Age Inference Using A Hierarchical Attention Neural Network

Yaguang Liu, Lisa Singh
Georgetown University
Washington, USA
{yl947,lisa.singh}@georgetown.edu

ABSTRACT

While demographic attributes, such as age, gender, and location, have been extensively studied, most previous studies usually combine different sources of data, such as the user's biography, pictures, posts, and the user's network to obtain reasonable inference accuracies. However, it is not always practical to collect all those different forms of data. Therefore, in this paper, we consider methods for inferring age that only use Twitter posts (tweet text and emojis). We propose a hierarchical attention neural model that integrates independent linguistic knowledge gained from text and emojis when making a prediction. This hierarchical model is able to capture the intra-post relationship between these different post components, as well as the inter-post relationships of a user's posts. Our empirical evaluation using a data set generated from Wikidata demonstrates that our model achieves better performance than the state-of-the-art models, and still performs well when the number of posts per user is reduced in the training data set.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**.

KEYWORDS

age inference; hierarchical attention neural network; BERT

ACM Reference Format:
Yaguang Liu, Lisa Singh. 2021. Age Inference Using A Hierarchical Attention Neural Network. In *Proceedings of the 30th ACM Int'l Conf. on Information and Knowledge Management (CIKM '21), November 1–5, 2021, Virtual Event, Australia.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3459637.3482055

1 INTRODUCTION

Demographic inference of age, gender, location and occupation using social media data has been extensively studied [1, 3, 8, 9, 15, 20, 22, 28]. The most prevalent studies use different combinations of user biography, posts, images, network, etc. However, for some use cases, it is not always practical to collect all these different types of data. For example, if a researcher uses the Twitter Streaming API, he/she may collect posts based on a keyword or hashtag. If the hashtag of interest is actively used, then millions of users may be connected to the posts, making it difficult to get profile and network data for each user. Therefore, we consider the problem of demographic inference in a constrained data environment, where we have 10s to 100s of posts associated with each user, but no other account information. More specifically, in this paper, we investigate the viability of using only post content to infer age bins of different Twitter users. Ultimately, we ask, are there sufficient linguistic differences in posts among users in different age groups to distinguish them? Given previous work that highlights the importance of emojis for gender inference [23], we are also interested in understanding the role emojis play for age inference and their importance within deep learning models. Finally, we consider the variation in activity level of Twitter users. According to Pew Research, the median user posts a tweet once a month, while more prolific users post over 150 times per month[32]. Given this wide range in user engagement, we are interested in determining how many posts are needed to maintain a high level of accuracy. Given our goals, we propose a deep learning hierarchical network that attempts to model the inter-post and intra-post relationships of both post text and emojis independently, before combining the knowledge.

Our contributions in the paper are as follows. (1) Our model explores inter-post relationships by using BERT and a hierarchical network with attention. (2) We incorporate emojis into the deep learning model more directly than has been done in previous literature. (3) We investigate the effect on F1 score of using different numbers of posts for users in the training data. (4) We release our preprocessed Wikidata data set for researchers to use.

2 RELATED LITERATURE

Algorithms for age inference can be divided into two groups: classic supervised learning models and deep learning models. Classic supervised models that have performed particularly well on this task include logistical regression, random forest, and support vector machines. Some of these models use user profile information (biography and/or profile image) to determine age group [4, 33]. Others build models using ngrams from posts as features or ngrams constructed from both the biography and post [19, 24]. Rosenthal and McKeown [27] also incorporate stylistic features, e.g. punctuation.

More recently, researchers have found deep learning models to be useful for this task. Wang and colleagues [33] investigate using profile-based features, such as profile image, within deep learning models to achieve state of the art performance. A graph-based Recursive Neural Networks (RNN) [7] using word embeddings to represent posts is proposed by Kim et al. [11]. Their model incorporates not only the posts of a user, but also the posts of a user's network. Liu et al. [15] incorporate user biographies and post content into their deep learning model by using sentence-level embeddings generated for each tweet from both original BERT and a fine-tuned BERT to capture semantically similar sentences. Our

CIKM '21, November 1–5, 2021, Virtual Event, Australia.

ACM ISBN 978-1-4503-8446-9/21/11…$15.00
https://doi.org/10.1145/3459637.3482055

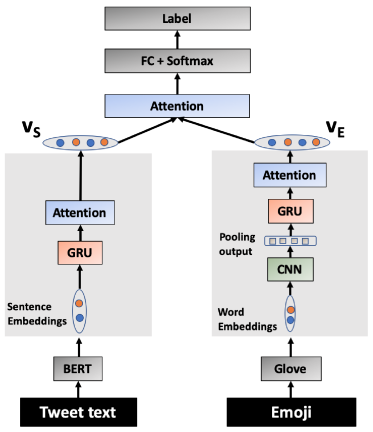


Figure 1: Overview of the proposed model

work differs because we only use post text when training and we propose using a hierarchical model.

Yang et al. [34] propose using a hierarchical network with attention for document classification. The hierarchical structure mirrors the hierarchical structure of documents. Specifically, a document can be split into sentences and each sentence can be split into words. The model was designed to capture both these levels, thereby allowing for the exploration of the intra-sentence and inter-sentence relationships. This network structure has been adapted for inference of location [9, 17]. Again, both of these works use more than post information to build their models. They also focus on word embeddings instead of sentence embeddings, thereby losing valuable contextual information.

Emojis have been used in different classic models for age inference, and are generally represented as a numeric feature [6, 16]. For example, Reifman and colleagues [25] use a log-linear model and show that young-adults preferred symbolic emojis when greeting others, while older/middle-aged adults opted for textual communication. Emojis have not been widely explored in deep learning models. Liu et al. [15] include percentage of emojis as a numeric feature in their deep learning model. In this paper, we want to explore emoji usage as a language construct within deep learning models by converting emojis to word embeddings and training them using a convolutional neural network (CNN) and a RNN.

3 MODEL DESCRIPTION

This section describes the construction of our proposed model and our approach for constructing features. Figure 1 presents an overview of the proposed model. Our model maps tweet text and emojis extracted from posts into two separate embedding spaces, allowing them to be trained as two independent components. Specifically, each type of data is sent into a RNN with attention. We use the gated recurrent unit (GRU) variant of the RNN. Each component is designed as an independent hierarchical neural network with attention, similar to Yang [34], to capture both posts that are important at the user level and sentences/phrases that are important at the post level. Finally, the last step combines the information from the two models to obtain the final prediction.

3.1 Tweet Text Hierarchical Network

BERT Sentence Embedding Previous work using hierarchical neural networks has used word embedding with a RNN to represent a sentence in the first layer of the hierarchy. However, since models using word embeddings analyze text one word at a time, they may miss important contextual differences at the sentence level. For example, suppose we have the following posts: 1) I just ate an apple. 2) I like Apple computers. Word embedding models will generate the same embedding for the two different contexts of the word 'apple'. In order to better represent the intra-text relationship, we use BERT, to generate contextualized word embeddings, therefore, capturing the contextual differences of the two sentences. We generate the sentence embedding for each tweet by averaging the word embeddings from the BERT output layer. Specifically, with the the pretrained uncased BERT-Base model,[1] we use SentenceBert [26] to map each tweet into a vector $s \subset R^d$ (d=the dimension of the vector) and get the embedding representation for the tweet.[2]

Tweet Text Encoder Suppose, we come across two posts from a user 1) I don't like crowded places. 2) However, I like the busy city, New York. How can we make the model draw the inference between 'crowded places' and 'busy cities'? To explore such inter-text relationship, we use the GRU [5] variant of a RNN to encode the tweet text, thereby avoiding the problem of vanishing and exploding gradients that can occur when using a RNN [2]. Given tweet i and tweet vector s_i, suppose each user has N posts. We use a bidirectional GRU to encode tweet text:

$$\overrightarrow{h_i} = \overrightarrow{GRU}(s_i), i \subset [1, N]$$

$$\overleftarrow{h_i} = \overleftarrow{GRU}(s_i), i \subset [N, 1]$$

We concatenate $\overrightarrow{h_i}$ and $\overleftarrow{h_i}$ to get an annotation of the tweet text i, i.e., $h_i = [\overrightarrow{h_i}, \overleftarrow{h_i}]$. Here, h_i summarizes the text of the tweets near tweet i, but still focuses on the text in tweet i.

Tweet Text Attention We use an attention mechanism to reward tweet texts that provide additional context for correctly classifying a user, and introduce a tweet/sentence level context vector u_S that measures the importance of the posts. This yields

$$u_i = tanh(W_S h_i + b_S),$$

$$\alpha_i = \frac{exp(u_i^T u_S)}{\sum_i exp(u_i^T u_S)},$$

$$v_S = \sum_i \alpha_i h_i$$

The tweet annotation h_i is fed through a MLP to get u_i as a hidden representation. Then we measure the importance of the tweet text with a context vector u_S and get a weight α_i through a softmax function. Finally, we compute the tweet document vector v_S as a weighted sum.

3.2 Emoji Hierarchical Network

In this section, we explain our emoji hierarchical network, focusing on the emoji encoding. After the encoding step, the remaining steps are the same as the text hierarchical network.

[1]The uncased BERT-Base model was pretrained on the BookCorpus dataset consisting of 11,038 unpublished books and English Wikipedia (excluding lists, tables and headers).
[2]We have tried Siamese BERT presented in [15] but it showed no improvement.

Demographics		Category	Count
Age	Bin 2	<45	9540
		>=45	4775
	Bin 3	<35	6526
		35-54	5047
		>=55	2742

Table 1: Ground truth data distribution for age

Embedding Layer Singh and colleagues show that incorporating emoji descriptions into the learning process can improve tweet classification for different sentiment analysis tasks [29]. So, here we apply the same strategy. For each post p_i, we extract emojis $e_1, e_2, ..., e_n$. Then we use Python library *emoji* [10] to convert emojis into a description. For example, the emoji 😄 is converted into *grinning_face_with_smiling_eyes*. We then extract the words from the description. Since they are just explanatory words instead of sentences that has contexts, word orders, etc., we use GloVe [21] with a dimension d_e. We introduce a hyper-parameter L as a threshold. The number of emoji words less than L are padded while the ones with more than L are truncated. Finally we concatenate the emoji vectors for each post p_i to generate the emoji matrix $M_{emoji}{}^{L \times d_e}$ as the representation of each tweet's emojis. Because emoji are mapped into meaningful words, there is not much noise introduced and we therefore, use the original pre-trained GloVe model to represent words. In cases where no word vector is found for a particular word, we randomly initialize them.

CNN Layer Because we believe emoji order has less importance than text order in a sentence, we choose to use a CNN model as opposed to a sequential one. In the embedding layer, we map each word into a word embedding space and get a vector $v_{e_i} \subset R^{d_e}$. For the convolutional layer, we apply a filter of window size k over the emoji matrix. Specifically, each filter $f \subset R^{k \times d_e}$ generates a feature vector $\theta = [\theta_1, \theta_2, ..., \theta_{L-k+1}] \subset R^{L-k+1}$, where $\theta_i = relu(u \odot v_{e_i:e_{i+k-1}+b})$ and b is a bias term and $\odot$ represents the convolution action between u and window $v_{e_i:e_{i+k-1}+b} \subset R^{k \times d_e}$ (a sum over element-wise multiplications). Next, we apply a Pooling Layer to the matrix and select the most representative feature $\hat{\theta} = max(\theta)$. With t windows each having m such filters, we get the tweet emoji embedding for each post $CNN_e(emoji) \subset R^{t*m}$.

Tweet Emoji Encoder Similar to the text hierarchical network, we apply a GRU network to the tweet emoji embedding of all posts of a user. Then, we concatenate the outputs from the GRU to get annotations of the emojis of the posts.

Tweet Emoji Attention To reward emojis that are more helpful for a document classification, we again use an attention mechanism as tweet text and generate the emoji vector v_E.

3.3 Feature Fusion Attention Layer

Since we do not know whether text or emojis are more important for the age inference task, we choose to consider them independently and then incorporate an additional attention layer to combine information from the text hierarchical network and the emoji hierarchical network. In this way, different weights are assigned to the independent components depending upon their value for the task.[3] Finally, we include a fully connected (FC) layer and Softmax, which generates a probability distribution for each class before returning the final prediction.

Parameter	Value
Batch size	32
Learning rate	0.0001
Word embedding dimension	50
Sentence embedding dimension	768
Filter window sizes	2, 3, 4
Filter number for each size	256
Emoji threshold length	30
Maximum number of tweets per users	200

Table 2: A summary of hyperparameter settings

4 EMPIRICAL EVALUATION

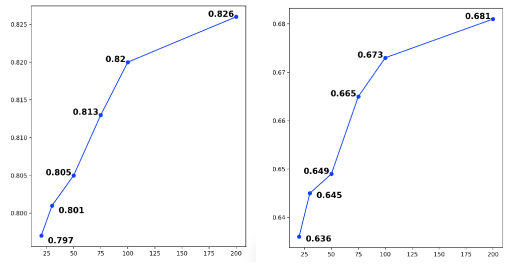
4.1 Experiment Setup

Data set Similar to Liu and colleagues [15], we use a data set from Wikidata [31], and collect data for the Twitter handles using the Twitter API.[4] For all the models, we remove users that have less than 20 English tweets. For the classic models, we also remove stopwords, handles, mentions, and lowercase all of the words. For the deep learning models, we only remove handles for privacy reasons. The average number of tweets per user is 553. We have 14,315 users in the training set.[5] Table 1 shows the number of users in each age category for our two different age groupings, 2-bin and 3-bin. Age 45 defines a new era of adulthood based on the Levinson adult development model [14]. Thus, we choose 45 as the 2-bin dividing line. Our 3-bin boundaries were identified by social science experts [15]. [6] Because we have significantly more younger users, we randomly sample from the group using the Python library *imblearn* [13] in order to create a more balanced data set.

Baseline Models We compare our model and a simple variant to five other models. (1) Nguyen and colleagues [19] use logistic regression (LR) with unigrams for age inference. (2) Morgan-Lopez and colleagues [18] incorporate bigrams and trigrams into a LR model. (3) Random forest (RF) has also been successful for this task [6, 30]. We also implemented both the (4) vanilla BERT and (5) Siamese BERT models from [15] as it achieves state of the art performance. The variant we present is the concatenation version of feature fusion (CFF), where we train the text and emojis within a single component as opposed to training them independently first.

Experiment settings We use 2 NVIDIA Tesla P4 GPUs each having 16 GBs of memory. Table 2 shows the hyper-parameters settings. We use the Adam update rule [12] to optimize our model. We randomly initialize the weight, bias and context vector for the attention mechanism and then normalize them with a mean value of 0 and a standard deviation of 0.05. They are jointly learned during

[3] In our empirical evaluation, we also show results from a model that concatenates the textual and emoji embedding before inputting them into a RNN with attention, thereby training the components together instead of separately.
[4] Other released data sets are very small. We are interested in larger data sets to avoid overfitting.
[5] The preprocessed data can be found at https://portals.mdi.georgetown.edu/public/demographic-inference
[6] Given the width of the bins, we are able to use more than one year of data.



(a) f1 score of 2-bins (b) F1 score of 3-bins

Figure 2: F1 score for 2-bins and 3-bins using different number of tweets per user

training. We use 5-fold cross validation with three different random seeds and report the average F1 score and standard deviations.

Model	Age (2 bins)	Age (3 bins)
Unigram-RF	0.812±0.003	0.643±0.003
Nguyen et al.	0.792±0.003	0.641±0.006
Morgan-Lopez et al.	0.794±0.005	0.643±0.002
Vanilla BERT	0.785±0.004	0.617±0.006
Siamese BERT	0.784±0.006	0.610±0.003
CFF	0.824±0.005	0.682±0.004
Proposed model	**0.833±0.007**	**0.684±0.004**

Table 3: F1 score for age

4.2 Experiment results

Results of Experiment 1 Table 3 presents our prediction results. Beginning with the 2-bin age results, we see that the best classic ML model is Random Forest with an F1 score of 0.812. Vanilla Bert has an F1 score of 0.785, which is lower than the classic ML models. This finding is consistent with previous research that uses text features, i.e., a general deep learning model results in little to no improvement in F1 score. For the proposed model, we observe a 2.1% improvement over the best classic model and a 4.8% improvement over the best neural network model. The proposed model has an F1 score that is 0.9% higher than CFF, which demonstrated that separating emojis and text can improve the result. Overall, we see that our proposed model performs 2% to 5% better than the state of the art, and 1% better than the combined model.

For the 3-bin age results, all the classic models perform similarly. Random Forest and the model proposed by Morgan-Lopez et al. perform slightly better with an F1 score of 0.643. The best BERT model has an F1 score of 0.617, demonstrating that it is difficult for a simpler deep learning model to identify more subtle linguistic differences for this task. For the proposed model, we can see there is a 4.1% improvement when compared to the best classic model and a 6.7% improvement when compared to the best BERT model. It is also 0.2% higher than CFF. These results highlight that the use of both the hierarchical model and emojis is important for capturing language difference by age.

Results of Experiment 2 In order to show how the number of posts per user impacts the model construction, we conduct a sensitivity analysis that varies the number of tweets per user from 20 to 200 (see Figure 2). We find that for the 2-bin inference task, there is a 3% difference in F1 score between our model built using 20 tweets per user and the one built using 200 tweets, indicating that the algorithm is fairly robust to the number of tweets. For the 3-bin model, the difference is closer to 4.5%, indicating that the algorithm is a little less robust to the number of tweets for the more difficult inference task. Still, in both cases, one can use 100 tweets and obtain an F1 score that is less than 1% lower. In general, given the situation, researchers may be willing to take a small hit in performance to collect less data.

Features	Age (2 bins)	Age (3 bins)
Text+emoji (proposed model)	0.833±0.007	0.684±0.004
Tweet text (hierarchical)	0.821±0.003	0.679±0.001
Tweet text (summing)	0.785±0.004	0.617±0.006

Table 4: Ablation study with F1 score

Ablation Study To evaluate the contribution of each group of features for our task, we conduct an ablation study. Table 4 shows the results. We can see that without using emojis, the model's F1 score is lowered by 1.2% for 2 bins and 0.5% for 3 bins. This shows that emojis do help improve the overall performance, but not by as much as we may have expected. We believe that this is because while emoji usage does differ between those who are younger and older, it is not as important for the 3-bin case. Our finding is consistent with Reifman and colleagues [25]. They find that emoji differences are more significant between old and young groups. Using tweet text alone with a hierarchical network, we see an improvement of 3.6% in the 2-bin case and 6.2% in the 3-bin case when compared to using tweet text summation only. This result highlights the importance of our hierarchical network.

5 CONCLUSIONS AND FUTURE WORK

In this paper we explore the use of hierarchical neural networks for capturing linguistic features from tweet text and emojis to predict age. When comparing to state of the art methods, we find that our approach achieves a higher F1 score for both the 2 class and 3 class problems. We also find that our model is robust to training with a smaller number of tweets per user. Future work includes exploring separating out different constructed features, trying other large-scale data sets, and building models that can handle even smaller amounts of training data.

6 ETHICAL CONSIDERATIONS

We acknowledge that the detection of user demographics poses unique ethical considerations. While automated methods can be valuable, error does exist in these models and there are possible equity and justice related consequences to imbalances in these errors. It is clear that public data should not be used to compromise reasonable privacy expectations. This is the reason we use Wikidata, where users choose to share their Twitter handle and age publicly.

ACKNOWLEDGEMENTS

This work was supported by National Science Foundation awards #1934925 and #1934494, the National Collaborative on Gun Violence Research (NCGVR) and the Massive Data Institute (MDI) at Georgetown University.

REFERENCES

[1] F. Al Zamal, W. Liu, and D. Ruths. 2012. Homophily and Latent Attribute Inference: Inferring Latent Attributes of Twitter Users from Neighbors. In *AAAI Conference on Weblogs and Social Media*.

[2] Y.a Bengio, P. Simard, and P. Frasconi. 1994. Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks* (1994).

[3] B. Chamberlain, C. Humby, and M. Deisenroth. 2017. Probabilistic inference of twitter users' age based on what they follow. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*.

[4] X. Chen, Y. Wang, E. Agichtein, and F. Wang. 2015. A Comparative Study of Demographic Attribute Inference in Twitter. In *AAAI Conference on Weblogs and Social Media*.

[5] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078* (2014).

[6] Joran Cornelisse and Reshmi Gopalakrishna Pillai. 2020. Age Inference on Twitter using SAGE and TF-IGM. In *Proceedings of the International Conference on Natural Language Processing and Information Retrieval*.

[7] J. Elman. 1990. Finding structure in time. *Cognitive science* (1990).

[8] J. Hinds and A. Joinson. 2018. What demographic attributes do our digital footprints reveal? A systematic review. *PloS one* (2018).

[9] B. Huang and K. Carley. 2019. A hierarchical location prediction neural network for twitter user geolocation. *arXiv preprint arXiv:1910.12941* (2019).

[10] Taehoon K. and Kevin W. 2021. *emoji terminal output for Python.* https://github.com/carpedm20/emoji/

[11] S. Kim, Q. Xu, L. Qu, S. Wan, and C. Paris. 2017. Demographic inference on twitter using recursive neural networks. In *Annual Meeting of the Association for Computational Linguistics*.

[12] D. Kingma and J. Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[13] Guillaume L., Fernando N., and Christos A. 2017. Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. *Journal of Machine Learning Research* (2017).

[14] D Levinson. 1986. A conception of adult development. *American psychologist* (1986).

[15] Y. Liu, L. Singh, and Z. Mneimneh. 2021. A Comparative Analysis of Classic and Deep Learning Models for Inferring Gender and Age of Twitter Users. In *Proceedings of the International Conference on Deep Learning Theory and Applications*.

[16] N. Ljubešić, D. Fišer, and T. Erjavec. 2017. Language-independent gender prediction on twitter. In *Proceedings of the Workshop on NLP and Computational Social Science*.

[17] Y. Miura, M. Taniguchi, T. Taniguchi, and T. Ohkuma. 2017. Unifying text, metadata, and user network representations with a neural network for geolocation prediction. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

[18] A. Morgan-Lopez, A. Kim, R. Chew, and P. Ruddle. 2017. Predicting age groups of Twitter users based on language and metadata features. *PloS one* (2017).

[19] D. Nguyen, R. Gravel, and T. Trieschnigg, D.and Meder. 2013. How old do you think I am? A study of language and age in Twitter. In *AAAI Conference on Weblogs and Social Media*.

[20] M. Pennacchiotti and A. Popescu. 2011. A machine learning approach to twitter user classification. In *AAAI Conference on Weblogs and Social Media*.

[21] J. Pennington, R. Socher, and C. Manning. 2014. Glove: Global vectors for word representation. In *Conference on empirical methods in natural language processing*.

[22] D. Preoţiuc-Pietro and L. Ungar. 2018. User-level race and ethnicity predictors from twitter text. In *Conference on Computational Linguistics*.

[23] F.o Rangel, P. Rosso, M. Potthast, and B. Stein. 2017. Overview of the 5th author profiling task at pan 2017: Gender and language variety identification in twitter. *Working notes papers of the CLEF* (2017).

[24] D. Rao, D. Yarowsky, A. Shreevats, and M. Gupta. 2010. Classifying latent user attributes in Twitter. In *International workshop on Search and Mining User-generated Contents*.

[25] A. Reifman, M. Ursua-Benitez, S. Niehuis, E. Willis-Grossmann, and M. Thacker. 2020. # HappyAnniversary: Gender and Age Differences in Spouses' and Partners' Twitter Greetings. *Interpersona: An International Journal on Personal Relationships* (2020).

[26] N. Reimers and I. Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Conference on Empirical Methods in Natural Language Processing*.

[27] S. Rosenthal and K. McKeown. 2011. Age prediction in blogs: A study of style, content, and online behavior in pre-and post-social media generations. In *Association for Computational Linguistics: Human Language Technologies*.

[28] S. Sakaki, Y. Miura, X. Ma, K. Hattori, and T. Ohkuma. 2014. Twitter user gender inference using combined analysis of text and image processing. In *Workshop on Vision and Language*.

[29] A. Singh, E. Blanco, and W. Jin. 2019. Incorporating emoji descriptions improves tweet classification. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*.

[30] P. Vijayaraghavan, S. Vosoughi, and D. Roy. 2017. Twitter demographic classification using deep multi-modal multi-task learning. In *Annual Meeting of the Association for Computational Linguistics*.

[31] D. Vrandečić and M. Krötzsch. 2014. Wikidata: A free collaborative knowledgebase. *Commun. ACM* (2014).

[32] Stefan W. and Adam H. 2019. *Sizing Up Twitter Users.* https://www.pewresearch.org/internet/2019/04/24/sizing-up-twitter-users/

[33] Z. Wang, S. Hale, D. Adelani, P. Grabowicz, T. Hartman, F. Flock, and D. Jurgens. 2019. Demographic inference and representative population estimates from multilingual social media data. In *The World Wide Web Conference*.

[34] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the conference of the North American chapter of the association for computational linguistics*.