AGILE: GIScience Series, 5, 36, 2024. https://doi.org/10.5194/agile-giss-5-36-2024
Proceedings of the 27th AGILE Conference on Geographic Information Science, 4–7 Sept. 2024.
Editors: Alison Heppenstall, Mingshu Wang, Urska Demsar, Rob Lemmens, and Jing Yao.
This contribution underwent peer review based on a full paper submission.

Occupation Prediction with Multimodal Learning from Tweet Messages and Google Street View Images

Xinyi Liu [1,2], Bo Peng [1,3], Meiliu Wu [1], Mingshu Wang [4], Heng Cai[5], and Qunying Huang [1]

[1]Department of Geography, University of Wisconsin-Madison, Madison, Wisconsin
[2]Zoox Inc., Foster City, California
[3]PAII Inc., Palo Alto, California
[4]School of Geographical & Earth Sciences, University of Glasgow, UK
[5]Department of Geography, Texas A&M University, College Station, Texas

Correspondence: Qunying Huang (qhuang46@wisc.edu)

Abstract.

Despite the development of various heuristic and machine learning models, social media user occupation predication remains challenging due to limited high-quality ground truth data and difficulties in effectively integrating multiple data sources in different modalities, which can be complementary and contribute to informing the profession or job role of an individual. In response, this study introduces a novel semi-supervised multimodal learning method for Twitter user occupation prediction with a limited number of training samples. Specifically, an unsupervised learning model is first designed to extract textual and visual embeddings from individual tweet messages (textual) and Google Street View images (visual), with the latter capturing the geographical and environmental context surrounding individuals' residential and workplace areas. Next, these high-dimensional multimodal features are fed into a multilayer transfer learning model for individual occupation classification. The proposed occupation prediction method achieves high evaluation scores for identifying Office workers, Students, and Others or Jobless people, with the F1 score for identifying Office workers surpassing the best previously reported scores for occupation classification using social media data.

Keywords. GeoAI, multimodal learning, deep learning, social media user profiling, demographic prediction, transformer

1 Introduction

Predicting the occupation of social media users, involves analyzing content posted on social media platforms (e.g., Twitter) with language processing (Preoţiuc-Pietro et al., 2015a; Liang et al., 2018; Pardo and Rosso, 2019; Das et al., 2021) and image processing techniques (Wieczorek et al., 2018; Hu et al., 2021; Li et al., 2021) to infer the profession or job role of individuals based on their online behavior, interactions, and profile information (Preoţiuc-Pietro et al., 2015a; Hu et al., 2021). As one of the important tasks for demographic inference of social media data, also known as social media user profiling (Liang et al., 2018; Ikeda et al., 2013), user occupation prediction and categorization are important in understanding and interpreting the behaviors of various user groups (Preoţiuc-Pietro et al., 2015a), thus enabling applications for a variety of disciplines, such as sociology, demography, and public health (Ghazouani et al., 2020; Khanam et al., 2021).

However, predicting the occupations of social media users, presents two primary challenges. First, non-biased and high-quality ground truth data are scarce. Previous studies have often collected both labels and features from selective users who actively disclosed job-related information on social media platforms. These users are typically associated with specific occupations (e.g., professionals in fields like medication and management, illegal drug dealers) (Preoţiuc-Pietro et al., 2015b; Hu et al., 2021; Khanam et al., 2021), leading to biases toward certain occupations and hindering the analysis of generic ones. Second, achieving precise profiling with limited data sources, especially in one modality (e.g., textual message) is challenging. While language models offer some insights, distinguishing between generic occupations based solely on text information remains difficult (Preoţiuc-Pietro et al., 2015b; Liang et al., 2018; Abitbol et al., 2018).

To address this, previous studies used additional data sources, such as GPS trajectories collected via social media platforms and associated demographic variables (e.g.,

household income, educational attainment), along with geographic data, such as remote sensing images, activity-related point of interests (POIs), and Google Street View (GSV) images describing the physical environment of locations, to aid in personal SES inference (Abitbol et al., 2019; Xu et al., 2020; Poulston, 2021). Consequently, multimodal learning models are developed to integrate data from different sources and modalities (e.g., visual and textual) for user profiling tasks (Li et al., 2021). The principle behind multimodal data fusion is to combine the strengths of different modalities, extracting meaningful patterns from complex datasets and improving the accuracy and robustness of ML models (Gao et al., 2020; Li et al., 2021; Wu and Huang, 2022). However, the exploration of novel datasets and approaches for occupation classification is limited (Abitbol et al., 2019; Xu et al., 2020). Previous research has mostly independently used social media textual content (e.g., tweet texts) and visual information (e.g., GSV images) for user SES inference (Gebru et al., 2017; Aletras and Chamberlain, 2018). Fusing both modalities is challenging and requires more effective models since tweet messages rarely directly describe visual objects in GSV images, presumed in previous vision-language joint models (Gao et al., 2020; Yang et al., 2022).

In response, this study explores the associations between multimodal data and develops a semi-supervised multimodal learning method for classifying generic personal occupations with higher accuracy using widely available data sources: Twitter message content and GSV images. Our method classifies individuals' occupations into 4 generic categories which capture the majority of occupation profiles posted by the U.S. Bureau of Labor Statistics (2023), including Office workers, Students, individuals engaged in Arts, Design, Entertainment, Sports, and Media (ADESM), and Others or Jobless. The proposed method involves extracting textual and visual embeddings from individual tweet messages and GSV images, with the former capturing personal attributes and the latter capturing the geographic and environmental context near individuals' residential and workplace locations. These embeddings are then mapped to a shared semantic space (Kiela et al., 2017; Merkx and Frank, 2019), and an unsupervised learning model is developed to minimize their distance in the high-dimensional space (Kiela et al., 2017). The aligned embeddings are concatenated and serve as the input for a supervised transfer learning model for individual occupation classification. Finally, we evaluate the resulting model and assess the feasibility of the proposed method.

2 Methods

This research proposes a semi-supervised learning method to effectively utilize both widely available image and tweet datasets. The method (Figure 1) consists of three steps: (1) **The unsupervised learning** focuses on a large number of unlabeled users, training a linear projector atop a pre-trained image encoder to extract visual features and a text encoder to extract textual features. The objective is to project data from both modalities into the same space for comparison and similarity measurement (Merkx and Frank, 2019); (2) **The supervised learning** leverages the majority of a relatively small number of labeled users, utilizing model weights learned in module 1 to generate image and text embeddings for each pair of image and tweet messages. These embeddings are then concatenated and fed into a multi-layer classifier trained with manually identified occupation labels; and (3) **The inference** employs the learned weights from both the multimodal embedding and the multilayer classifier to create a testing model, which classifies remaining labeled users into the predefined 4 occupation categories (i.e., Office worker, Student, ADESM, Others or Jobless).

Overall, this method enables cross-modal learning by integrating images and texts into a unified representation by ensuring that both labeled and unlabeled data contribute to occupation classification. The subsequent sections will detail two key techniques of the proposed method, including multimodal embedding and multilayer classifier.

2.1 Multimodal Embedding

In the proposed semi-supervised learning method, the weights of the multimodal embedding component are obtained through unsupervised learning. This component includes an image encoder and a text encoder. To process images, we resize them to ensure the smaller side is 256 pixels while maintaining the original aspect ratio (Harwath et al., 2016). We extract ten 224x224 crops from each image, including one from each corner, one from the center, and five additional crops from the mirrored image (Harwath et al., 2016). These crops are then processed using a pre-trained ResNet-152 model trained on ImageNet (Wu et al., 2019), excluding the last classification layer. The resulting visual features are averaged to obtain a single vector with 2048 high-dimensional features. To consolidate features from all selected images of the same user, we employ element-wise mean pooling, averaging all visual feature vectors. This pooling generates a final set of 2048 features. These features are linearly projected and normalized to produce the image embeddings (i.e., emb_{img}) for each user, as defined in Equation 1:

$$emb_{img} = \frac{img \cdot A^T + b}{\|img \cdot A^T + b\|_2} \tag{1}$$

where A and b represent the learned weights and bias terms, respectively. The variable img denotes the pooled vector derived from ResNet image features, with a size of 2048 in our model. $\|img \cdot A^T + b\|_2$ represents the $L2$ norm of the original embedding vector. By dividing each component of the embedding vector by its $L2$ norm, the re-

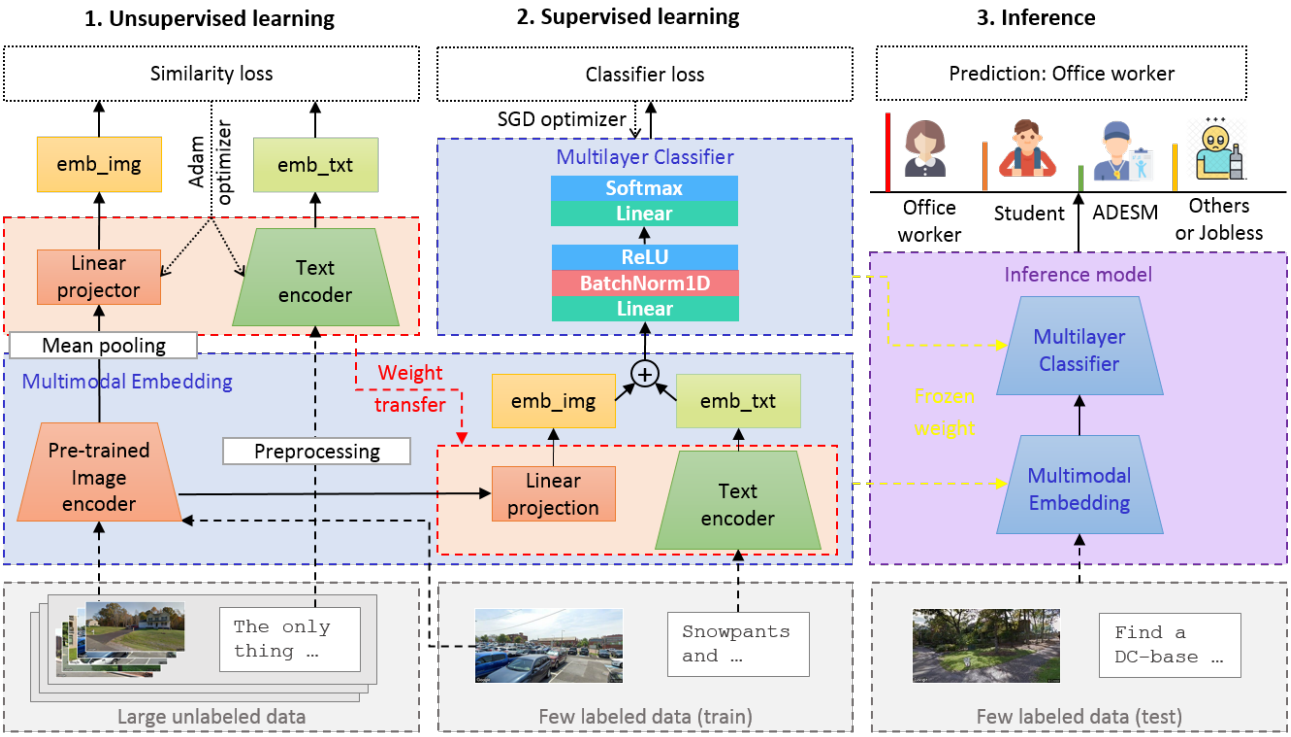

Figure 1. The proposed semi-supervised multimodal learning method for identifying Twitter users' generic occupation classes using their individual tweets and the GSV images collected around their home and work locations.

sulting unit vector retains the direction of the original vector but has a magnitude of 1. This normalization is commonly used in ML tasks to ensure that vectors are comparable and eliminates biases based on their original magnitudes (He et al., 2015).

Simultaneously, a text sentence encoder is trained on tweet content, treating it as word-level input(Harwath et al., 2018). This encoder (Equation 2) starts with an embedding layer that generates embeddings $(e_1, ..., e_t)$ for the t words present in the input sentence. Subsequently, these embeddings pass through a Long Short-Term Memory (LSTM) unit (Che et al., 2016), followed by a self-attention layer and normalization to possess a unit $L2$ norm (Merkx and Frank, 2019).

$$emb_{txt} = \frac{Att(LSTM(e_1, ..., e_t))}{\|Att(LSTM(e_1, ..., e_t))\|_2} \quad (2)$$

where $e_1, \ldots, e_t$ is used to represent word embeddings that capture the text content of a tweet. These embeddings assign dense vectors to individual words, encoding their semantic and contextual information. Att represents a self-attention layer with a single head, details of which can be referred to in the previous work (Merkx et al., 2022). The LSTM layer captures long-range dependencies within each tweet bidirectionally, producing 1024 embedding features in each direction. These results are concatenated to create a single embedding of size 2048.

Cosine similarity is used to compare image and text embeddings for each user(Merkx et al., 2019). An Adam optimizer is employed during training epochs to optimize this similarity Merkx et al. (2019). The training process enables the text encoder to understand the semantic meanings of text sentences with context provided by the corresponding GSV images, reflecting the user's living or working environment. These image-grounded text embeddingsKiela et al. (2017) emphasize shared features between different modalities, implicitly indicating the user's SES attributes, particularly occupation types.

2.2 Multilayer Classifier

In the supervised learning process (Step 2 in Figure 1), weights from the image embedding linear projector and the entire text encoder are transferred to the multimodal encoders. This yields emb_{img} and emb_{txt}, representing visual features and image-grounded textual features within the same semantic space (Kiela et al., 2017). Textual features can be directly utilized in a classifier for a transfer learning occupation classification task. Alternatively, in this study, visual and textual features are concatenated, forming a visual-enhanced image-grounded textual embedding to be fed into a multilayer occupation classifier.

The multilayer classifier is a sequential neural network model with two tiers of linear transformations (Figure 1). The first tier consists of three layers, while the second tier has two layers, effectively processing input data. Initially, input data is concatenated into a 4096-dimensional vector to undergo a linear mapping for dimension reduction (i.e.,

from 4096 to 2048), followed by the ReLU activation and batch normalization layers for stabilizing and speeding up training. In the second tier, another linear layer with a bias term maps the 2048-dimensional input to a 4-dimensional space, representing one-hot encoding of the 4 distinct occupation labels. A softmax activation layer converts values into a probability distribution denoting the probability of belonging to each occupation class.

Binary cross-entropy loss is computed by combining a sigmoid activation function, which avoids the need for separate application of the sigmoid for numerical stability and computational efficiency (Ruby and Yendapalli, 2020). This loss function quantifies the difference between predicted logits and one-hot encodings of target labels, aiming to minimize the loss by adjusting predicted logits. During training, a stochastic gradient descent optimizer is used for back-propagation and weight updating. This optimizer applies the gradient descent algorithm iteratively with small batches of training data, updating classifier weights to minimize the loss function and enhance model performance based on computed gradients.

3 Datasets

This study collects tweets and GSV image datasets from 103 eligible users for model training and evaluation. Initially, tweets from the Washington D.C. area were streamed using Twitter's API over 5 months. Subsequently, historical tweets of each DC user were re-harvested every six months between 2014 and 2016, with a maximum of 3200 tweets per re-harvesting. A total of 7660 eligible users met the criteria of having public accounts and posting over 40 geotagged tweets.

To refine the text data, a post-processing step removed common English stop words, such as "to," "as," and "is" and duplicated messages. Specific location names irrelevant to user profiling were also excluded. Messages with fewer than 5 words and users with fewer than 5 messages were discarded to ensure data representativeness. The dataset was reduced to 208 users, each with an average of 530 geotagged messages. Geotags consisted of latitude and longitude coordinates, facilitating the creation of individual online footprints with timestamps represented as (x, y, t). These footprints were aggregated using spatial clustering technique (Liu et al., 2019).

Next, GSV images are collected for these 208 users at distinct locations indicated by latitude and longitude pairs within 50 meters of the center location associated with each aggregated footprint cluster from geotagged tweet messages. However, the final selection only includes 103 users who have GSV images posted around at least one of their home or work locations. This is because higher accuracy is typically associated with identifying dwelling and work activities at these two types of locations. Figure 2 displays the image collections around home and work locations and tweet messages posted online for a selected user with the job title of a golf coach within the ADESM generic occupation category. It can be observed that one of the GSV images collected at the work location displays a large green space resembling a golf course. In addition, the tweet messages include keywords (e.g., 'tiger' and 'play') and emojis indicating golf-related activities.

Figure 2. Google Street View image collection around individual home and work locations and tweet message corpora for a selected user with the job title of a golf coach within the ADESM generic occupation category.

To encode the textual data, we created a word dictionary consisting of 6,828 unique meaningful words to build word indices (Merkx et al., 2022). The entire dataset was utilized in the unsupervised learning process to train the multimodal encoders for mapping visual and textual embeddings into a shared semantic space.

During training in the unsupervised module, 55 users were used to train the image encoder linear projector and the text encoder, with 12 users as validation data to prevent overfitting. In the supervised learning process, the remaining 36 users were split into 28 for model training and 8 for testing (Figure 1).

4 Performance Evaluation

The proposed occupation inference model (Figure 1) is implemented using a single tweet message and its corresponding pooled image embeddings as input. These inputs are fed into the previously trained multimodal encoders, which were trained over 16 epochs, and multilayer classifiers. Training data comprises 28 labeled users out of 36 total users. Specifically, a dataset of 100 data points for testing is created by randomly selecting 100 pairs of text messages and the pooled image embeddings from the remaining 8 test users, using stratified sampling based on their labeled occupation classes. The inferred results are compared with predefined manual labels using three evaluation scores: precision, recall, and F1 score (Table 1). Additionally, the inference model is also assessed using the same input data but with less training on the multimodal encoders, specifically, with only 4 or 8 training epochs. The examination demonstrates that, when trained for a higher number of epochs (e.g., 16 epochs), all occupation types except for ADESM achieve an F1 score

Table 1. Precision, recall, and F1 scores for identifying the 4 predefined generic occupation classes (i.e., Arts, Design, Entertainment, Sports, and Media (ADESM), Office worker, Student, Others or Jobless) with increasing numbers (i.e., 4, 8, 16) of training epochs

Occupation Class	Pr			Re			$F1$			Support
	4ep	8ep	16ep	4ep	8ep	16ep	4ep	8ep	16ep	
ADESM	**0.39**	**0.39**	0.13	0.76	0.34	**0.87**	**0.514**	0.36	0.23	21
Office worker	0.22	0.28	**0.69**	0.16	0.48	**0.78**	0.19	0.35	**0.73**	25
Student	0.29	0.14	**0.94**	**0.86**	0.11	0.33	0.22	0.12	**0.48**	29
Others/Jobless	**0.60**	0.39	0.58	0.36	0.53	**0.59**	0.29	0.44	**0.6**	25

exceeding or around 0.5. Among these, the identification of Office workers exhibits the best performance, achieving an F1 score of 0.73. This indicates the stronger alignment of multimodal embeddings through extensive training. However, evaluation scores for identifying ADESM occupations decrease with more training epochs, unlike other occupation types. This suggests that stronger alignment negatively affects ADESM detection, possibly due to the complexity of this type. Breaking down ADESM into smaller groups could enhance the investigation. Furthermore, reverse similarity metrics between multimodal embeddings might offer updated identification results, revealing underlying correlations.

ADESM shows high recall but low precision, with Students often misclassified as ADESM workers due to textual similarity. Distinguishing between Others or Jobless and Students with part-time jobs adds to classification challenges. Conversely, Office workers exhibit distinctive features in language usage and visual environments, achieving the best evaluation scores.

Occupation type classification in this study is dataset-specific with limited labeling. Nonetheless, the proposed method effectively extracts features and identifies occupations based on users' travel trajectories.

5 Conclusions and Future Work

Previous studies have faced challenges for social media user occupation prediction due to relying on biased datasets generated from self-reported biographies or job-related keywords as occupation labels. Additionally, prior investigations have focused on extracting features either from people's text messages or images that portray their living environment. Multimodal learning has shown promise in faster language processing by incorporating image context. Nevertheless, the application of multimodal learning to associate image data with short sentences of human-spoken languages that do not explicitly describe the same physical objects or phenomena has been largely unexplored.

In response, the proposed occupation inference method aligns visual embeddings from GSV images with textual embeddings from tweets, despite the texts not directly describing the images. Utilizing a multilayer classifier, the method fuses multimodal embeddings to classify users into generic occupation classes without relying on self-reported biographies. Experimental results 103 Twitter users indicate that the method achieves high evaluation scores for identifying Office worker, Student, and Others or Jobless people, surpassing previous scores for social media data classification.

Although the method shows promise, further exploration is needed to optimize architectures for multimodal encoders and the multilayer classifier. In particular, further investigation of optimal architectures for both multimodal encoders and the multilayer classifier is necessary to improve user profiling effectiveness. This includes further exploration of the correlation between embedding alignment and multimodal classification. For example, additional loss functions could be examined to define the optimal alignment for profiling generic occupation types with suboptimal performance (e.g., ADESM) in this study. In addition, future efforts could incorporate additional datasets and features for enhanced user profiling.

Acknowledgment

This research is funded by the National Institute of Health (R01DA047315), and National Institute of Food and Agriculture (WIS04084). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the funders.

References

2023: May 2021 National Occupational Employment and Wage Estimates, https://www.bls.gov/oes/current/oes_stru.htm#31-0000.

Abitbol, J., Fleury, E., and Karsai, M.: Optimal Proxy Selection for Socioeconomic Status Inference on Twitter, Complexity, 2019, 1–15, https://doi.org/10.1155/2019/6059673, 2019.

Abitbol, J. L., Karsai, M., and Fleury, E.: Location, occupation, and semantics based socioeconomic status inference on twitter, in: 2018 IEEE International Conference on Data Mining Workshops (ICDMW), pp. 1192–1199, IEEE, 2018.

Aletras, N. and Chamberlain, B.: Predicting Twitter User Socioeconomic Attributes with Network and Language Information, pp. 20–24, https://doi.org/10.1145/3209542.3209577, 2018.

Che, Z., Purushotham, S., Cho, K., Sontag, D., and Liu, Y.: Recurrent Neural Networks for Multivariate Time Series with Missing Values, 2016.

Das, K. G., Patra, B. G., and Naskar, S. K.: Profiling Celebrity Profession from Twitter Data, in: 2021 International Conference on Asian Language Processing (IALP), pp. 207–212, https://doi.org/10.1109/IALP54817.2021.9675260, 2021.

Gao, J., Li, P., Chen, Z., and Zhang, J.: A Survey on Deep Learning for Multimodal Data Fusion, Neural Computation, 32, 829–864, https://doi.org/10.1162/neco_a_01273, 2020.

Gebru, T., Krause, J., Wang, Y., Chen, D., Deng, J., Aiden, E. L., and Fei-Fei, L.: Using deep learning and Google Street View to estimate the demographic makeup of neighborhoods across the United States, Proceedings of the National Academy of Sciences, 114, 13 108–13 113, https://doi.org/10.1073/pnas.1700035114, 2017.

Ghazouani, D., lancieri, l., ounalli, h., and Chaker, J.: Assessing Socioeconomic Status of Twitter Users: A Survey, Advanced Composites Letters, 2020.

Harwath, D., Torralba, A., and Glass, J.: Unsupervised Learning of Spoken Language with Visual Context, in: Advances in Neural Information Processing Systems, edited by Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., and Garnett, R., vol. 29, Curran Associates, Inc., https://proceedings.neurips.cc/paper_files/paper/2016/file/82b8a3434904411a9fdc43ca87cee70c-Paper.pdf, 2016.

Harwath, D., Recasens, A., Surís, D., Chuang, G., Torralba, A., and Glass, J.: Jointly Discovering Visual Objects and Spoken Words from Raw Sensory Input, 2018.

He, K., Zhang, X., Ren, S., and Sun, J.: Deep Residual Learning for Image Recognition, 2015.

Hu, C., Yin, M., Liu, B., Li, X., and Ye, Y.: Identifying Illicit Drug Dealers on Instagram with Large-Scale Multimodal Data Fusion, ACM Trans. Intell. Syst. Technol., 12, https://doi.org/10.1145/3472713, 2021.

Ikeda, K., Hattori, G., Ono, C., Asoh, H., and Higashino, T.: Twitter user profiling based on text and community mining for market analysis, Knowledge-Based Systems, 51, 35–47, https://doi.org/https://doi.org/10.1016/j.knosys.2013.06.020, 2013.

Khanam, K. Z., Srivastava, G., and Mago, V.: Identifying health related occupations of Twitter Users through word embedding and deep neural networks, 2021.

Kiela, D., Conneau, A., Jabri, A., and Nickel, M.: Learning Visually Grounded Sentence Representations, CoRR, abs/1707.06320, http://arxiv.org/abs/1707.06320, 2017.

Li, L., Hu, K., Zheng, Y., Liu, J., and Lee, K. A.: COOP-Net: Multi-Modal Cooperative Gender Prediction in Social Media User Profiling, in: ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4310–4314, https://doi.org/10.1109/ICASSP39728.2021.9414808, 2021.

Liang, S., Zhang, X., Ren, Z., and Kanoulas, E.: Dynamic Embeddings for User Profiling in Twitter, in: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD '18, p. 1764–1773, Association for Computing Machinery, New York, NY, USA, https://doi.org/10.1145/3219819.3220043, 2018.

Liu, X., Huang, Q., and Gao, S.: Exploring the uncertainty of activity zone detection using digital footprints with multi-scaled DBSCAN, International Journal of Geographical Information Science, 33, 1196–1223, https://doi.org/10.1080/13658816.2018.1563301, 2019.

Merkx, D. and Frank, S. L.: Learning semantic sentence representations from visually grounded language without lexical knowledge, Natural Language Engineering, 25, 451–466, https://doi.org/10.1017/S1351324919000196, 2019.

Merkx, D., Frank, S. L., and Ernestus, M.: Language Learning Using Speech to Image Retrieval, in: Interspeech 2019, ISCA, https://doi.org/10.21437/interspeech.2019-3067, 2019.

Merkx, D., Frank, S. L., and Ernestus, M.: Seeing the advantage: visually grounding word embeddings to better capture human semantic knowledge, 2022.

Pardo, F. M. R. and Rosso, P.: Overview of the 7th Author Profiling Task at PAN 2019: Bots and Gender Profiling in Twitter, in: Conference and Labs of the Evaluation Forum, 2019.

Poulston, A. R. S.: User profiling with geo-located social media and demographic data, Ph.D. thesis, University of Sheffield, 2021.

Preoţiuc-Pietro, D., Lampos, V., and Aletras, N.: An analysis of the user occupational class through Twitter content, 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), pp. 1754–1764, Association for Computational Linguistics, Beijing, China, https://doi.org/10.3115/v1/P15-1169, 2015b.

Preoţiuc-Pietro, D., Volkova, S., Lampos, V., Bachrach, Y., and Aletras, N.: Studying User Income through Language, Behaviour and Affect in Social Media, PLOS ONE, 10, 1–17, https://doi.org/10.1371/journal.pone.0138717, 2015a.

Ruby, U. and Yendapalli, V.: Binary cross entropy with deep learning technique for Image classification, International Journal of Advanced Trends in Computer Science and Engineering, 9, https://doi.org/10.30534/ijatcse/2020/175942020, 2020.

Wieczorek, S., Filipiak, D., and Filipowska, A.: Semantic Image-Based Profiling of Users' Interests with Neural Networks, https://doi.org/10.3233/978-1-61499-894-5-179, 2018.

Wu, M. and Huang, Q.: IM2City: Image Geo-Localization via Multi-Modal Learning, in: Proceedings of the 5th ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery, GeoAI '22, p. 50–61, Association for Computing Machinery, New York, NY, USA, https://doi.org/10.1145/3557918.3565868, 2022.

Wu, Z., Shen, C., and Van Den Hengel, A.: Wider or deeper: Revisiting the resnet model for visual recognition, Pattern Recognition, 90, 119–133, 2019.

Xu, F., Lin, Z., Xia, T., Guo, D., and Li, Y.: Sume: Semantic-enhanced urban mobility network embedding for user demographic inference, Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 4, 1–25, 2020.

Yang, J., Duan, J., Tran, S., Xu, Y., Chanda, S., Chen, L., Zeng, B., Chilimbi, T., and Huang, J.: Vision-Language Pre-Training with Triple Contrastive Learning, 2022.