Multiple-Aspect Attentional Graph Neural Networks for Online Social Network User Localization

TING ZHONG, TIANLIANG WANG, JIAHAO WANG, JIN WU, AND FAN ZHOU
School of Information and Software Engineering, University of Electronic Science and Technology of China, Chengdu 610054, China
Corresponding author: Jiahao Wang (wangjh@uestc.edu.cn)

This work was supported by the National Natural Science Foundation of China under Grant 61602097 and Grant 61472064.

ABSTRACT
Identifying the geographical locations of online social media users, a.k.a. user geolocation (UG), is an essential task for many location-based applications such as advertising, social event detection, emergency localization, etc. Due to the unwillingness of revealing privacy information for most users, it is challenging to directly locate users with the ground-truth geotags. Recent efforts sidestep this limitation through retrieving users’ locations by alternatively unifying user generated contents (e.g., texts and public profiles) and online social relations. Though achieving some progress, previous methods rely on the similarity of texts and/or neighboring nodes for user geolocation, which suffers the problems of: (1) location-agnostic problem of network representation learning, which largely impedes the performance of their prediction accuracy; and (2) lack of interpretability w.r.t. the predicted results that is crucial for understanding model behavior and further improving prediction performance. To cope with such issues, we proposed a Multiple-aspect Attentional Graph Neural Networks (MAGNN) – a novel GNN model unifying the textual contents and interaction network for user geolocation prediction. The attention mechanism of MAGNN has the ability to capture multi-aspect information from multiple sources of data, which makes MAGNN inductive and easily adapt to few label scenarios. In addition, our model is able to provide meaningful explanations on the UG results, which is crucial for practical applications and subsequent decision makings. We conduct comprehensive evaluations over three real-world Twitter datasets. The experimental results verify the effectiveness of the proposed model compared to existing methods and shed lights on the interpretable user geolocation.

INDEX TERMS
Attention mechanism, user geolocation, online social network, graph neural networks.

I. INTRODUCTION
With the popularity of online social network (OSN), e.g., Twitter, Facebook, Wikipedia and Instagram, unprecedented volumes of heterogeneous data have been generated, e.g., published message contents, mention tags and follow/followee relations, which could be leveraged to geolocating OSN users. For example, people from San Francisco may frequently mention “49ers” and “Warrios” and those from New York City have high probability of tweeting contents referring to the words “Knicks” and “Yankees”. As such, the problem of user geolocation (UG) has received a lot of research attention in the past decade [1]–[7]. Successfully locating OSN users has become a key Internet service for many downstream applications, including location-based targeted advertising, emergency location identification, flu trend prediction, political election, local event/place recommendation, restricted content delivery following regional policies, natural disaster response, etc [8].

Since the social media data is unstructured, learning useful representations for both users and their generated contents becomes a key step for geolocation and downstream tasks. A plethora of works have been proposed towards structuring heterogeneous data towards better OSN user geolocation. Early efforts [1], [3], [9]–[13] mainly focus on mining indicative information from user posting contents, such as tweets and microblogs. These approaches rely on indicative words that can link users to their home locations via...
various natural language processing (NLP) techniques, e.g., topic models and statistic models. For example, TF-IDF (term frequency–inverse document frequency) [14] is a commonly used method to measure the distribution of location words [1], [15]. Besides publishing text, users are usually involved in OSNs to establish relationship and interact with friends to share work/life experience. Therefore, the locations of users can be inferred by the clues extracting from their social networks on the OSN, which has spurred a variety of network-based approaches [16]–[19]. In general, these methods leverage the user interactions, including followee relationships and mutual/unidirectional mentions, to learn users’ online proximity with various graph learning methods.

While achieving promising performance, previous works fail to tackle two main issues in user geolocation. During user content learning, the posting contents are either manually associated with indicative location words [1], [9]–[12], or are simply embedded into low-dimensional vectors using NLP techniques such as TF-IDF and doc2vec [5], [6], [20], both of which fail to capture users’ writing style especially their preference over meaningful location-related words. Furthermore, existing user geolocation methods, especially those deep learning based ones, work as a “black-box” model, fail to provide explanations regarding the model behavior and prediction results. These limitations substantially prevent previous methods from many safety-critical applications, e.g., identifying the epidemic propagation and accurate/personalized advertising.

Inspired by the recent success of graph neural networks (GNN) [21], [22] and attention mechanism [23], [24], we propose a novel GNN-based user geolocation model, called Multiple-aspect Attentional GNN (MAGNN), to address aforementioned limitations. It is a multi-view UG model that captures both linguistic and interactive information for interpretable user geolocation. The main contributions of this work can be summarized as follows:

- A novel multi-aspect GNN model for efficient user generated content and network information fusion. The proposed model exploits the interactions among tweets words and the relationship information of social networks in an end-to-end manner.
- By stacking multi-head attention layers, our model is able to distinguish different aspects of user publishing preference and the different importance of on-line interactions in a dynamic learning way rather than a fixed representation in previous work.
- We conducted extensive experiments to evaluate the proposed model on three large-scale real Twitter datasets. The experimental results demonstrate that our MAGNN model significantly improves the user location prediction accuracy compared with the state-of-the-art baselines with explainable results.

The remainder of this paper is organized as follows. We discuss the related work in Sec.II, and introduce the problem and provide the necessary background in Sec.III. The details of MAGNN will be explained in Sec.IV, followed by the extensive experimental evaluations in Sec.V. We conclude this work in Sec.VI.

II. RELATED WORK

In previous work on geolocating online social networks (OSN), the models be broadly categorized as three groups according to the type of data used to make prediction. We now review relevant works and position our paper in the literature.

A. CONTENT-BASED APPROACHES

User generated contents (UGC) such as textual posts and photos may be casually attached with real-time locations facilitated by increasing popularity of GPS-equipped devices. However, these geo-tagged tweets are extremely sparse, e.g., no more than 1% of published tweets are tagged with geographical locations [25]. A plethora of works [1], [3], [9]–[13] have studied the possibility of leveraging UGC for locating users. These methods address the geolocation problem by inferring locations from the location-relevant words with various classification models. Therefore, identifying meaningful indicative words is an important step towards accurate user geolocation, where TF-IDF (term frequency–inverse document frequency) [14] is a widely adopted textual content representation method in the literature [1], [6], [15], [26], [27]. For example, inverse location/city frequency has been used to measure the location words in the content [1], [15] while probabilistic models are used to characterize the users’ location distributions w.r.t. their published UGC, which, however, requires extensive manually labeled location-related words to achieve satisfactory results.

Inspired by recent advances in applying deep learning in natural language processing, a few studies turn to model users’ textual contents with various neural networks based models in order to learn the tweet representation in an end-to-end manner [4], [5], [20], [28]. Among these methods, doc2vec [29] and recurrent neural networks (RNNs) are simple yet effective choices for learning vector representation of textual contents. For example, Do et al. [4] combine TF-IDF and doc2vec representations of textual information to enhance the prediction performance. Miura et al. [5] use GRU [30] with attention mechanism [31] to model user tweet content and obtain a timeline representations. Though doc2vec and RNN-based methods can learn language efficiently without manual location feature engineering, a recent study [32] find that TF-IDF is consistently superior to doc2vec due to the location-indicative words captured in TF-IDF.

B. NETWORK-BASED METHODS

Online social relationships are also important indicators for user geolocation under the homophily assumption [16]–[19], i.e., people prefer to interact with others in nearby areas. Backstrom et al. [16] examine the relationship between users’ geographical proximity and online friendships on Facebook, and find that the likelihood of friendship between any user pair drops monotonically as a function of distance. Rather
than solely relying on friendships, more and more works utilize various types of connections, such as the co-mention tags and mentions between non-friends, to construct closer social interactions beyond friendships [8], [20]. In this way, similar interests among users can be retrieved from such implicit networks to improve geolocation accuracy [28], [33], [34]. Moreover, researchers also identify some noisy interaction factors that may degrade the prediction performance. For example, social influence of celebrities is a distracting factor that may confuse the prediction and thus is removed from the built user network [28], [35]. Though explicitly modeling location dependency between social connected users, some challenges of these works have not been properly addressed, e.g., sparsity of geo-tagged users and inaccurate label propagation, and, most important, the locations of friends are usually contradicting with each other, which hinder these approaches from practical applications.

C. MULTI-VIEW MODELS
Recent efforts have leveraged deep graph learning methods to model user interaction network by fusing user generated contents and various meta-data, such as user profiles, tweeting time and user timezone. For example, MENET [4] exploits node2vec [36] to learn user representations, combined with text representation learned by doc2vec, for predicting users’ locations. Another work [6] employs GCNs [21] for learning network structures with the graph convolution and pooling operations, which has achieved the state-of-the-art geolocation performance. A recent work [32] investigate several graph embedding methods and found that NetMF [37] performs better than node2vec and GraphSAGE [38] on user geolocation task, but does not show superior performance than GCN-based models [6], [32].

It is worthwhile to note that there are some works making use of various meta-data (e.g., self-declared location in profile and timezone information) for improving the prediction performance. For example, user timezone, as well as UTC offset and country noun, have been used for user geolocation [4], [5], [20], [26], [39]. While such auxiliary information is a strong indicator for regularizing the locations the model predicted, a majority of users are not willing to open these privacy information, which are sometimes camouflaged or posted casually. We further note that there is another line of efforts [7], [17], [40]–[42] studying the Twitter message geolocation problem which try to identify the tweeting locations rather than the Twitter user location discussed in this work.

TABLE 1. NOTATIONS

| Notation | Description |
|----------|-------------|
| $X_v$    | tweets content representation of all users / user v. |
| $X_v'$   | fusion representation of all users / user v. |
| $d/d'$   | dimension of $x_v / x_v'$. |
| $G/V/E$  | mention network with its nodes and edges. |
| $Y/Y'$   | label / prediction of all users. |
| $n/c$    | number of all users / clusters. |
| $e_i/b_i$ | initial / learned vector of i-th element in sequence. |
| $Q/K/K_v$ | Query / Key / Value matrix. |
| $O/O_G/O_v$ | transformation matrix of sentences / contents. |
| $d/g/d_v$ | dimension of initial / learned element vector. |
| $d_k/d_n$ | column number of Key / Value matrix. |
| $m$      | number of elements in sequence. |
| $N_v$    | the neighborhood nodes of node v in G. |
| $T$      | the number of sentences (tweets) of per user. |
| $W/R$    | number of attention heads in content / network learning. |
| $\alpha_{ij}$ | importance score of j-th to i-th element in w-th head. |
| $\beta_{vw}$ | correlations of user u and v in v-th head. |
| $W'$     | learnable matrix of r-th head in network learning. |
| $\ell$   | loss of model. |

III. PRELIMINARIES

In this section, we introduce the problem definition as well as basic notations used throughout this paper, cf. Table 1.

In this work, we consider the problem of locating the users in Twitter, where a tweet is a short text (within 140 characters) with some other content, e.g., photos and emojis. The extra information are usually associated with a tweet describing specific meanings, e.g., “@” used to mention people who are already in Twitter, and words starting with “#” are hashtags used to mention a topic.

Definition 1 (Tweet Content): For each user, we collect his/her tweets as linguistic content, including both tweet messages by himself and retweets forwarding other users’ postings. Following previous works [4]–[6], [26], we filter out the photos and symbols for each user. We denote the learned content embedding vector of user v as $x_v$, and the tweet content for all users as $X$.

In addition to posting text, we construct the mention graph to represent the social relationships among users by extracting mention (@-somebody) information from tweet messages.

Definition 2 (Mention Network): The mention network is defined as $G = (V, E)$, where $V$ is a set of all users (nodes) and $E$ is a set of edges between nodes. Each node $v \in V$ is associated with a tweet content vector $x_v$ as its feature.

We focus on predicting the “home” location of users [8], i.e., the location that a user most probably resides in. Since each user location is described by a pair of numbers (longitude and latitude), we convert this problem to the classification problem by dividing the surface of earth into closed and non overlapping clusters using k-d trees. Each user is therefore tagged with one (and only one) label indicating the cluster he/she belongs to. Each label is encoded as a one-hot vector, and we denote all labels (clusters) as $Y \in \mathbb{R}^{n \times c}$, where $n$ is the number of users, and $c$ is the number of clusters. Now, we formally define the user geolocation problem as:

Definition 3 (User Geolocation Prediction): Given all user tweet content and mention graph $G$, as well as partially labeled users, we are interested in identifying the geographical locations of unlabeled users.

IV. METHODOLOGY: MAGNN

The proposed MAGNN model is shown in Figure 1. It consists of three main components, i.e., attention-based content learning, graph neural networks based interact relation learning and geolocation predictor. First, multi-head self-attention
is utilized to learn user posting content embedding from tweets content. Secondly, the user tweet content embedding and the topological features of mention network are fused with attention mechanism on graph. After that, the fully connected layer with softmax is used to predict the locations for users.

### A. LEARNING USER CONTENT WITH MULTI-HEAD ATTENTION

To represent user generated text, TF-IDF [14] and doc2vec [29] are two widely used techniques in previous works [4], [6], [26], [32]. TF-IDF is a relative frequency approach that captures linguistic information at the word-level, while doc2vec embeds user content into a low-dimensional latent space. However, in some situation, we need to capture the meaning behind the language to achieve good performance. For example, the tweet “Too unlucky, I’ll never come to Alaska again” in Figure 1 implies that this user had negative impression on Alaska according to the emotion conveyed by her tweet. Meanwhile, it is very likely that she is just a tourist or on a business trip to Alaska, i.e., the user probably does not reside in Alaska. However, these information can’t be captured by the traditional methods such as TF-IDF and doc2vec.

In addition, tweets sent by the same user always have some irrelevant information that would be confounding factors for geolocation prediction, e.g., users usually (re)tweet some information having nothing to do with indicative location names. Thus, different tweets sent by the same user may also have different importance in representing the user gelocations. For example, the first two tweets are more informative than the last one in Figure 1.

Inspired by recent advances in natural language representation learning [23], [24], we propose to utilize multi-head self-attention to deal with user tweets content which could capture plentiful syntactic features w.r.t. user posting behaviors, and more importantly, the more significant location related information. First, we use multi-head self-attention to learn user tweet sentence embedding by paying attention to informative words and building correlation with other relevant words. Then, we apply a learnable matrix transformation to sentence embedding so as to form the user tweet content representation which will be considered as user tweet content features associated with the node in the mention network.

Specifically, we first tokenize the tweet sentence and convert the sequence of words into a sequence of low-dimensional embedding vectors, of which the $i$-th word is denoted as $e_i$. Next, the relative importance score of $j$-th word to $i$-th word under a specific attention head $w$ is computed with softmax function over the tweet sentence:

$$
\alpha^{w}_{ij} = \text{softmax} \left( \frac{(e_i \Theta^{w}_Q)(e_j \Theta^{w}_K)^T}{\sqrt{d_k}} \right),
$$

(1)

where $\Theta^{w}_Q \in \mathbb{R}^{d_e \times d_k}$ and $\Theta^{w}_K \in \mathbb{R}^{d_e \times d_k}$ are Query and Key parameter matrices [23], respectively, and $d_k$ is its column number. In addition, the softmax operation and division with the square root $d_k$ enable the score to have more stable gradients. Then, we update the $i$-th word’s representation in head $w$ through combining features of all relevant words guided by importance score $\alpha^{w}_{ij}$:

$$
\mathbf{h}^{w}_i = \sum_{j=1}^{m} \alpha^{w}_{ij}(e_j \Theta^{w}_V),
$$

(2)
where $\Theta_v^w \in \mathbb{R}^{d_v \times d_v}$ is the value parameter matrix [23] and $d_v$ is its column number, $m$ represents the length of the sentence, i.e., the number of word token in the sentence. Furthermore, the different heads are supposed to focus on different words and learn different aspects of the sentence. And the new representation of $i$-th word is calculated by combining combinatorial features learned in each head as:

$$\hat{e}_i = (h_i^1 \oplus h_i^2 \oplus \ldots \oplus h_i^w) \Theta_O,$$

where $\Theta$ represents the concatenation operator, $W$ is the number of total heads, and $\Theta_O \in \mathbb{R}^{Wd_v \times d_v}$ is the output weight matrix. With such attentional operation, the embedding of $i$-th word $e_i$ is updated into $\hat{e}_i$, which captures multi-aspects meanings guided by multi-heads attention among all words. And the final embedding of this sentence is the summation of the contextual word representations, formulated as:

$$s = \sum_{i=1}^{m} \hat{e}_i,$$

where $s \in \mathbb{R}^{1 \times d_v}$. In order to select and learn more informative information from multiple tweet sentences sent by the same user automatically, we design an additive linear transformation network to generate the tweet content representation for users. The formulation is shown as follows:

$$x = (s_1 \oplus s_2 \oplus \ldots \oplus s_i \oplus \ldots \oplus s_T) \Theta_S,$$

where $T$ is the number of tweets of each user, which is fixed in our datasets. $\Theta_S \in \mathbb{R}^{T \times d_v \times d_v}$ is the learnable matrix transforming the multiple sentence embeddings into a single vector. The tweet content representations of all users are denoted by $X (X \in \mathbb{R}^{n \times d_v})$, which is the input features of networks learning in MAGNN.

### B. MULTI-ASPECT INFORMATION FUSION USING GNN

GNN is a powerful tool for graph representation learning, which has received increasing attention over the past years [21], [22], [43], [44]. A GNN model consists of a stack of neural network layers, where each layer aggregates neighborhood information around each node and then passes the aggregated message into the next layer. Given a network $G = (V, \mathcal{E})$ and the initial features $x_v$ of corresponding node $v$, a general GNN architecture updating the node representation in $k$-th ($k > 0$) layer can be implemented as [38]:

$$x_v^{(k)} = f_{\text{merge}}^{(k)}(x_v^{(k-1)}, f_{\text{agg}}^{(k)}(\{x_u^{(k-1)} | u \in N_v\}),$$

where $\theta_1$ and $\theta_2$ are trainable parameters optimized via stochastic gradient descent, and $N_v$ represents the neighbors of node $v$. $f_{\text{agg}}^{(k)}$ aggregates the features from neighbors with various operations (e.g., Mean and Pooling) while $f_{\text{merge}}^{(k)}$ merges node’s representations from the $k-1$ step and the aggregated features of neighbors. The learned node embeddings can be used for downstream tasks such as link prediction, node/graph classification.

There are many variants of GNNs which are used to deal with graph structure data, cf. [43], [44] for comprehensive reviews. For example, GAT [22] introduces attention mechanism in the process of GNN learning, where a node involves most relevant information from its neighborhoods and update its own features with the learned attention weights, which enables the model to focus on the most informative features while alleviating noise signals during message passing in the network. Here we extend GAT with multi-head attention to learn the structural representation while propagating the content features in network.

Specifically, we first compute the relevant coefficients between a pair of nodes with multi-head attention. Namely, the correlations between node $u$ and node $v$ in the $r$-th head can be calculated as ($r > 0$):

$$e_{uv}^r = \sigma([x_u W_v || x_u W_v] a^T),$$

where $W_v \in \mathbb{R}^{d_v \times d_v'}$ is a linear transformation matrix of the $r$-th head which maps input features into high-level representations, and $||$ denotes the concatenation operation. Here we use a feedforward neural network with parameters $a \in \mathbb{R}^{1 \times 2d_v'}$ as the attention layer and $\sigma(\cdot)$ as the non-linear activation function (LeakyReLU(\cdot) in our implementation). In order to make the correlation computation stable, softmax is applied over all nodes in $N_v$:

$$\beta_{uv} = \frac{\exp(e_{uv}^r)}{\sum_{q \in N_v} \exp(e_{qv}^r)},$$

where the coefficient $\beta_{uv}$ is expected to capture the most relevant features while dynamically filtering out the useless features for node $v$. Subsequently, linear combination is used to fuse the neighboring features with the built coefficient in $r$-th head as:

$$f_v^r = \sum_{u \in N_v} \beta_{uv} x_u.$$

Next, we calculate the new representation of each node by averaging the features of all multi-heads through a non-linear transformation:

$$x_v^r = \text{LeakyReLU}(\frac{1}{R} \sum_{r=1}^{R} f_v^r W_r),$$

where $R$ is the number of attention heads and $W_r$ is the linear transformation weight matrix of $r$-th head. The new representation of all users is represented as $X' (X' \in \mathbb{R}^{n \times d_v'})$ which will be fed into the geolocation predictor for predicting the final results. Note that we mask the labels of validation and testing samples during training, i.e., the labels of the data in validation and testing set are invisible when learning user representation.

### C. GEOLOCATION PREDICTOR

The objective of geolocation predictor is to predict the highest probability of a location the user belongs to. Here, we adopt a multilayer perceptron layer (MLP) to make the predictions based on the learned user representations $X'$:

$$Y' = \text{softmax}(\text{MLP}(X')),$$
where \( Y' \in \mathbb{R}^{n \times c} \) is predictions for all users. Here we adopt cross entropy as the loss function:
\[
\ell = - \sum_{i=1}^{n} \sum_{j=1}^{c} y_{ij} \log y'_{ij},
\]
where \( y_{ij} \) denotes the probability that the \( i \)-th user belongs to the \( j \)-th cluster. During training, Adam [45] is adopted as the stochastic gradient descent optimizer.

**Algorithm 1 Training Algorithm of MAGNN**

**Input:** Tweet content, Mention network \( G = (\mathcal{V}, \mathcal{E}) \).

**Output:** Trained model.

/* Content Learning */
1. for user id \( i = 1 \) to \( n \) do
2.   for sentence id \( j = 1 \) to \( T \) do
3.     Initialize \( j \)-th sentence tokens’ embeddings;
4.     foreach token do
5.       Calculate new token embedding \( \hat{e} \) via Eq. (1), (2) and (3);
6.     end
7.     Compute the \( j \)-th sentence embedding \( s_j \) via Eq. (4);
8.   end
9.   Compute the \( i \)-th user content representation \( x_i \) via Eq. (5);
10. end
11. Concatenate content representation of all users into matrix \( X \);

/* Network Learning */
12. foreach user \( v \in \mathcal{V} \) do
13.   Get \( v \)'s neighborhoods \( \mathcal{N}_v \) in \( G \);
14.   for head id \( r = 1 \) to \( R \) do
15.     Compute attention scores \( \beta_{vw}^r \) among \( \mathcal{N}_v \) via Eq. (7) and (8);
16.     Compute \( v \)'s representation \( \mathbf{f}_v^r \) with \( \beta_{vw}^r \) via Eq. (9);
17.   end
18.   Compute fusion representation \( \mathbf{x}_v' \) of user \( v \) via Eq. (10);
19. end
20. Concatenate fusion representation of all users into matrix \( X' \);

/* Geolocation Predictor */
21. Compute predictions \( Y' \) with \( X' \) via Eq. (11);
22. Calculate loss with Eq. (12);
23. Update model parameters with Adam.

**D. COMPLEXITY**

Algorithm 1 outlines the training procedure of MAGNN. In order to simplify the formula, we assume \( d_e = d_k = d_v = \hat{d}_e = d \) for content learning. Thus, the complexity of single self-attention is \( O(m^2d) \), where \( m \) is the sequence length. Since multiple heads can be computed in parallel, the complexity of calculating the content presentation for one user only introduces the extra cost \( O(md^2) \) in Eq. (3) and \( O(md^2) \) in Eq. (5). Therefore, the complexity for content learning is \( O(n(m^2d + md^2)) \), where \( n \) is the number of all users. As for network learning, the operation of calculating attention score (cf. Eq. (7) and Eq. (8)) and output features (i.e., \( \mathbf{f}_v^r \)) of each head can be paralleled across all nodes and the time complexity of attentional GNN learning with one attention head is \( O(nmd' + |E|d') \), where \( |E| \) is the number of edges in the mention network. By contrast, the time complexity of GCN4Geo is \( O(LA_0 F + LNF^2) \), where \( L \) is the number of layers, \( N \) is the number of users, \( A_0 \) represents the number of non-zeros in the adjacency matrix of mention network and \( F \) is the dimension of features.

**V. EXPERIMENTS**

In this section, we conduct experiments on three real-world datasets to evaluate our model against baselines. Specifically, we aim to answer the following research questions:

- **Q1.** How does MAGNN perform compared to the state-of-the-art approaches/baselines on geolocation prediction?
- **Q2.** How does each component (context embedding and network embedding) contributes to the performance of MAGNN?
- **Q3.** How do the key hyper-parameters affect MAGNN’s performance?
- **Q4.** How can we explain the prediction results made by our MAGNN?

**A. DATASET**

To evaluate the performance of our model, we conduct experiments on three real-world Twitter datasets which have been widely used for evaluating the user geolocation models. The datasets are listed below and their statistics are summarized in Table 2.

- **GeoText** [46] is a Twitter dataset consisting of 9.5K users from 49 states and Washington D.C. in U.S., which is originally compiled by the authors in [46]. The dataset has already been divided into the training, development and testing set with 5,685, 1,895 and 1,895 users, respectively.
- **Twitter-US** [11] is a larger dataset consisting of 449K users from the U.S., which was created by the authors in [11]. This dataset is also referred to as UTGeo2011 in some papers [4], [11]. Following previous works, 10K users are held out for validation and 10K users left for testing.

**TABLE 2. Statistics of datasets.**

| Dataset       | # Nodes | # Edges | # Clusters |
|---------------|---------|---------|------------|
| GeoText       | 9,475   | 77,155  | 129        |
| Twitter-US    | 449,200 | 18,498,702 | 256        |
| Twitter-World | 1,386,766 | 1,001,181 | 930        |
### B. METRICS

We evaluate all approaches using the following three metrics that are commonly used for user geolocation performance evaluation:

1. **Mean** prediction error, measured in kilometres, gives the averaged error between the predicted cluster centers and the ground-truth geolocations for all testing samples.
2. **Median** prediction error reports the median value of the predicted errors for all testing samples.
3. **Acc@161** measures the accuracy of the classification. Namely, if the distance between the predicted cluster center and ground-truth is within 161 km (or 100 miles), the result will be considered as a correct prediction.

Note that the distance of coordinates is computed using the Haversine formula. The lower values of Mean and Median error indicate a better prediction. Conversely, achieving higher value of Acc@161 is desirable.

### C. BASELINES

We compare MAGNN with the following user geolocation models:

**Text-based:**

- **HierLR** [2] is a text-based geolocation model, which adopts a grid representation of locations and resorts to hierarchical classification using logistic regression (LR).
- **MLP4Geo** [20] is a text-based model which uses dialectal terms to improve the prediction performance. A simple MLP network is used to predict the locations.
- **DocSim** [11] uses a method of matching the similarity (KL divergence) of the subject document for prediction.
- **LocWords** [1] is a text-based model which uses several methods to find the location indicative words (LIWs) for prediction.
- **MixNet** [27] is a text-based model which applies mixture density network (MDN) for embedding coordinates in a continuous vector space with shared parameters.

**Network-based:**

- **MADCEL** [26] is a network-based model, which applies Modified Adsorption with celebrity removal. Only the results of weighted network will be reported since it performs better than binary network.
- **GCN-LP** [6] is a GCN-based model and it is similar to label propagation. It performs the convolution operation on network for prediction and user’ features are represented by one-hot encoding of its neighbours.

**Multiview-based:**

- **MADCEL-LR** [26] combines the text and network information and uses LR for location prediction.
- **MENET** [4] concatenates the features from textual information (tf-idf [49], doc2vec [50]), user interaction network (node2vec [36]) and metadata (timestamp) and use fully connected networks for location prediction. For a fair comparison, we only use text and network information in MENET.
- **GeoAtt** [5] models the textual context with RNN and attention mechanism. We remove the location descriptions in GeoAtt for fair comparisons.
TABLE 3. Performance comparison among different algorithms on three datasets.

| Type   | Method  | GeoText |  | Twitter-US |  | Twitter-World |
|--------|---------|---------|  |           |  |              |
|        |         | Mean    | Median | Acc@161 | Mean    | Median | Acc@161 | Mean    | Median | Acc@161 |
| Text   | HerLR   | 834     | 403   | 41      | 703     | 170    | 49      | 1714    | 490    | 32      |
|        | MLP4Geo | 844     | 389   | 38      | 554     | 120    | 54      | 1456    | 415    | 34      |
|        | DocSim  | 897     | 432   | 35      | 860     | 463    | 34      | 1839    | 530    | 30      |
|        | LocWords| –       | –     | –       | 814     | 260    | 45      | 3002    | 926    | 25      |
|        | MixNet  | 865     | 412   | 39      | 655     | 216    | 42      | 1805    | 512    | 31      |
| Network| MADCEL  | 586     | 60    | 58      | 705     | 116    | 54      | 2525    | 279    | 45      |
|        | GCN-LP  | 576     | 56    | 58      | 653     | 126    | 53      | 2357    | 279    | 45      |
| MultiView| MADCEL-LP | 581   | 57    | 59      | 529     | 78     | 60      | 1403    | 111    | 53      |
|        | MENET   | 643     | 125   | 55      | 526     | 93     | 52      | 1290    | 126    | 52      |
|        | GeoAtt  | 612     | 81    | 57      | 545     | 91     | 54      | 1263    | 214    | 50      |
|        | DCCA    | 627     | 79    | 56      | 516     | 90     | 58      | 2095    | 913    | 21      |
|        | GCN4Geo | 546     | 45    | 60      | 485     | 71     | 62      | 1130    | 108    | 54      |
|        | KB-emb  | 780     | 339   | 43      | 601     | 197    | 45      | 1409    | 372    | 35      |
|        | GausMix | 593     | 69    | 56      | 540     | 89     | 56      | 1358    | 200    | 51      |

D. Q1: OVERALL PERFORMANCE COMPARISON

The overall performance of all methods across three datasets are presented in Table 3, from which we have following major observations.

First, relying only on tweet content [2], [20] is not enough for user geolocation prediction, which usually exhibits extremely high prediction bias. This result is intuitive since neither indicative words [1], [11] nor topic-based language models [27], [46] can filter out noisy signals from user tweeting content. For example, users usually publish short texts containing acronyms and misspellings which are difficult to be identified. Moreover, estimating spatial word distribution often confronts sparsity problem, i.e., some location words w.r.t. less populated locations are unobserved during training, which further obfuscate the geolocation models. On the other hand, user interaction network plays a key role in predicting the home locations. However, we also cannot rely only on user networks for accurate user geolocation. This is mainly because many accounts use Twitter, as well as other OSN platforms such as Facebook and Instagram, for the purpose of propagating information like advertising and commercial usage – e.g., there are many official accounts associating with various companies and NGOs. Also, a large number of personal accounts use Twitter for information dissemination and knowledge sharing instead of building social relationships. In both cases, homophily assumption is not held anymore.

Second, the performance of deep learning-based multi-view models, including MENET, GeoAtt, DCCA and GCN4Geo, is very similar if both text and network features are used. Surprisingly, their performance are very close to the models using simple classification methods, e.g., the LR in MADCEL. This result implicates that meaningful features are more important than complicated models in the user geolocation prediction task. This observation can be further proven in previous work [4], [5] that incorporates more strong indicators such as timezone of users and description in the location field – however, improving MAGNN with more features, e.g., self-declared locations and timezone, is beyond the scope of this work and is left for our future work. Furthermore, previous multi-view models fail to improve the model performance largely due to the ignorance of node importance when modeling the user interaction network. For example, MENET uses node2vec to embed the network while GCN4Geo directly leverages GCNs for modeling the user interactions. However, both node2vec and GCNs do not discriminate the relative influence of nodes when aggregating the local structural information. For example, if two users are topologically the same, they would be located to the same region (without considering their tweet content) even they are residing in geographically different locations.

Third, our MAGNN consistently outperforms the baselines on all metrics, which proves the effect of addressing the user geolocation problem with the proposed multi-head attention based neural networks. This is mainly because the multi-heads attention could capture multiple aspects meaning for dynamic features aggregation, while filtering out the noise of content and structure information to reduce prediction bias. Compared to RNN-based attention models in [5],
MAGNN can capture long-range dependencies in textual information and adaptively adjust interaction learning. Finally, the Macro-Recall and Macro-F1 results of MAGNN and GCN4Geo are shown in Figure 2—we omitted other methods because GCN4Geo usually performs best among the baselines. Clearly, MAGNN outperforms GCN4Geo slightly due to its ability of distinguishing the importance of neighboring nodes when aggregating features from social friends. We note that the number of training samples in different clusters are extremely imbalanced, e.g., people generally live in densely popular cities (e.g., NY City and Los Angles in GeoText Data) while only few users live in the rural areas. Therefore, how to address the class imbalance issue inherent in user geolocation is a challenging problem requiring further examinations, which is left as our future work.

E. Q2: ABLATION STUDY
To investigate the effect of different components in our model, we implement two variants of MAGNN, including: (1) MAGNN-content, which only utilizes user content features for prediction; and (2) MAGNN-network which only relies on interaction network for user geolocation. The performance of two variants, as well as MAGNN, is shown in Table 4. The result suggests that network information plays more important role than content features, which is also observed in recent experimental comparisons [32]. It also points out the promising ways of improving geolocation performance in future studies, i.e., focusing more on users' interactions rather than their publishing contents. Another potential way of further improving MAGNN is to explore more auxiliary features of user (e.g., user profile, timezone) which have been proven in previous work [4] to be strong indicators for better regularizing the geolocation results.

F. Q3: PARAMETER SENSITIVITY
As the multi-head attention mechanism is used in our model, different heads are supposed to capture multi-aspect features and make our model more stable. In this section, we analyze the performance of MAGNN w.r.t. the head number. Since network plays more important role, we fixed an optimal number of heads in content learning and investigate the performance of MAGNN by varying the head number \( R \) in network learning. In particular, we use 8 attention heads to embed the content of each user into a 512-dimensional vector, and then investigate the influence of the number of attention heads \( R \). Tables 5 shows the results carried on GeoText and Twitter-US. On GeoText, the more heads the better performance of our model, when \( R \leq 32 \). This result suggests that increasing \( R \) is a direct optimization method for a smaller dataset. However, further increasing \( R \) (e.g., greater than 32) does not imply higher performance, which means one should carefully tune this hyperparameters to balance the effectiveness vs. efficiency – the computational cost surges considerably with the value of \( R \). This can be further proven by the results on Twitter-US, a significantly larger dataset, where a smaller value of \( R \) is enough for our model to achieve best results.

G. Q4: QUALITATIVE ANALYSIS
In this section, we provide qualitative interpretation of the results made by MAGNN from the latent space – the learned latent space which reflects how expressive and distinct representation our model can learn.

In this section, we provide qualitative interpretation of the results made by MAGNN. We randomly select four cluster from GeoText and their corresponding users and use t-SNE [51] to map the learned latent representation into 2D space. Figure 3 illustrates the results of MLP4Geo and our MAGNN, from which we can easily observe the clustering effect in the latent space learned by MAGNN. MLP4Geo, in contrast, is a plain model which simply concatenates content features \( X \) and network adjacent matrix \( A \) and then feed them to MLPs for geolocation prediction. Therefore, it is difficult for this simple model to capture non-linear interactions among samples from different classes, and, more importantly, to discriminate the users using uniformly scattered representations. This result also explains the performance gain made by our GNN based model which aggregates important signals.
aggressively and can effectively group the users from same regions together in the latent space.

H. DISCUSSION

From the empirical observations on three real-world datasets, the proposed method MAGNN is able to estimate the geolocation of Twitter users with higher accuracy than previous methods. MAGNN achieves superior performance due to its ability of effectively fusing content features and network features within the attentive graph neural networks architectures. This also demonstrates the power of the proposed method on mining hidden features regarding Twitter content and user mention network, while filtering out noisy signals that have been ignored in previous methods. Nevertheless, it is worthwhile to note that the multi-head attention used in MAGNN may require more memory cost, especially when the number of attention heads increases, which restricts the application of our model in resource-limited setting. One of the promising ways of improving memory efficiency is to replace multi-head attention in MAGNN with multi-linear attention mechanism as suggested in [52]. However, this is beyond the scope of this work and left as our future work.

VI. CONCLUSION REMARKS

In this work, we presented a new social user geolocation framework, which is built upon the twitter content and user social network, without requiring any explicit user profile information. With the proposed graph neural networks with multi-head attention mechanism, our model can filter out the noise from the content information and confounding user contacts, so as to focus on the most important information, both linguistic and structural, to alleviate the problem of inference bias when geolocating the users. Extensive experiments have been conducted on large scale datasets which demonstrate the superior performance of our model against previous state-of-the-art UG methods. We also provide interpretable results regarding our model and its performance. One of our immediate future work is to further improve the UG performance by exploiting multi-aspect features, such as profile and timezone. In addition, how to better distill spatio-temporal knowledge and geographical semantics from user published content – in addition to indicative words – is another topic of our ongoing work.

REFERENCES

[1] B. Han, P. Cook, and T. Baldwin, “Geolocation prediction in social media data by finding location indicative words,” in Proc. Int. Conf. Comput. Linguistics (COLING), 2012, pp. 1045–1062.

[2] B. Wing and J. Baldridge, “Hierarchical discriminative classification for text-based geolocation,” in Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP), 2014, pp. 336–348.

[3] B. Han, P. Cook, and T. Baldwin, “Text-based Twitter user geolocation prediction,” J. Artif. Intell. Res., vol. 49, pp. 451–500, Mar. 2014.

[4] T. Huu Do, D. Minh Nguyen, E. Tsigliani, B. Cornelis, and N. Deligiannis, “Multiview deep learning for predicting Twitter Users’ location,” 2017, arXiv:1712.08091. [Online]. Available: http://arxiv.org/abs/1712.08091

[5] Y. Miura, M. Taniguchi, T. Taniguchi, and T. Ohkuma, “Unifying text, metadata, and user network representations with a neural network for geolocation prediction,” in Proc. Annu. Meeting Assoc. Comput. Linguistics (ACL), 2017, pp. 1260–1272.

[6] A. Rahimi, T. Cohn, and T. Baldwin, “Semi-supervised user geolocation via graph convolutional networks,” in Proc. Annu. Meeting Assoc. Comput. Linguistics (ACL), 2018, pp. 2009–2019.

[7] J. Bakerman, K. Pazdernik, A. Wilson, G. Fairchild, and R. Bahram, “Twitter geolocation: A hybrid approach,” ACM Trans. Knowl. Discovery from Data, vol. 12, no. 3, p. 34, Apr. 2018.

[8] X. Zheng, J. Han, and A. Sun, “A survey of location prediction on Twitter,” IEEE Trans. Knowl. Data Eng., vol. 30, no. 9, pp. 1652–1671, Sep. 2018.

[9] E. Amitay, N. Har El, R. Sivan, and A. Soffer, “Web-a-where: Geotagging Web content,” in Proc. Int. Conf. Res. Develop. Inf. Retr. (SIGIR), 2004, pp. 273–280.

[10] B. P. Wing and J. Baldridge, “Simple supervised document geolocation with geodesic grids,” in Proc. Annu. Meeting Assoc. Comput. Linguistics (ACL), 2011, pp. 955–964.

[11] S. Roller, M. Speriosu, S. Rallapalli, B. Wing, and J. Baldridge, “Supervised text-based geolocation using language models on an adaptive grid,” in Proc. Annu. Meeting Assoc. Comput. Linguistics (ACL), 2012, pp. 1500–1510.

[12] A. Ahmed, L. Hong, and A. J. Smola, “Hierarchical geographical modeling of user locations from social media posts,” in Proc. 22nd Int. Conf. World Wide Web (WWW), 2013, pp. 25–36.

[13] W.-H. Chong and E.-P. Lim, “Tweet geolocation: Leveraging location, user and peer signals,” in Proc. ACM Conf. Inf. Knowl. Manage. CIKM, Nov. 2017, pp. 1279–1288.

[14] K. S. Jones, “A statistical interpretation of term specificity and its application in retrieval,” J. Document., vol. 28, no. 1, pp. 11–21, Jan. 1972.

[15] K. Ren, S. Zhang, and H. Lin, “Where are you settling down: Geo-locating Twitter users based on tweets and social networks,” in Proc. Asia Inf. Retr. Symp. Springer, 2012, pp. 150–161.
T. Zhong et al.: Multiple-Aspect Attentional Graph Neural Networks for Online Social Network User Localization

[16] L. Backstrom, E. Sun, and C. Marlow, “Find me if you can: Improving geographical prediction with social and spatial proximity,” in Proc. 19th Int. Conf. World Wide Web (WWW), 2010, pp. 61–70.

[17] C. A. Davis, Jr., G. L. Pappa, D. R. R. de Oliveira, and F. de L. Arcanjo, “Inferring the location of Twitter messages based on user relationships,” Trans. GIS, vol. 15, no. 6, pp. 735–751, Dec. 2011.

[18] L. Kong, Z. Liu, and Y. Huang, “SPOT: Locating social media users based on social network context,” Proc. VLDB Endowment, vol. 7, no. 13, pp. 1681–1684, Aug. 2014.

[19] E. Rodrigues, R. Assunção, G. L. Pappa, D. Renno, and W. Meira, Jr., “Exploring multiple evidence to infer users’ location in Twitter,” Neurocomputing, vol. 171, pp. 30–38, Jan. 2016.

[20] A. Rahimi, T. Cohn, and T. Baldwin, “A neural model for user geolocation and lexical dialectology,” in Proc. 55th Annu. Meeting Assoc. Comput. Linguistics, 2017, pp. 209–216.

[21] T. N. Kipf and M. Welling, “Semi-supervised classification with graph convolutional networks,” in Proc. Int. Conf. Learn. Represent. (ICLR), 2017.

[22] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, “Graph attention networks,” in Proc. Int. Conf. Learn. Represent. (ICLR), 2018.

[23] A. Vasznavi, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” in Proc. Adv. Neural Inf. Process. Syst. (NIPS), 2017, pp. 5998–6008.

[24] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in Proc. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol. (NAACL-HLT), 2019, pp. 4171–4186.

[25] Z. Cheng, J. Caverlee, and K. Lee, “You are where you tweet: A content-based approach to geo-locating Twitter users,” in Proc. 19th ACM Int. Conf. Inf. Knowl. Manage. (CIKM), 2010, pp. 759–768.

[26] A. Rahimi, T. Cohn, and T. Baldwin, “Twitter user geolocation using a unified text and network prediction model,” in Proc. Annu. Meeting Assoc. Comput. Linguistics (ACL), 2015, pp. 630–636.

[27] A. Rahimi, T. Baldwin, and T. Cohn, “Continuous representation of location for geolocation and lexical dialectology using mixture density networks,” in Proc. Conf. Empirical Methods Natural Lang. Process., 2017, pp. 167–176.

[28] A. Rahimi, D. Vu, T. Cohn, and T. Baldwin, “Exploiting text and network context for geolocation of social media users,” in Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics Hum. Lang. Technol., 2015, pp. 1362–1367.

[29] Q. Le and T. Mikolov, “Distributed representations of sentences and documents,” in Proc. Int. Conf. Mach. Learn. (ICML), 2014, pp. 1188–1196.

[30] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, “Empirical evaluation of gated recurrent neural networks on sequence modeling,” 2014, arXiv:1412.3555. [Online]. Available: http://arxiv.org/abs/1412.3555

[31] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” in Proc. Int. Conf. Learn. Represent. (ICLR), 2015.

[32] P. Hamouni, T. Khazaiei, and E. Arjadian, “TF-MF: Improving multi-view representation for Twitter user geolocation prediction,” in Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining (ASONAM), Aug. 2019, pp. 543–545.

[33] J. McGee, J. Caverlee, and Z. Cheng, “Location prediction in social media based on tie strength,” in Proc. ACM Int. Conf. Inf. Knowl. Manage. (CIKM), 2013, pp. 459–468.

[34] D. Jurgens, “That’s what friends are for: Inferring location in online social media platforms based on social relationships,” in Proc. Int. AAAI Conf. Weblogs Social Media, 2013.

[35] R. Li, S. Wang, H. Deng, R. Wang, and K. C.-C. Chang, “Towards social user profiling: Unified and discriminative influence model for inferring home locations,” in Proc. SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD), 2012, pp. 1023–1031.

[36] A. Grover and J. Leskovec, “node2vec: Scalable feature learning for networks,” in Proc. 22nd ACM SIGKDD Conf. Knowl. Discovery Data Mining (KDD), Aug. 2016, pp. 855–864.

[37] J. Qiu, Y. Dong, H. Ma, J. Li, K. Wang, and J. Tang, “Network embedding as matrix factorization: Unifying DeepWalk, LINE, PTE, and node2vec,” in Proc. 11th ACM Int. Conf. Web Search Data Mining (WSDM), 2018, pp. 459–467.

[38] W. Hamilton, Z. Ying, and J. Leskovec, “Inductive representation learning on large graphs,” in Proc. Adv. Neural Inf. Process. Syst. (NIPS), 2017, pp. 1024–1034.
JIAHAO WANG received the M.S. and Ph.D. degrees from the University of Electronic Science and Technology of China, in 2004 and 2007, respectively. He was a Postdoctoral Research Associate with The Hong Kong University, from 2007 to 2009. He was a Visiting Scholar with the Illinois Institute of Technology (IIT), USA, from 2013 to 2014. He is currently an Associate Professor with the School of Information and Software Engineering, University of Electronic Science and Technology of China. His research interests include the IoT data mining, the IoT security, and deep learning for IoT.

JIN WU received the B.S. degree in automatic control and the M.S. and Ph.D. degrees from the University of Electronic Science and Technology of China, in 1993, 1996, and 2004, respectively. She is currently an Associate Professor with the University of Electronic Science and Technology of China. Her research interests include machine learning, knowledge mapping, software development techniques, and process technology.

FAN ZHOU received the B.S. degree in computer science from Sichuan University, China, in 2003, and the M.S. and Ph.D. degrees from the University of Electronic Science and Technology of China, in 2006 and 2011, respectively. He is currently an Associate Professor with the School of Information and Software Engineering, University of Electronic Science and Technology of China. His research interests include machine learning, spatio-temporal data management, and social network knowledge discovery.