intra-modality dynamics, Yu and Jiang (2019) first apply Bidirectional Encoder Representations from Trans- formers (BERT) (Devlin et al., 2018) to get target-sensitive textual representations. In the context of multimodal language understanding, a majority of multimodal research (Rahman et al., 2019) find this model outperforms several highly competitive approaches.

Translation: The second difficulty announces how to translate, or say, map, data from two or more modalities. On one hand, the data is disparate due to the way of representations between modalities. On the other hand, the relationship between modalities is often open-ended or subjective. For instance, although there are several ways of translation to explain an image, there may not be one. In addition, the evaluation and characterization of the multimodal translation may be subjective.

Alignment: The third obstacle indicates how to confirm the direct relationships between (sub)elements of instances from one modality to another. For example, given an image and a caption, the mission is to align which part of an image could be corresponded to the caption's representation. With reference to combining information from visual image and text, incorporating two synergistic modalities in a combined model is a high-efficient way adapted by most researchers, no matter in ICPs (Kruk et al., 2019), emoji-text pair (Barbieri et al., 2018), or feature-extraction for multimodal sentiment analysis (Soleymani et al., 2017). These studies, in fact, most studies, employ a competitive architecture in many image classification tasks, Convolutional Neural Networks (CNN) (Baltrušaitis et al., 2018; Lin et al., 2014; Russakovsky et al., 2015) or Random Forest (Breiman, 2001). What's more, Residual Networks (abbreviated as ResNets afterward) (He et al., 2016) is involved with CNN showed to be one of the best CNN models for image recognition. However, existing approaches to this task primarily rely on the textual content, but ignoring the other increasingly vibrant multimodal data sources, like images. This kind of ignoring will somehow enhance the robustness of these text-based models. Inspired by the recently proposed BERT architecture, a multimodal BERT architecture is applied, firstly by (Yu & Jiang, 2019), to obtain target-sensitive textual representations in order to model intra-modality dynamics.

Fusion: The fourth face-off remarks on how to combine information from two or more different modalities to perform a prediction, discrete or continuous. Take an image for example, the visual description of an image is fused with a caption to predict authors intent. The varying predictive power and noise typology may be the consequence of information coming from different modalities. Baltrušaitis et al. (2018) claim that two types of multimodal representations - joint and coordinated. Joint representation often projects multimodal input into a common space while coordinated representation project each modality into a separate but coordinated space where only one modality is present at test time.

Co-learning: The last challenge suggests transferring knowledge between different modalities, their representation, and their predictive models. It will be unexpectedly dominant when one of the modalities has limited resources such as a lack of annotated data, noisy data, and unreliable labels.

3. Methodology

3.1 Datasets

The study comprises data labeling and text analysis of a corpus published posts from the discourse community Instagram. The primary criterion for selecting objects was that they are 100 famous Key Opinion Leader (KOLs) in Taiwan in 2019. Ten posts are crawled employing each KOL's Instagram official web page using a python package beautifulSoup. Corresponding posts from April 20 in 2020 are collected with the goal of developing a rich and diverse set of posts. Since the posts contain not only texts and images, but hashtags, emojis, and name-taggings, those features would be directly analyzing by being integrated. Although the posts include a variety of texts, images, hashtags, emojis, and name-taggings to convey information, only under two circumstances are the posts collected. First, to ensure some homogeneity of meta-data, this study only includes images of photos, rather than any short video or long video in a post. The second circumstance is that we only recruit the first photo of multiple photos in a post if have ones. Currently, an amount of 906 posts are extracted from Instagram.

3.2 Annotation

Data were pre-processed, converting all albums to single ICPs. A simple annotation toolkit built on an online google sheet was developed and displayed with the form of ICPs. The annotators, who are non-expert in linguistics, are asked to confirm whether the data was acceptable and if so, to identify the post's intent, the contextual relationship, and the semiotic relationship. Every image was labeled by at least two independent human annotators. We retained only those images on which all annotators agreed. From the collected datasets, exploratory analysis can be conducted by an amount of analysis and visualizations. We take advantage of Cohen's Kappa to measure the agreement between different annotators who classify 936 items into three taxonomies. The scores of Cohen's Kappa calculated on the three taxonomies are 0.5803 (the contextual relationship), 0.5834 (the semiotic relationship), and 0.6223 (the author's intent). The scores for three taxonomies are all complied to the moderate agreement. The scores are used to ensure that annotators have high enough reliability in giving the same degree of annotating.

3.3 Model

After obtaining and annotating the caption and image on Instagram, it is necessary to compute embeddings when working with both contextual and image data in the machine learning pipeline. For text embedding, we utilize Sentence-BERT (Reimers & Gurevych, 2019) which is pretrained character-based contextual embeddings. For image embedding, we use ResNet50 (He et al., 2016) which has a model pretrained on ImageNet as the image encoder by implementing a Python package pic2vec to convert the image modality to embedding. After, based on the collected datasets, two Machine Learning models are applied to train the classifiers: Decision Tree Classifier (DTC's) (Swain & Hauska, 1977) and Random Forest (RF) (Breiman, 2001). In this study, both Machine Learning models were trained on both multimodal features. Our model takes input image (Img), text (Txt), or both (Img+Txt), plus modality-specific encoders, a fusion layer, and a class prediction layer. Consistent with our purpose that caption is seen as an integration, BERT sentence embedding fits the most because it considers caption as a whole sentence. For the combined encoding model, we take advantage of a simple fusion strategy that linearly projects encoded vectors from both modalities in the same embedding space and then adds two vectors (Kruk et al., 2019). According to Kruk et al. (2019), despite naive, this simple strategy has demonstrated high effectiveness at different related tasks. At last, we use the fused vector to predict scores with a fully connected layer.

4. Result and Discussion

We use a 906-sample dataset and only use a corresponding image and text information which is aligned manually for each post. Due to the small dataset, 10-fold cross-validation is conducted in our implementation. In order to report the result, the classification accuracy (ACC) and area under the ROC curve (AUC) are reported, using micro-average across all classes (Jeni et al., 2013; Stager et al., 2006). Additionally, for image, we use 2048 dimensional vectors trained from scratch. For character-based embeddings, we use a pretrained model with layers resulting in a 512-dimensional vector. The results will be shown after training in DTC and RF models. Due to the small dataset, we conducted a 10-fold cross-validation. To report the result, I'll present the classification accuracy (ACC) and area under the ROC curve (AUC) with micro-average across all classes. There will be two tables summarizing the main results as follows.

Table 1. Results with DTC's models – image only (Img), text-only (Text-BERT) and combined model (Img+Txt-BERT)

Method	**Contextual**		**Semiotic**		**Intent**	
	ACC	AUC	ACC	AUC	ACC	AUC
Img	50.54	50.70	51.37	49.86	67.23	50.00
Txt-BERT	72.67	69.08	73.63	67.67	82.20	65.14
Combined	70.97	67.88	71.94	64.49	81.82	64.89

First, the result with Decision Tree Classifier is shown in Table 1. Overall, the result shows a striking effect of text embedding on performance. For all the taxonomies, text embedding was significantly superior to the other models. It outperforms consistently than just using images embedding and the combined model. Following Kruk's work, we hypothetically assume that the combined model would help across the board. However, it has been disproved from several 'aspects of the result. For the contextual and semiotic relationship, text embedding and combined model both performs much better than image model for more than 20%. Toward this result, this might because Sentence-BERT is originally designed for text classification, so it outperforms in the text classification task. As for author's intent, the performance of text embedding (82.20%) is just slightly higher than image embedding (81.02%), and reaches almost 15% difference from the combined model (67.23%). The reason might be that the data of author's intent is more homogeneous than the others. Digging into the annotation details, when annotating the author's intent, annotators tend to label the target with high identical consistency, hence making the high performance in the author's intent.

Table 2. Results with RF models – image only (Img), text-only (Text-BERT) and combined model (Img+Txt-BERT)

Method	**Contextual**	**Semiotic**	**Intent**

	ACC	AUC	ACC	AUC	ACC	AUC
Img	62.45	84	60.37	83	81.02	92
Txt-BERT	84.43	94	83.99	91	86.23	95
Combined	78.31	93	80.48	90	82.36	95

Next, the results with the RF models are given in Table 2. Overall, like in Decision Tree Classifier, the result also shows a powerful performance of the text embedding. For the three taxonomies, text embedding (average 84.88) was significantly superior to those other models. It outperforms than the image embedding for 16.93% and than the combined model for 4.5%. For the contextual relationship, text embedding (84.83%) and the combined model (78.31%) both perform much better than Image model (62.45%) for at least 15.86%. As for semiotic relationship, text embedding (83.99%%) and the combined model (80.48%) both perform also much better than image model (60.37%) for more than 20%. Likewise, sentence-BERT again performs its advantage in the text classification experiment. As for author's intent, the performance of text embedding (86.23%) is just slightly higher than image embedding (82.36 %) and the combined model (81.02 %). The result of the Random Forest model tends to show its advantage in training the features, especially in the intent relationship (average 83.20%). Concerning the RF model, the author intent again displays its advantage of this task.

On one hand, compare the result of DTC's and RF, the accuracy of the text embedding of RF is even 8.71% higher than that of DTC's. The image embedding is 11.57% higher, and the combined model is 5.7% higher, in that RF is constructed on the foundation of multiple decision trees. Resulted from the growing ensemble of Decision trees, in this study, RF combines 100 tree predictors and makes them vote for the most popular class. By selecting features on randomly training and creating an amount of decision trees, it has significantly improved the classification accuracy. On the other hand, in comparison to the result of Kruk, this study performs better than Kruk's in both single modalities and combined modality. For text embedding, Kruk uses ELMO model, while this study uses Sentence-BERT architecture, showing an absolute advantage by improving the performance for 24.95%. For image embedding, this study utilizes ResNet50 instead of ResNet18 and makes progress on the performance by 15.65%. For the combined model, RF in this study outperforms Kruk's DCNN for 17.04%. Kruk has set the baseline classifier models as a preliminary effort, while this study examines dataset from a different domain and adapts more sounder classifiers, making outstanding performance in the multimodal document classification.

5. Conclusion

This study utilizes a computational model to capture the complex relationship between text and image modality, and how they cue authors intent in Instagram posts. In response to the first research question that if the combined modality stronger than individual modalities, text did assign a subordinate role to image. Toward this result, this might because Sentence-BERT is originally designed for text classification, and it performs very strong in the text classification task. Up to this point, these results are consistent with the prior studies which indicate that either text or image should be assigned a subordinate role. Although these two modalities encode different information on the use of classifying the relationship, the result suggests that the incorporating of text and image needs more effort to complement each other. The second main finding is that we present the results of the relationships of author's intent, the contextual relationship and the semiotic relationship between the ICPs. Furthermore, among the all studied modalities, the captions are no doubt the strongest feature for classifying relationships, in that the caption encoder (Sentence-BERT) shows its powerful advantages in text classification. In addition, we make two comparisons: the results between the two Machine Learning models and the result between this study and that of Kruk et al (2019).

Even though there are a variety of tasks being carried in the study, the design of the present study is not without limitations. The first limitation concerns the data size used in this current study. 906 posts might be too small to make a classification with high accuracy. The deep learning model used in these tasks needs more data to get better training. This may probably be one of the reasons that we cannot reach better performance in classifying the relationship with the image embedding model. The second limitation is rooted in the labor for annotation. The Kappa score in this study is to the average of 0.5953 (moderate agreement). If it may reach a substantial agreement (0.61-0.80) or even almost perfect agreement (0.81- 1.00), the result should appear to be better. In order to raise agreement between annotators, we should have completed more annotating norms with an abundance of details besides the two norms applied in this study. Third, a more stable architecture to build the image describer is needed. To characterize the images by representing an n-dimensional feature vector, it is necessary to explore the image describer to better generalize the results. Last but not least, the synergistic model in our task did not outperform the model of single modality. This issue might because the linearly project vectors did not highlight the significance of image embedding. Having acknowledged the limitations above, future studies should be alerted to the disadvantages of this study.

For future studies, new possibilities can be listed. Since sentence-BERT is good at processing the text input, the literal meaning of the image automatically generated from the computer vision technique should be involved as a part of the image encoder to reach a possibly better performance. Additionally, it is suggested that future research should explore more linguistic features of multimodal documents to improve the working. More solid visual features and other meta-data features are needed. Further, considering a large amount of literature in predicting sentiments of multimodal documents, the key challenge is to collect a sufficient amount of training labels to train a discriminative model for multimodal prediction. Although preliminary research in the area is already being undertaken by researchers, more extensive research would be necessary to make any definite claims along these lines.

References

[1] Wason, P. C., & Jones, S. (1963). Negatives: denotation and connotation. British Journal of Psychology, 54(4), 299–307.

[2] Jin, X., Gallagher, A., Cao, L., Luo, J., & Han, J. (2010). The wisdom of social multimedia: using flickr for prediction and forecast, In Proceedings of the 18th ACM international conference on Multimedia.

[3] Kruk, J., Lubin, J., Sikka, K., Lin, X., Jurafsky, D., & Divakaran, A. (2019). Integrating text and image: determining multimodal document intent in instagram posts. arXiv preprint arXiv:1904.09073.

[4] Castro, S., Hazarika, D., Pérez-Rosas, V., Zimmermann, R., Mihalcea, R., & Poria, S. (2019). Towards multimodal sarcasm detection (an _obviously_ perfect paper). arXiv preprint arXiv:1906.01815.

[5] O'Halloran, K. L., Tan, S., Wignell, P., Bateman, J. A., Pham, D.-S., Grossman, M., & Moere, A. V. (2019). Interpreting text and image relations in vio- lent extremist discourse: a mixed methods approach for big data analytics. Terrorism and Political Violence, 31(3), 454–474.

[6] Phan, T.-T., Muralidhar, S., & Gatica-Perez, D. (2019). # drink or# drunk: multimodal signals and drinking practices on instagram, In Proceedings of the 13th EAI International Conference on Pervasive Computing Technologies for Healthcare.

[7] Lemke, J. (1998). Multiplying meaning. Reading science: Critical and functional perspectives on discourses of science, 87–113.

[8] Summaries, P. E., & Panel, C. (n.d.). Esrc centre for corpus approaches to social science (cass).

[9] Chancellor, S., Kalantidis, Y., Pater, J. A., De Choudhury, M., & Shamma, D. A. (2017). Multimodal classification of moderated online pro-eating disorder content, In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems.

[10] Illendula, A., & Sheth, A. (2019). Multimodal emotion classification, In Companion Proceedings of The 2019 World Wide Web Conference.

[11] Zeppelzauer, M., & Schopfhauser, D. (2016). Multimodal classification of events in social media. Image and Vision Computing, 53, 45–56.

[12] Bateman, J. (2014). Text and image: a critical introduction to the visual/verbal divide. Routledge.

[13] Marsh, E. E., & White, M. D. (2003). A taxonomy of relationships between images and text. Journal of Documentation.

[14] Kloepfer, R. (1976). Komplementarität von sprache und bild am beispiel von comic, karikatur und reklame.(la complémentarité de la langue et de l'image. l'exemple des bandes dessinées, des caricatures et des réclames). Sprache in Technischen Zeitalter Stuttgart, (57), 42–56.

[15] Zhang, M., Hwa, R., & Kovashka, A. (2018). Equal but not the same: understanding the implicit relationship between persuasive images and text. arXiv preprint arXiv:1807.08205.

[16] Goffman, E. et al. (1978). The presentation of self in everyday life. Harmondsworth London.

[17] Mahoney, J., Feltwell, T., Ajuruchi, O., & Lawson, S. (2016). Constructing the visual online political self: an analysis of instagram use by the scottish electorate, In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems.

[18] Bongers, B., & van der Veer, G. C. (2007). Towards a multimodal interaction space: categorisation and applications. Personal and Ubiquitous Computing, 11(8), 609–619.

[19] Baltrušaitis, T., Ahuja, C., & Morency, L.-P. (2018). Multimodal machine learning: a survey and taxonomy. IEEE transactions on pattern analysis and machine intelligence, 41(2), 423–443.

[20] Morency, L.-P., & Baltrušaitis, T. (2017). Multimodal machine learning: integrating language, vision and speech, In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts.

[21] Le, Q., & Mikolov, T. (2014). Distributed representations of sentences and documents, In International conference on machine learning.

[22] Singla, K., Mukherjee, N., Koduvely, H. M., & Bose, J. (2018). Evaluating usage of images for app classificatio, In 2018 15th IEEE India Council International Conference (INDICON). IEEE.

[23] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition, In Proceedings of the IEEE conference on computer vision and pattern recognition.

[24] Yu, J., & Jiang, J. (2019). Adapting bert for target-oriented multimodal sentiment classification.

[25] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[26] Rahman, W., Hasan, M. K., Zadeh, A., Morency, L.-P., & Hoque, M. E. (2019). M-bert: injecting multimodal information in the bert structure. arXiv preprint arXiv:1908.05787.

[27] Barbieri, F., Ballesteros, M., Ronzano, F., & Saggion, H. (2018). Multimodal emoji prediction. arXiv preprint arXiv:1803.02392.

[28] Soleymani, M., Garcia, D., Jou, B., Schuller, B., Chang, S.-F., & Pantic, M. (2017). A survey of multimodal sentiment analysis. Image and Vision Computing, 65, 3–14.

[29] Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft coco: common objects in context, In European conference on computer vision. Springer.

[30] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. (2015). Imagenet large scale visual recognition challenge. International journal of computer vision, 115(3), 211–252.

[31] Breiman, L. (2001). Random forests. Machine learning, 45(1), 5–32.

[32] Reimers, N., & Gurevych, I. (2019). Sentence-bert: sentence embeddings using Siamese bert-networks. arXiv preprint arXiv:1908.10084.

[33] Swain, P. H., & Hauska, H. (1977). The decision tree classifier: design and potential. IEEE Transactions on Geoscience Electronics, 15(3), 142–147.

[34] Jeni, L. A., Cohn, J. F., & De La Torre, F. (2013). Facing imbalanced data–recommendations for the use of performance metrics, In 2013 Humaine association conference on affective computing and intelligent interaction. IEEE.

[35] Stager, M., Lukowicz, P., & Troster, G. (2006). Dealing with class skew in context recognition, In 26th IEEE International Conference on Distributed Computing Systems Workshops (ICDCSW'06). IEEE.