Unifying Text, Metadata, and User Network Representations with a Neural Network for Geolocation Prediction

Yasuhide Miura[†,‡]
yasuhide.miura@fujixerox.co.jp

Motoki Taniguchi[†]
motoki.taniguchi@fujixerox.co.jp

Tomoki Taniguchi[†]
taniguchi.tomoki@fujixerox.co.jp

Tomoko Ohkuma[†]
ohkuma.tomoko@fujixerox.co.jp

[†]Fuji Xerox Co., Ltd.
[‡]Tokyo Institute of Technology

Abstract

We propose a novel geolocation prediction model using a complex neural network. Our model unifies text, metadata, and user network representations with an attention mechanism to overcome previous ensemble approaches. In an evaluation using two open datasets, the proposed model exhibited a maximum 3.8% increase in accuracy and a maximum of 6.6% increase in accuracy@161 against previous models. We further analyzed several intermediate layers of our model, which revealed that their states capture some statistical characteristics of the datasets.

1 Introduction

Social media sites have become a popular source of information to analyze current opinions of numerous people. Many researchers have worked to realize various automated analytical methods for social media because manual analysis of such vast amounts of data is difficult. Geolocation prediction is one such analytical method that has been studied widely to predict a user location or a document location. Location information is crucially important information for analyses such as disaster analysis (Sakaki et al., 2010), disease analysis (Culotta, 2010), and political analysis (Tumasjan et al., 2010). Such information is also useful for analyses such as sentiment analysis (Martínez-Cámara et al., 2014) and user attribute analysis (Rao et al., 2010) to undertake detailed region-specific analyses.

Geolocation prediction has been performed for Wikipedia (Overell, 2009), Flickr (Serdyukov et al., 2009; Crandall et al., 2009), Facebook (Backstrom et al., 2010), and Twitter (Cheng et al., 2010; Eisenstein et al., 2010). Among these sources, Twitter is often preferred because of its characteristics, which are suited for geolocation prediction. First, some tweets include geotags, which are useful as ground truth locations. Secondly, tweets include *metadata* such as timezones and self-declared locations that can facilitate geolocation prediction. Thirdly, a user network is obtainable by consideration of the interaction between two users as a network link.

Herein, we propose a neural network model to tackle geolocation prediction in Twitter. Past studies have combined text, metadata, and user network information with ensemble approaches (Han et al., 2013, 2014; Rahimi et al., 2015a; Jayasinghe et al., 2016) to achieve state-of-the-art performance. Our model combines text, metadata, and user network information using a complex neural network. Neural networks have recently shown effectiveness to capture complex representations combining simpler representations from large-scale datasets (Goodfellow et al., 2016). We intend to obtain unified text, metadata, and user network representations with an attention mechanism (Bahdanau et al., 2014) that is superior to the earlier ensemble approaches. The contributions of this paper are the following:

1. We propose a neural network model that learns unified text, metadata, and user network representations with an attention mechanism.
2. We show that the proposed model outperforms the previous ensemble approaches in two open datasets.
3. We analyze some components of the proposed model to gain insight into the unification processes of the model.

Our model specifically emphasizes geolocation prediction in Twitter to use benefits derived from the characteristics described above. However, our

Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, pages 1260–1272
Vancouver, Canada, July 30 - August 4, 2017.
https://doi.org/10.18653/v1/P17-1116

model can be readily extended to other social media analyses such as user attribute analysis and political analysis, which can benefit from metadata and user network information.

In subsequent sections of this paper, we explain the related works in four perspectives in Section 2. The proposed neural network model is described in Section 3 along with two open datasets that we used for evaluations in Section 4. Details of an evaluation are reported in Section 5 with discussions in Section 6. Finally, Section 7 concludes the paper with some future directions.

2 Related Works

2.1 Text-based Approach

Probability distributions of words over locations have been used to estimate the geolocations of users. Maximum likelihood estimation approaches (Cheng et al., 2010, 2013) and language modeling approaches minimizing KL-divergence (Wing and Baldridge, 2011; Kinsella et al., 2011; Roller et al., 2012) have succeeded in predicting user locations using word distributions. Topic modeling approaches to extract latent topics with geographical regions (Eisenstein et al., 2010, 2011; Hong et al., 2012; Ahmed et al., 2013) have also been explored considering word distributions.

Supervised machine learning methods with word features are also popular in text-based geolocation prediction. Multinomial Naive Bayes (Han et al., 2012, 2014; Wing and Baldridge, 2011), logistic regression (Wing and Baldridge, 2014; Han et al., 2014), hierarchical logistic regression (Wing and Baldridge, 2014), and a multilayer neural network with stacked denoising autoencoder (Liu and Inkpen, 2015) have realized geolocation prediction from text. A semi-supervised machine learning approach by Cha et al. (2015) has also been produced using a sparse-coding and dictionary learning.

2.2 User-network-based Approach

Social media often include interactions of several kinds among users. These interactions can be regarded as links that form a network among users. Several studies have used such user network information to predict geolocation. Backstrom et al. (2010) introduced a probabilistic model to predict the location of a user using friendship information in Facebook. Friend and follower information in Twitter were used to predict user locations with a most frequent friend algorithm (Davis Jr. et al., 2011), a unified descriptive model (Li et al., 2012b), location-based generative models (Li et al., 2012a), dynamic Bayesian networks (Sadilek et al., 2012), a support vector machine (Rout et al., 2013), and maximum likelihood estimation (McGee et al., 2013). Mention information in Twitter is also used with label propagation models (Jurgens, 2013; Compton et al., 2014) and an energy and social local coefficient model (Kong et al., 2014). Jurgens et al. (2015) compared nine user-network-based approaches targeting Twitter, controlling data conditions.

2.3 Metadata-based Approach

Metadata such as location fields are useful as effective clues to predict geolocation. Hecht et al. (2011) reported that decent accuracy of geolocation prediction can be achieved using location fields. Approaches to combine metadata with texts are also proposed to extend text-based approaches. Combinatory approaches such as a dynamically weighted ensemble method (Mahmud et al., 2012), polygon stacking (Schulz et al., 2013), stacking (Han et al., 2013, 2014), and average pooling with a neural network (Miura et al., 2016) have strengthened geolocation prediction.

2.4 Combinatory Approach Extending User-network-based Approach

Several attempts have been made to combine user-network-based approaches with other approaches. A text-based approach with logistic regression was combined with label propagation approaches to enhance geolocation prediction (Rahimi et al., 2015a,b, 2016). Jayasinghe et al. (2016) combined nine components including text-based approaches, metadata-based approaches, and a user-network-based approach with a cascade ensemble method.

2.5 Comparisons with Proposed Model

A model we propose in Section 3 which combines text, metadata, and user network information with a neural network, can be regarded as an alternative to approaches using text and metadata (Mahmud et al., 2012; Schulz et al., 2013; Han et al., 2013, 2014; Miura et al., 2016), approaches with text and user network information (Rahimi et al., 2015a,b), and an approach with text, metadata, and user network information (Jayasinghe et al., 2016). In Section 5, we demonstrate that our model outperforms earlier models.

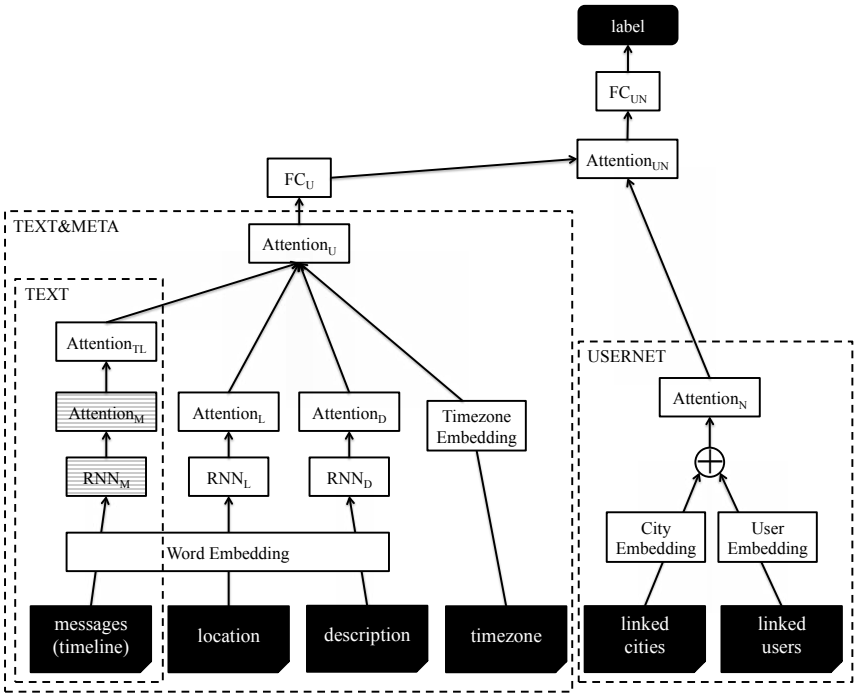


Figure 1: Overview of the proposed model. RNN denotes a recurrent neural network layer. FC denotes a fully connected layer. The striped layers are message-level processes. $\oplus$ represents element-wise addition.

In terms of machine learning methods, our model is a neural network model that shares some similarity with previous neural network models (Liu and Inkpen, 2015; Miura et al., 2016). Our model and these previous models have two key differences. First, our model integrates user network information along with other information. Secondly, our model combines text and metadata with an attention mechanism (Bahdanau et al., 2014).

3 Model

3.1 Proposed Model

Figure 1 presents an overview of our model: a complex neural network for classification with a city as a label. For each user, the model accepts inputs of messages, a location field, a description field, a timezone, linked users, and the cities of linked users.

User network information is incorporated by city embeddings and user embeddings of linked users. User embeddings are introduced along with city embeddings because linked users with city information[1] are limited. We chose to let the model learn geolocation representations of linked users directly via user embeddings. The model can be broken down to several components, details of which are described in Section 3.1.1–3.1.4.

3.1.1 Text Component

We describe the text component of the model, which is the "TEXT" section in Figure 1. Figure 2 presents an overview of the text component. The component consists of a recurrent neural network (RNN) (Graves, 2012) layer and attention layers. An input of the component is a *timeline* of a user, which consists of messages in a time sequence.

As an implementation of RNN, we used Gated Recurrent Unit (GRU) (Cho et al., 2014) with a bi-directional setting. In the RNN layer, word embeddings $\boldsymbol{x}$ of a message are processed with the following transition functions:

$$\boldsymbol{z}_t = \sigma\left(\boldsymbol{W}_z \boldsymbol{x}_t + \boldsymbol{U}_z \boldsymbol{h}_{t-1} + \boldsymbol{b}_z\right) \quad (1)$$

$$\boldsymbol{r}_t = \sigma\left(\boldsymbol{W}_r \boldsymbol{x}_t + \boldsymbol{U}_r \boldsymbol{h}_{t-1} + \boldsymbol{b}_r\right) \quad (2)$$

$$\tilde{\boldsymbol{h}}_t = \tanh\left(\boldsymbol{W}_h \boldsymbol{x}_t + \boldsymbol{U}_h \left(\boldsymbol{r}_t \odot \boldsymbol{h}_{t-1}\right) + \boldsymbol{b}_h\right) \quad (3)$$

$$\boldsymbol{h}_t = \left(\boldsymbol{1} - \boldsymbol{z}_t\right) \odot \boldsymbol{h}_{t-1} + \boldsymbol{z}_t \odot \tilde{\boldsymbol{h}}_t \quad (4)$$

where $\boldsymbol{z}_t$ is an update gate, $\boldsymbol{r}_t$ is a reset gate, $\tilde{\boldsymbol{h}}_t$ is a candidate state, $\boldsymbol{h}_t$ is a state, $\boldsymbol{W}_z, \boldsymbol{W}_r, \boldsymbol{W}_h, \boldsymbol{U}_z, \boldsymbol{U}_r, \boldsymbol{U}_h$ are weight matrices, $\boldsymbol{b}_z, \boldsymbol{b}_r, \boldsymbol{b}_h$ are bias vectors, σ is a logistic sigmoid function, and $\odot$ is an element-wise multiplication operator. The bi-directional GRU outputs $\overrightarrow{\boldsymbol{h}}$

[1] City information are provided by a dataset. The detail of the city information is explained in Section 4.

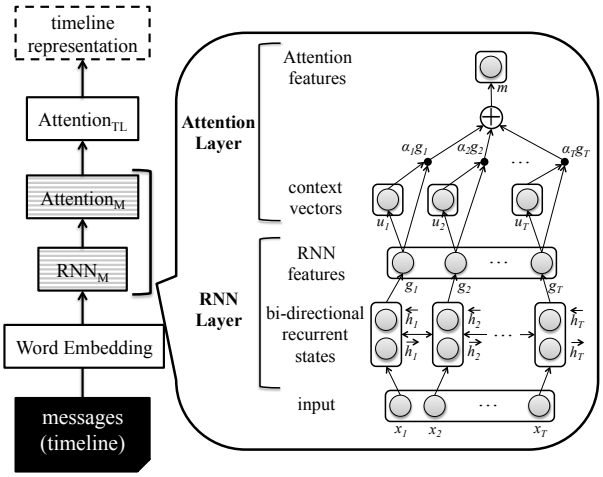


Figure 2: Overview of the text component with detailed description of $\mathrm{RNN_M}$ and $\mathrm{Attention_M}$.

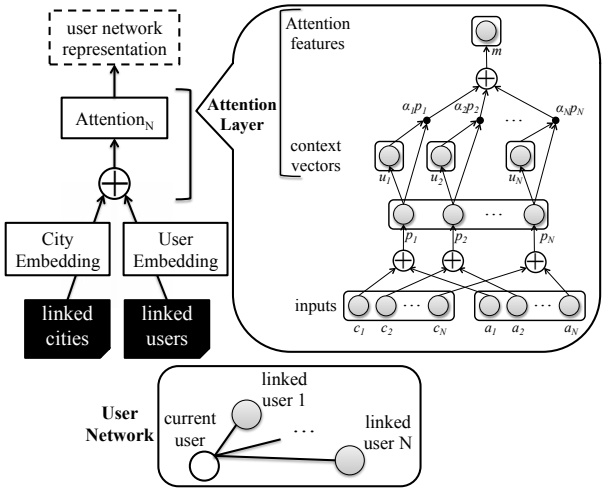


Figure 3: Overview of the user network component with a detailed description of the element-wise addition and $\mathrm{Attention_N}$.

and $\overleftarrow{\boldsymbol{h}}$ are concatenated to form $\boldsymbol{g}$ *where* $\boldsymbol{g}_t = \overrightarrow{\boldsymbol{h}_t} \| \overleftarrow{\boldsymbol{h}_t}$ and are passed to the first attention layer $\mathrm{Attention_M}$.

$\mathrm{Attention_M}$ computes a message representation $\boldsymbol{m}$ as a weighted sum of $\boldsymbol{g}_t$ with weight α_t:

$$\boldsymbol{m} = \sum_t \alpha_t \boldsymbol{g}_t \tag{5}$$

$$\alpha_t = \frac{\exp\left(\boldsymbol{v}_\alpha^T \boldsymbol{u}_t\right)}{\sum_t \exp\left(\boldsymbol{v}_\alpha^T \boldsymbol{u}_t\right)} \tag{6}$$

$$\boldsymbol{u}_t = \tanh\left(\boldsymbol{W}_\alpha \boldsymbol{g}_t + \boldsymbol{b}_\alpha\right) \tag{7}$$

where $\boldsymbol{v}_\alpha$ is a weight vector, $\boldsymbol{W}_\alpha$ is a weight matrix, and $\boldsymbol{b}_\alpha$ a bias vector. $\boldsymbol{u}_t$ is an attention context vector calculated from $\boldsymbol{g}_t$ with a single fully-connected layer (Eq. 7). $\boldsymbol{u}_t$ is normalized with softmax to obtain α_t as a probability (Eq. 6). The message representation $\boldsymbol{m}$ is passed to the second attention layer $\mathrm{Attention_{TL}}$ to obtain a timeline representation from message representations.

3.1.2 Text and Metadata Component

We describe text and metadata components of the model, which is the "TEXT&META" section in Figure 1. This component considers the following three types of metadata along with text: **location** a text field in which a user is allowed to write the user location freely, **description** a text field a user can use for self-description, and **timezone** a selective field from which a user can choose a timezone. Note that certain percentages of these fields are not available[2], and *unknown* tokens are used for inputs in such cases.

We process location fields and description fields similarly to messages using an RNN layer and an attention layer. Because there is only one location and one description per user, a second attention layer is not required, as it is in the text component. We also chose to share word embeddings among the messages, the location, and the description processes because these inputs are all textual information. For the timezone, an embedding is assigned for each timezone value. A processed timeline representation, a location representation, and a description representation are then passed to the attention layer $\mathrm{Attention_U}$ with a timezone representation. $\mathrm{Attention_U}$ combines these four representations and outputs a user representation. This combination is done as in $\mathrm{Attention_{TL}}$ with four representations as $g_1 \ldots g_4$ in Eq. 5.

3.1.3 User Network Component

We describe the user network component of the model, which is the "USERNET" section in Figure 1. Figure 3 presents an overview of the user network component. The model has two inputs *linked cities* and *linked users*. Users connected with a user network are extracted as linked users. We treat their cities[3] as linked cities. Linked cities and linked users are assigned with city embeddings $\boldsymbol{c}$ and user embeddings $\boldsymbol{a}$ respectively. $\boldsymbol{c}$ and $\boldsymbol{a}$ are then processed to output $\boldsymbol{p} = \boldsymbol{c} \oplus \boldsymbol{a}$, where $\oplus$ is an element-wise addition operator. $\boldsymbol{p}$ is then passed to the subsequent attention layer $\mathrm{Attention_N}$ to obtain a user network representa-

[2]Han et al. (2014) reported missing percentages of 19% for location, 24% for description, and 25%for timezone.

[3]A user with city information implies that the user is included in a training set.

	TwitterUS (train)	W-NUT (train)
#user	279K	782K
#tweet	23.8M	9.03M
tweet/user	85.6	11.6
#edge	3.69M	3.21M
#reduced-edge	2.11M	1.01M
reduced-edge/user	7.04	1.29
#city	339	3028

Table 1: Some properties of TwitterUS (train) and W-NUT (train). We were able to obtain approximately 70–78% of the full datasets because of accessibility changes in Twitter.

tion as in $\mathrm{Attention_U}$.

3.1.4 Model Output

An output of the text and metadata component and an output of the mention network component are further passed to the final attention layer $\mathrm{Attention_{UN}}$ to obtain a merged user representation as in $\mathrm{Attention_U}$. The merged user representation is then connected to labels with a fully connected layer $\mathrm{FC_{UN}}$.

3.2 Sub-models of the Proposed Model

SUB-NN-TEXT We prepare a sub-model SUB-NN-TEXT by adding $\mathrm{FC_U}$ and $\mathrm{FC_{UN}}$ to the text component. This sub-model can be considered as a variant of a neural network model by Yang et al. (2016), which learns a representation of hierarchical text.

SUB-NN-UNET We prepare a sub-model SUB-NN-UNET by connecting the text component and the user network component with $\mathrm{FC_U}$, $\mathrm{Attention_{UN}}$, and $\mathrm{FC_{UN}}$. This model can be regarded as a model that uses text and user network information.

SUB-NN-META We prepare a sub-model SUB-NN-META by adding $\mathrm{FC_U}$ and $\mathrm{FC_{UN}}$ to the metadata component. This model is a text-meta-based model that uses text and metadata.

4 Data

4.1 Dataset Specifications

TwitterUS The first dataset we used is TwitterUS assembled by Roller et al. (2012), which consists of 429K training users, 10K development users, and 10K test users in a North American region. The ground truth location of a user is set to the first geotag of the user in the dataset. We collected TwitterUS tweets using TwitterAPI to reconstruct TwitterUS to obtain metadata along with text. Up to date versions in November–December 2016 were used for the metadata[4]. We additionally assigned city centers to ground truth geotags using the city category of Han et al. (2012) to make city prediction possible in this dataset. TwitterUS (train) in Table 1 presents some properties related to the TwitterUS training set.

W-NUT The second dataset we used is W-NUT, a user-level dataset of the geolocation prediction shared task of W-NUT 2016 (Han et al., 2016). The dataset consists of 1M training users, 10K development users, and 10K test users. The ground truth location of a user is decided by majority voting of the closest city center. Like in TwitterUS, we obtained metadata and texts using TwitterAPI. Up to date versions in August–September 2016 were used for the metadata. W-NUT (train) in Table 1 presents some properties related to the W-NUT training set.

4.2 Construction of the User Network

We construct mention networks (Jurgens, 2013; Compton et al., 2014; Rahimi et al., 2015a,b) from datasets as user networks. To do so, we follow the approach of Rahimi et al. (2015a) and Rahimi et al. (2015b) who use uni-directional mention to set edges of a mention network. An edge is set between the two users nodes if a user mentions another user. The number of uni-directional mention edges for TwitterUS and W-NUT can be found in Table 1.

The uni-directional setting results to large numbers of edges, which often are computationally expensive to process. We restricted edges to satisfy one of the following conditions to reduce the size: (1) both users have ground truth locations or (2) one user has a ground truth location and another user is mentioned 5 times or more in a training set. The number of reduced-edges with these conditions in TwitterUS and W-NUT can be confirmed in Table 1.

5 Evaluation

5.1 Implemented Baselines

5.1.1 LR

LR is an l_1-regularized logistic regression model with k-d tree regions (Roller et al., 2012) used

[4]TwitterAPI returns the current version of metadata even for an old tweet.

in Rahimi et al. (2015a). The model uses tf-idf weighted bag-of-words unigrams for features. This model is simple, but it has shown state-of-the-art performance in cases when only text is available.

5.1.2 MADCEL-B-LR

MADCEL-B-LR, a model presented by (Rahimi et al., 2015a), combines LR with Modified Adsorption (MAD) (Talukdar and Crammer, 2009). MAD is a graph-based label propagation algorithm that optimizes an objective with a prior term, a smoothness term, and an uninformativeness term. LR is combined with MAD by introducing LR results as dongle nodes to MAD.

This model includes an algorithm for the construction of a mention network. The algorithm removes *celebrity users*[5] and *collapses a mention network*[6]. We use binary edges for user network edges because they performed slightly better than weighted edges by accuracy@161 metric in Rahimi et al. (2015a).

5.1.3 LR-STACK

LR-STACK is an ensemble learning model that combines four LR classifiers (LR-MSG, LR-LOC, LR-DESC, LR-TZ) with an l_2-regularized logistic regression meta-classifier (LR-2ND). LR-MSG, LR-LOC, LR-DESC, and LR-TZ respectively use messages, location fields, description fields, and timezones as their inputs. This model is similar to the stacking (Wolpert, 1992) approach taken in Han et al. (2013) and Han et al. (2014), which showed superior performance compared to a feature concatenation approach.

The model takes the following three steps to combine text and metadata: **Step 1** LR-MSG, LR-LOC, LR-DESC, and LR-TZ are trained using a training set, **Step 2** the outputs of the four classifiers on the training set are obtained with 10-fold cross validation, and **Step 3** LR-2ND is trained using the outputs of the four classifiers.

5.1.4 MADCEL-B-LR-STACK

MADCEL-B-LR-STACK is a combined model of MADCEL-B-LR and LR-STACK. LR-STACK results are introduced as dongle nodes to MAD instead of LR results to combine text, metadata, and network information.

[5]Users with more than t unique mentions.

[6]Users not included in training users or test users are removed and disconnected edges with the removals are converted to direct edges.

5.2 Model Configurations

5.2.1 Text Processor

We applied a lower case conversion, a unicode normalization, a Twitter user name normalization, and a URL normalization for text pre-processing. The pre-processed text is then segmented using Twokenizer (Owoputi et al., 2013) to obtain words.

5.2.2 Pre-training of Embeddings

We pre-trained word embeddings using messages, location fields, and description fields of a training set using fastText (Bojanowski et al., 2016) with the skip-gram algorithm. We also pre-trained user embeddings using the non-reduced mention network described in Section 4.2 of a training set with LINE (Tang et al., 2015). The detail of pre-training parameters are described in Appendix A.1.

5.2.3 Neural Network Optimization

We chose an objective function of our models to cross-entropy loss. l_2 regularization was applied to the RNN layers, the attention context vectors, and the FC layers of our models to avoid overfitting. The objective function was minimized through stochastic gradient descent over shuffled mini-batches with Adam (Kingma and Ba, 2014).

5.2.4 Model Parameters

The layers and the embeddings in our models have unit size and embedding dimension parameters. Our models and the baseline models have regularization parameter α, which is sensitive to a dataset. The baseline models have additional k-d tree bucket size c, celebrity threshold t, and MAD parameters μ_1, μ_2, and μ_3, which are also data sensitive.

We chose optimal values for these parameters in terms of accuracy with a grid search using the development sets of TwitterUS and W-NUT. Details of the parameter selection strategies and the selected values are described in Appendix A.2.

5.2.5 Metrics

We evaluate the models in the following four commonly used metrics in geolocation prediction: **accuracy** the percentage of correctly predicted cities, **accuracy@161** a relaxed accuracy that takes prediction errors within 161 km as correct predictions, **median error distance** median value of error distances in predictions, and **mean error distance** mean value of error distances in predictions.

	Model	Sign. Test ID	Accuracy	Accuracy @161	Error Distance	
					Median	Mean
Baselines (reported)	Han et al. (2012)		26.0	45.0	260	814
	Wing and Baldridge (2014)		-	49.2	170.5	703.6
	LR (Rahimi et al. 2015b)		-	50	159	686
	LR-NA (Rahimi et al. 2016)		-	51	148	636
	MADCEL-B-LR (Rahimi et al. 2015a)		-	60	77	533
	MADCEL-W-LR (Rahimi et al. 2015a)		-	60	78	529
Baselines (implemented)	LR	i	42.0	52.7	121.1	666.6
	MADCEL-B-LR	ii	50.2	60.1	66.5	582.8
	LR-STACK	iii	50.8	64.1	42.3*	427.7
	MADCEL-B-LR-STACK	iv	55.7	67.7	45.1	412.7
Our Models	SUB-NN-TEXT	i	44.9**	55.6**	110.5	585.1**
	SUB-NN-UNET	ii	51.0	61.5*	65.0	481.5**
	SUB-NN-META	iii	54.6**	67.2**	46.8	356.3**
	Proposed Model	iv	58.5**	70.1**	41.9*	335.7**

Table 2: Performances of our models and the baseline models on TwitterUS. Significance tests were performed between models with same Sign. Test IDs. The shaded lines represent values copied from related papers. Asterisks denote significant improvements against paired counterparts with 1% confidence (**) and 5% confidence (*).

	Model	Sign. Test ID	Accuracy	Accuracy @161	Error Distance	
					Median	Mean
Baselines (reported)	Miura et al. (2016)		47.6	-	16.1	1122.3
	Jayasinghe et al. (2016)		52.6	-	21.7	1928.8
Baselines (implemented)	LR	i	34.1	46.7	248.7	2216.4
	MADCEL-B-LR	ii	36.2	49.7	166.3	2120.6
	LR-STACK	iii	51.2	64.9	0.0	1496.4
	MADCEL-B-LR-STACK	iv	51.6	65.3	0.0	1471.9
Our Models	SUB-NN-TEXT	i	35.4**	50.3**	155.8**	1592.6**
	SUB-NN-UNET	ii	38.1**	53.3**	99.9**	1498.6**
	SUB-NN-META	iii	54.7**	70.2**	0.0	825.8**
	Proposed Model	iv	56.4**	71.9**	0.0	780.5**

Table 3: Performance of our models and baseline models on W-NUT. The same notations as those in Table 2 are used in this table.

5.3 Result

Performance on TwitterUS

Table 2 presents results of our models and the implemented baseline models on TwitterUS. We also list values from earlier reports (Han et al., 2012; Wing and Baldridge, 2014; Rahimi et al., 2015a,b, 2016) to make our results readily comparable with past reported values.

We performed some statistical significance tests among model pairs that share the same inputs. The values in the Sign. Test ID column of Table 2 represent the IDs of these pairs. As a preparation of statistical significance tests, accuracies, accuracy@161s, and error distances of each test user were calculated for each model pair. Two-sided Fisher-Pittman Permutation tests were used for testing accuracy and accuracy@161. Mood's median test was used for testing error distance in terms of median. Paired t-tests were used for testing error distance in terms of mean.

We confirmed the significance of improvements in accuracy@161 and mean distance error for all of our models. Three of our models also improved in terms of accuracy. Especially, the proposed model achieved a 2.8% increase in accuracy and a 2.4% increase in accuracy@161 against the counterpart baseline model MADCEL-B-LR-STACK. One negative result we found was the median error distance between SUB-NN-META and LR-STACK. The baseline model LR-STACK performed 4.5 km significantly better than our model.

Performance on W-NUT

Table 3 presents the results of our models and the implemented baseline models on W-NUT. As for TwitterUS, we listed values from Miura et al. (2016) and Jayasinghe et al. (2016). We tested the significance of these results in the same way as we did for TwitterUS.

We confirmed significant improvement in the four metrics for all of our models. The proposed model achieved a 4.8% increase in accuracy and a

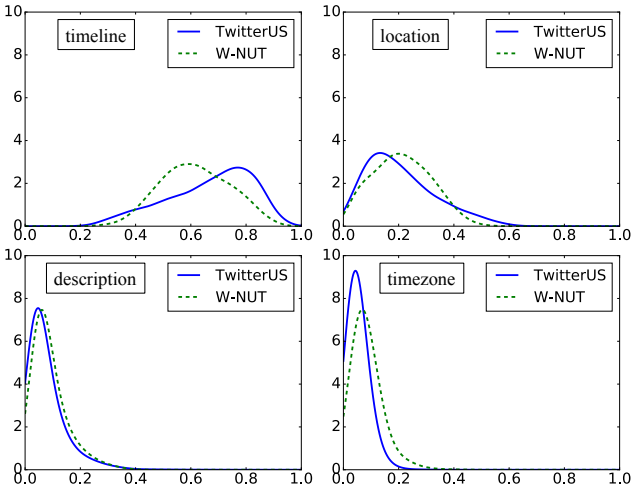


Figure 4: Estimated probability density functions of the four representations in $\text{Attention}_{\text{U}}$.

6.6% increase in accuracy@161 against the counterpart baseline model MADCEL-B-LR-STACK. The accuracy is 3.8% higher against the previously reported best value (Jayasinghe et al., 2016) which combined texts, metadata, and user network information with an ensemble method.

6 Discussion

6.1 Analyses of Attention Probabilities

6.1.1 Unification Strategies

In the evaluation, the proposed model has implicitly shown effectiveness at unifying text, metadata, and user network representations through improvements in the four metrics. However, details of the unification processes are not clear from the model outputs because they are merely the probabilities of estimated locations. To gain insight into the unification processes, we analyzed the states of two attention layers: $\text{Attention}_{\text{U}}$ and $\text{Attention}_{\text{UN}}$ in Figure 1.

Figure 4 presents the estimated probability density functions (PDFs) of the four input representations for $\text{Attention}_{\text{U}}$. These PDFs are estimated with kernel density estimation from the development sets of TwitterUS and W-NUT, where all four representations are available. From the PDFs, it is apparent that the model assigns higher probabilities to time line representations than to other three representations in TwitterUS compared to W-NUT. This finding is reasonable because timelines in TwitterUS consist of more tweets (tweet/user in Table 1) and are likely to be more informative than in W-NUT.

Figure 5 presents the estimated PDFs of user network representations for $\text{Attention}_{\text{UN}}$. These

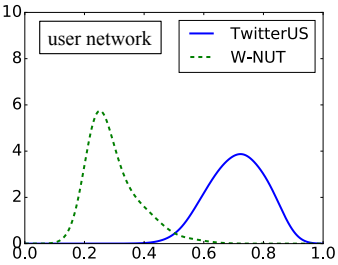


Figure 5: Estimated probability density functions of user network representations in $\text{Attention}_{\text{UN}}$.

PDFs are estimated from the development sets of TwitterUS and W-NUT, where both input representations are available. Strong preference of network representation for TwitterUS against W-NUT is found in the PDFs. This finding is intuitive because TwitterUS has substantially more user network edges (reduced-edge/user in Table 1) than W-NUT, which is likely to benefit more from user network information.

6.1.2 Attention Patterns

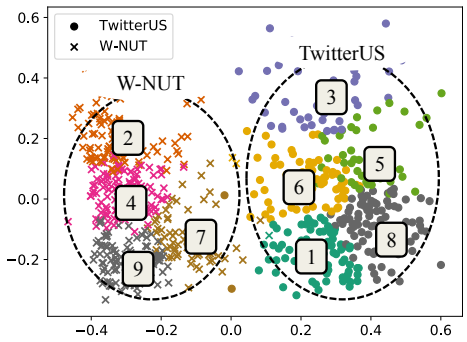
We further analyzed the proposed model by clustering attention probabilities to capture typical attention patterns. For each user, we assigned six attention probabilities of $\text{Attention}_{\text{U}}$ and $\text{Attention}_{\text{UN}}$ as features for a clustering. A k-means clustering was performed over these users with 9 clusters. The clustering clearly separated the users to 5 clusters for TwitterUS users and 4 clusters for W-NUT users. We extracted typical users of each cluster by selecting the closest users of the cluster centroids. Figure 6 shows a clustering result and the attention probabilities of these users.

These attention probabilities can be considered as typical attention patterns of the proposed model and match with the previously estimated PDFs. For example, cluster 2 and 3 represent an attention pattern that processes users by balancing the representations of locations along with the representations of timelines. Additionally, the location probabilities in this pattern are in the right tail region of the location PDF.

6.2 Limitations of Proposed Model

6.2.1 City Prediction

The evaluation produced improvements in most of our models in the four metrics. One exception we found was the median distance error between SUB-NN-META and LR-STACKING in TwitterUS. Because the median distance error of SUB-NN-META was quite low (46.8 km), we



Cluster ID	Dataset	Timeline	Location	Description	Timezone	User	User Network
1	TwitterUS	0.843	0.082	0.040	0.035	0.359	0.641
2	W-NUT	0.517	0.317	0.081	0.085	0.732	0.268
3	TwitterUS	0.432	0.430	0.069	0.069	0.319	0.681
4	W-NUT	0.637	0.160	0.097	0.105	0.737	0.263
5	TwitterUS	0.593	0.219	0.114	0.075	0.230	0.770
6	TwitterUS	0.672	0.214	0.069	0.045	0.365	0.635
7	W-NUT	0.741	0.077	0.080	0.102	0.605	0.395
8	TwitterUS	0.766	0.099	0.068	0.067	0.222	0.778
9	W-NUT	0.800	0.067	0.056	0.078	0.730	0.270

Figure 6: A k-means clustering result and the attention probabilities of users that are closest to the cluster centroids. The underlined values are the max values of the two datasets for each column.

Model	Error Distance		
	Median	Mean	σ
Oracle	23.3	31.4	30.1

Table 4: Error distance values in TwitterUS with oracle predictions. σ in the table denotes the standard deviation.

measured the performance of an oracle model where city predictions are all correct (accuracy of 100%) in the test set.

Table 4 denotes this oracle performance. The oracle mean error distance is 31.4 km. Its standard deviation is 30.1. Note that ground truth locations of TwitterUS are geotags and will not exactly match the oracle city centers. These oracle values imply that the current median error distances are close to the lower bound of the city classification approach and that they are difficult to improve.

6.2.2 Errors with High Confidences

The proposed model still contains 28–30% errors even in accuracy@161. A qualitative analysis of errors with high confidences was performed to investigate cases that the model fails. We found two common types of error in the error analysis. The first is a case when a location field is incorrect due to a reason such as a house move. For example, the model predicted "Hong Kong" for a user with a location field of "Hong Kong" but has the gold location of "Toronto". The second is a case when a user tweets a place name of a travel. For example, the model predicted "San Francisco" for a user who tweeted about a travel to "San Francisco" but has the gold location of "Boston".

These two types of error are difficult to handle with the current architecture of the proposed model. The architecture only supports single location field which disables the model to track location changes. The architecture also treats each tweet independently which forbids the model to express a temporal state like traveling.

7 Conclusion

As described in this paper, we proposed a complex neural network model for geolocation prediction. The model unifies text, metadata, and user network information. The model achieved the maximum of a 3.8% increase in accuracy and a maximum of 6.6% increase in accuracy@161 against several previous state-of-the-art models. We further analyzed the states of several attention layers, which revealed that the probabilities assigned to timeline representations and user network representations match to some statistical characteristics of datasets.

As future works of this study, we are planning to expand the proposed model to handle multiple locations and a temporal state to capture location changes and states like traveling. Additionally, we plan to apply the proposed model to other social media analyses such as gender analysis and age analysis. In these analyses, metadata like location fields and timezones may not be effective like in geolocation prediction. However, a user network is known to include various user attributes information including gender and age (McPherson et al., 2001) which suggests the unification of text and user network information to result in a success as in geolocation prediction.

Acknowledgments

We would like to thank the members of Okumura–Takamura Group at Tokyo Institute of Technology for having insightful discussions about user profiling models in social media. We would also like to thank the anonymous reviewer for their comments to improve this paper.

References

Amr Ahmed, Liangjie Hong, and Alexander J. Smola. 2013. Hierarchical geographical modeling of user locations from social media posts. In *Proceedings of the 22nd International Conference on World Wide Web*. pages 25–36.

Lars Backstrom, Eric Sun, and Cameron Marlow. 2010. Find me if you can: Improving geographical prediction with social and spatial proximity. In *Proceedings of the 19th International Conference on World Wide Web*. pages 61–70.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *Computing Research Repository* abs/1409.0473. http://arxiv.org/abs/1409.0473.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606* .

Miriam Cha, Youngjune Gwon, and H. T. Kung. 2015. Twitter geolocation and regional classification via sparse coding. In *Proceedings of the Ninth International AAAI Conference on Web and Social Media*.

Zhiyuan Cheng, James Caverlee, and Kyumin Lee. 2010. You are where you tweet: A content-based approach to geo-locating Twitter users. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*. pages 759–768.

Zhiyuan Cheng, James Caverlee, and Kyumin Lee. 2013. A content-driven framework for geolocating microblog users. *ACM Transactions on Intelligent Systems and Technology* 4(1):1–27. Article 2.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*. pages 1724–1734.

Ryan Compton, David Jurgens, and David Allen. 2014. Geotagging one hundred million Twitter accounts with total variation minimization. In *Proceedings of the 2014 IEEE International Conference on Big-Data*. pages 393–401.

David J. Crandall, Lars Backstrom, Daniel Huttenlocher, and Jon Kleinberg. 2009. Mapping the world's photos. In *Proceedings of the 18th International Conference on World Wide Web*. pages 761–770.

Aron Culotta. 2010. Towards detecting influenza epidemics by analyzing Twitter messages. In *Proceedings of the First Workshop on Social Media Analytics*. pages 115–122.

Clodoveu A. Davis Jr., Gisele L. Pappa, Diogo Rennó Rocha de Oliveira, and Filipe de L. Arcanjo. 2011. Inferring the location of Twitter messages based on user relationships. *Transactions in GIS* 15(6):735–751.

Jacob Eisenstein, Amr Ahmed, and Eric P. Xing. 2011. Sparse additive generative models of text. In *Proceedings of the 28th International Conference on Machine Learning*. pages 1041–1048.

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

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT Press.

Alex Graves. 2012. *Supervised Sequence Labelling with Recurrent Neural Networks*, volume 385 of *Studies in Computational Intelligence*. Springer-Verlag Berlin Heidelberg.

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

Bo Han, Paul Cook, and Timothy Baldwin. 2013. A stacking-based approach to twitter user geolocation prediction. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. pages 7–12.

Bo Han, Paul Cook, and Timothy Baldwin. 2014. Text-based Twitter user geolocation prediction. *Journal of Artificial Intelligence Research* 49(1):451–500.

Bo Han, Afshin Rahimi, Leon Derczynski, and Timothy Baldwin. 2016. Twitter geolocation prediction shared task of the 2016 workshop on noisy user-generated text. In *Proceedings of the Second Workshop on Noisy User-generated Text*. pages 213–217.

Brent Hecht, Lichan Hong, Bongwon Suh, and Ed H. Chi. 2011. Tweets from Justin Bieber's heart: the dynamics of the location field in user profiles. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. pages 237–246.

Liangjie Hong, Amr Ahmed, Siva Gurumurthy, Alexander J. Smola, and Kostas Tsioutsiouliklis. 2012. Discovering geographical topics in the Twitter stream. In *Proceedings of the 21st International Conference on World Wide Web*. pages 769–778.

Gaya Jayasinghe, Brian Jin, James Mchugh, Bella Robinson, and Stephen Wan. 2016. CSIRO Data61 at the WNUT geo shared task. In *Proceedings of the Second Workshop on Noisy User-generated Text*. pages 218–226.

David Jurgens. 2013. That's what friends are for: Inferring location in online social media platforms based on social relationships. In *Proceedings of the Seventh International AAAI Conference on Web and Social Media*.

David Jurgens, Tyler Finethy, James McCorriston, Yi Xu, and Derek Ruths. 2015. Geolocation prediction in Twitter using social networks: A critical analysis and review of current practice. In *Proceedings of the Ninth International AAAI Conference on Web and Social Media*.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* .

Sheila Kinsella, Vanessa Murdock, and Neil O'Hare. 2011. "I'm eating a sandwich in Glasgow": Modeling locations with tweets. In *Proceedings of the Third International Workshop on Search and Mining User-generated Contents*. pages 61–68.

Longbo Kong, Zhi Liu, and Yan Huang. 2014. SPOT: Locating social media users based on social network context. *Proceedings of the VLDB Endowment* 7(13):1681–1684.

Rui Li, Shengjie Wang, and Kevin Chen-Chuan Chang. 2012a. Multiple location profiling for users and relationships from social network and content. *Proceedings of the VLDB Endowment* 5(11):1603–1614.

Rui Li, Shengjie Wang, Hongbo Deng, Rui Wang, and Kevin Chen-Chuan Chang. 2012b. Towards social user profiling: Unified and discriminative influence model for inferring home locations. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. pages 1023–1031.

Ji Liu and Diana Inkpen. 2015. Estimating user location in social media with stacked denoising autoencoders. In *Proceedings of the First Workshop on Vector Space Modeling for Natural Language Processing*. pages 201–210.

Jalal Mahmud, Jeffrey Nichols, and Clemens Drews. 2012. Where is this tweet from? Inferring home locations of Twitter users. In *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media*.

Eugenio Martínez-Cámara, Maria Teresa Martín-Valdivia, Luis Alfonso Ureña López, and Arturo Montejo Raéz. 2014. Sentiment analysis in Twitter. *Natural Language Engineering* 20(1):1–28.

Jeffrey McGee, James Caverlee, and Zhiyuan Cheng. 2013. Location prediction in social media based on tie strength. In *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*. pages 459–468.

Miller McPherson, Lynn Smith-Lovin, and James M Cook. 2001. Birds of a feather: Homophily in social networks. *Annual review of sociology* 27(1):415–444.

Yasuhide Miura, Motoki Taniguchi, Tomoki Taniguchi, and Tomoko Ohkuma. 2016. A simple scalable neural networks based model for geolocation prediction in Twitter. In *Proceedings of the Second Workshop on Noisy User-generated Text*. pages 235–239.

Simon E. Overell. 2009. *Geographic Information Retrieval: Classification, Disambiguation, and Modelling*. Ph.D. thesis, Imperial College London.

Olutobi Owoputi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A. Smith. 2013. Improved part-of-speech tagging for online conversational text with word clusters. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. pages 380–390.

Afshin Rahimi, Trevor Cohn, and Timothy Baldwin. 2015a. Twitter user geolocation using a unified text and network prediction model. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. pages 630–636.

Afshin Rahimi, Trevor Cohn, and Timothy Baldwin. 2016. pigeo: A python geotagging tool. In *Proceedings of ACL-2016 System Demonstrations*. pages 127–132.

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

Delip Rao, David Yarowsky, Abhishek Shreevats, and Manaswi Gupta. 2010. Classifying latent user attributes in Twitter. In *Proceedings of the Second International Workshop on Search and Mining User-generated Contents*. pages 37–44.

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

Dominic Rout, Kalina Bontcheva, Daniel Preoţiuc-Pietro, and Trevor Cohn. 2013. Where's @wally?: A classification approach to geolocating users based on their social ties. In *Proceedings of the 24th ACM Conference on Hypertext and Social Media*. pages 11–20.

Adam Sadilek, Henry Kautz, and Jeffrey P. Bigham. 2012. Finding your friends and following them to where you are. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*. pages 723–732.

Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. 2010. Earthquake shakes Twitter users: Real-time event detection by social sensors. In *Proceedings of the 19th International Conference on World Wide Web*. pages 851–860.

Axel Schulz, Aristotelis Hadjakos, Heiko Paulheim, Johannes Nachtwey, and Max Mühlhäuser. 2013. A multi-indicator approach for geolocalization of tweets. In *Proceedings of the Seventh International AAAI Conference on Web and Social Media*.

Pavel Serdyukov, Vanessa Murdock, and Roelof van Zwol. 2009. Placing Flickr photos on a map. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. pages 484–491.

Partha Pratim Talukdar and Koby Crammer. 2009. New regularized algorithms for transductive learning. In *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases: Part II*. pages 442–457.

Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. 2015. LINE: Large-scale information network embedding. In *Proceedings of the 24th International Conference on World Wide Web*. pages 1067–1077.

Andranik Tumasjan, Timm O. Sprenger, Philipp G. Sandner, and Isabell M. Welpe. 2010. Predicting elections with Twitter: What 140 characters reveal about political sentiment. In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*. pages 178–185.

Benjamin Wing and Jason Baldridge. 2011. Simple supervised document geolocation with geodesic grids. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. pages 955–964.

Benjamin Wing and Jason Baldridge. 2014. Hierarchical discriminative classification for text-based geolocation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*. pages 336–348.

David H. Wolpert. 1992. Stacked generalization. *Neural Networks* 5(2):241–259.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. pages 1480–1489.

A Supplemental Materials

A.1 Parameters of Embedding Pre-training

Word embeddings were pre-trained with the parameters of learning rate=0.025, window size=5, negative sample size=5, and epoch=5. User embeddings were pre-trained with the parameters of initial learning rate=0.025, order=2, negative sample size=5, and training sample size=100M.

A.2 Model Parameters and Parameter Selection Strategies

Unit Sizes, Embedding Dimensions, and a Max Tweet Number

The layers and the embeddings in our models have unit size and embedding dimension parameters. We also restricted the maximum number of tweets per user for TwitterUS to reduce memory footprints. Table 5 shows the values for these parameters. Smaller values were set for TwitterUS because TwitterUS is approximately 2.6 times larger in terms of tweet number. It was computationally expensive to process TwiiterUS in the same settings as W-NUT.

Regularization Parameters and Bucket Sizes

We chose optimal values of α using a grid search with the development sets of TwitterUS and W-NUT. The range of α was set as the following:
$\alpha \in \{1e^{-4}, 5e^{-5}, 1e^{-5}, 5e^{-6}, 1e^{-6}, 5e^{-7}, 1e^{-7}, 5e^{-8}, 1e^{-8}\}$.

We also chose optimal values of c using grid search with the development sets of TwitterUS and W-NUT for the baseline models. The range of c was set as the following for TwitterUS:
$c \in \{50, 100, 150, 200, 250, 300, 339\}$.
The following was set for W-NUT:
$c \in \{100, 200, 300, 400, 500, 600, 700, 800, 900, 1000, 1500, 2000, 2500, 3000, 3028\}$.
Table 6 presents selected values of α and c. For LR-STACK and MADCEl-B-LR-STACK, different parameters of α and c were selected for each logistic regression classifier.

MAD Parameters and Celebrity Threshold

The MAD parameters μ_1, μ_2, and μ_3 and celebrity threshold t were also chosen using grid search with the development sets of TwitterUS and W-NUT. The ranges of μ_1, μ_2, and μ_3 were set as the following:
$\mu_1 \in \{1.0\}$, $\mu_2 \in \{0.001, 0.01, 0.1, 1.0, 10.0\}$,
$\mu_3 \in \{0.0, 0.001, 0.01, 0.1, 1.0, 10.0\}$.
The range of t for TwitterUS was set as $t \in \{2, \ldots, 16\}$. The range of t for W-NUT was set as $t \in \{2, \ldots, 6\}$. Table 6 presents selected values of μ_1, μ_2, μ_3, and t.

	TwitterUS	W-NUT
RNN unit size	100	200
Attention context vector size	200	400
FC unit size	200	400
Word embedding dimension	100	200
Timezone embedding dimension	200	400
City embedding dimension	200	400
User embedding dimension	200	400
Max tweet number per user	200	-

Table 5: Unit sizes, embedding dimensions, and max tweet numbers of our models.

Model	Parameter	TwitterUS	W-NUT
SUB-NN-TEXT	α	1e-8	1e-7
SUB-NN-UNET		1e-6	5e-8
SUB-NN-META		1e-8	5e-8
Proposed Model		1e-6	5e-8
LR MADCEL-B-LR	α	1e-6	5e-7
	c	300	3000
LR-STACK MADCEL-B-LR-STACK	α_{MSG}	1e-6	5e-7
	α_{LOC}	1e-6	1e-6
	α_{DESC}	5e-6	1e-6
	α_{TZ}	1e-4	5e-6
	α_{2ND}	1e-6	1e-7
	c_{MSG}	300	3000
	c_{LOC}	300	3000
	c_{DESC}	250	1500
	c_{TZ}	100	2500
	c_{2ND}	300	2000

Table 6: Regularization parameters and bucket sizes selected for our models and baseline models.

Model	Parameter	TwitterUS	W-NUT
MADCEL-B-LR	μ_1	1.0	1.0
	μ_2	1.0	10.0
	μ_3	0.01	0.1
	t	5	4
MADCEL-B-LR-STACK	μ_1	1.0	1.0
	μ_2	1.0	1.0
	μ_3	0.1	0.0
	t	4	2

Table 7: MAD parameters and celebrity threshold selected for baseline models.