On Predicting Geolocation of Tweets using Convolutional Neural Networks

Binxuan Huang

binxuanh@cs.cmu.edu

Kathleen M. Carley

kathleen.carley@cs.cmu.edu

Outline

- Introduction and problem description
- Our method:
  - Useful Features
  - Our neural network architecture
- Experiments:
  - Country-level prediction
  - City-level prediction
- A case study

Outline

- Introduction and problem description
- Our method:
  - Useful Features
  - Our neural network architecture
- Experiments:
  - Country-level prediction
  - City-level prediction
- A case study

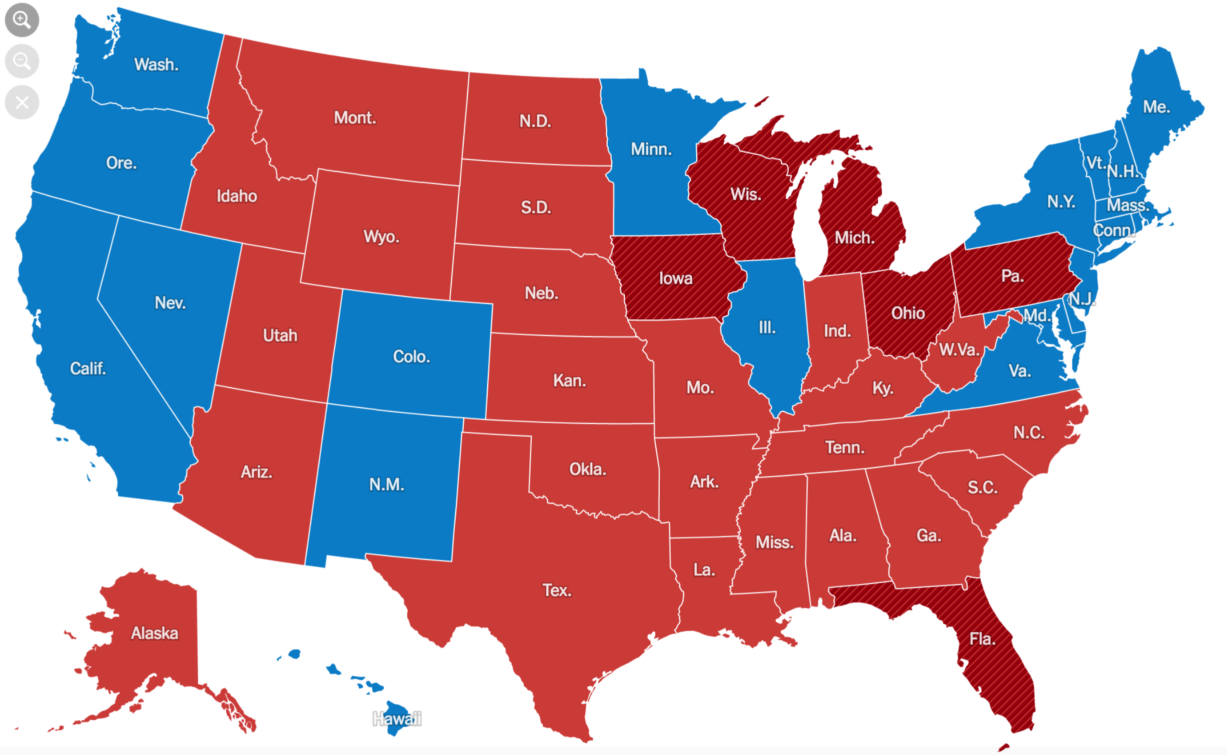
Carnegie Mellon

isr institute for SOFTWARE RESEARCH

Introduction

- Why we care about online users' location?
  - Understand online user's opinion across different regions.
  - A typical example is US president election: we are interested in the regional user opinion.

Understanding what regional online users are doing or thinking requires location information for each user.



https://www.nytimes.com/elections/results/president

CASOS

Introduction

- Twitter has become a popular platform for researchers studying social phenomenon.
- A common way for researcher collecting Twitter data is using Twitter's streaming API[3]
  - Following users
  - Following terms
  - Following Geo-bounding boxes. As reported in [4], less than 1% of tweets contain structured geolocation information.
- Using a geo-bounding box means we will lose the majority of information.

Introduction

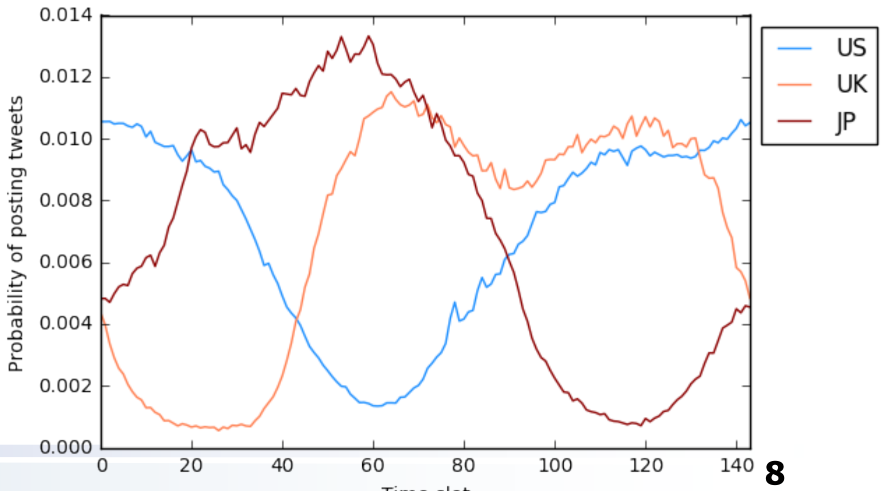
- Task: predict user's location from public accessible information in one single tweet.

Outline

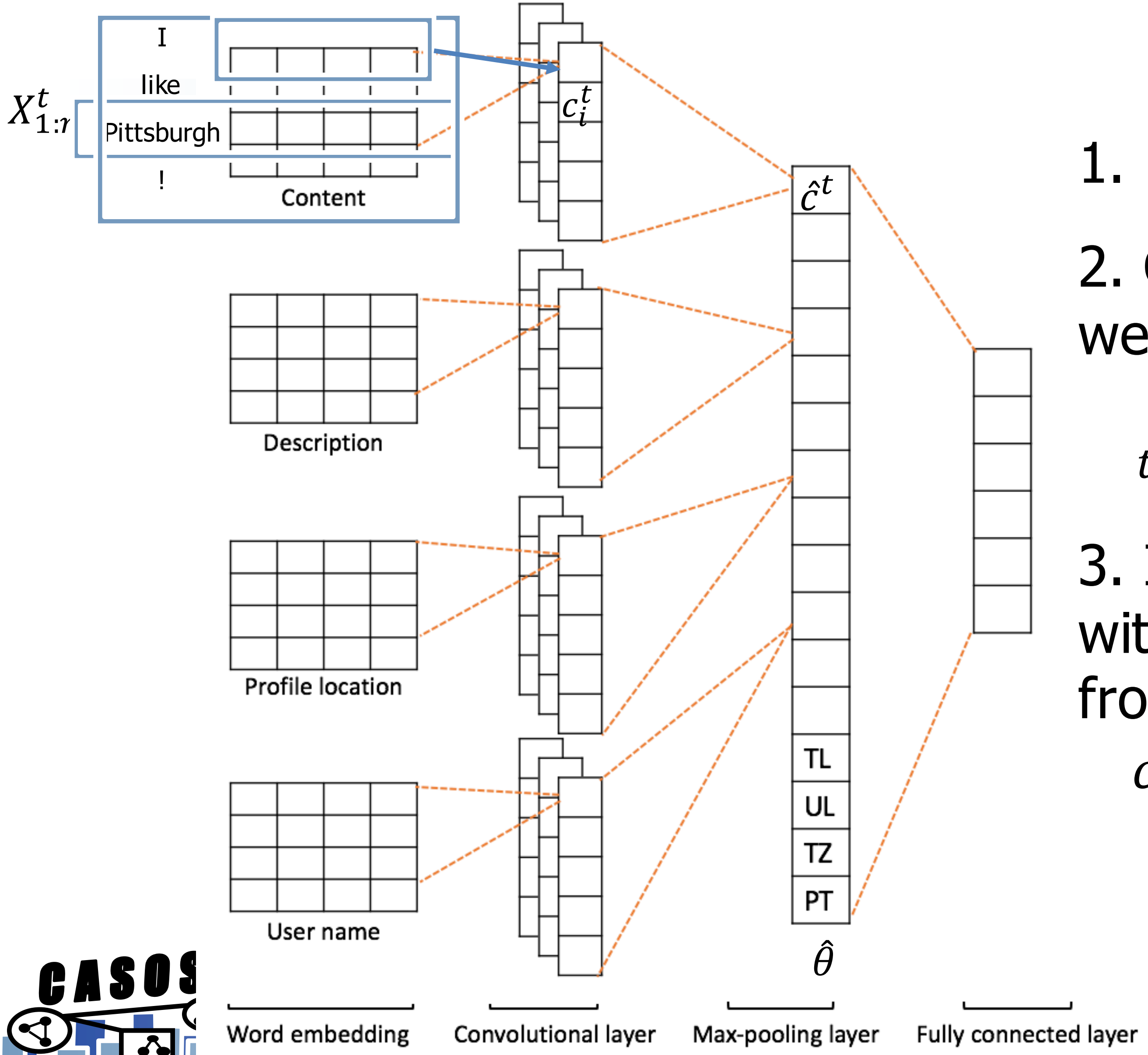
- Introduction and problem description
- Our method:
  - Useful Features
  - Our neural network architecture
- Experiments:
  - Country-level prediction
  - City-level prediction
- Conclusion

Useful Features in Tweet Json

- Text features
  - Tweet content
  - Personal description in user's profile
  - Location in user's profile
  - Username
- Categorical features
  - Tweet language(TL)
  - User language in user's profile(UL)
  - User timezone in user's profile(TZ)
  - Posting time(PT)



Tweet Location Prediction Architecture



1. Each word is represented by a vector.

2. Concatenating the word embedding vectors, we got the text representation.

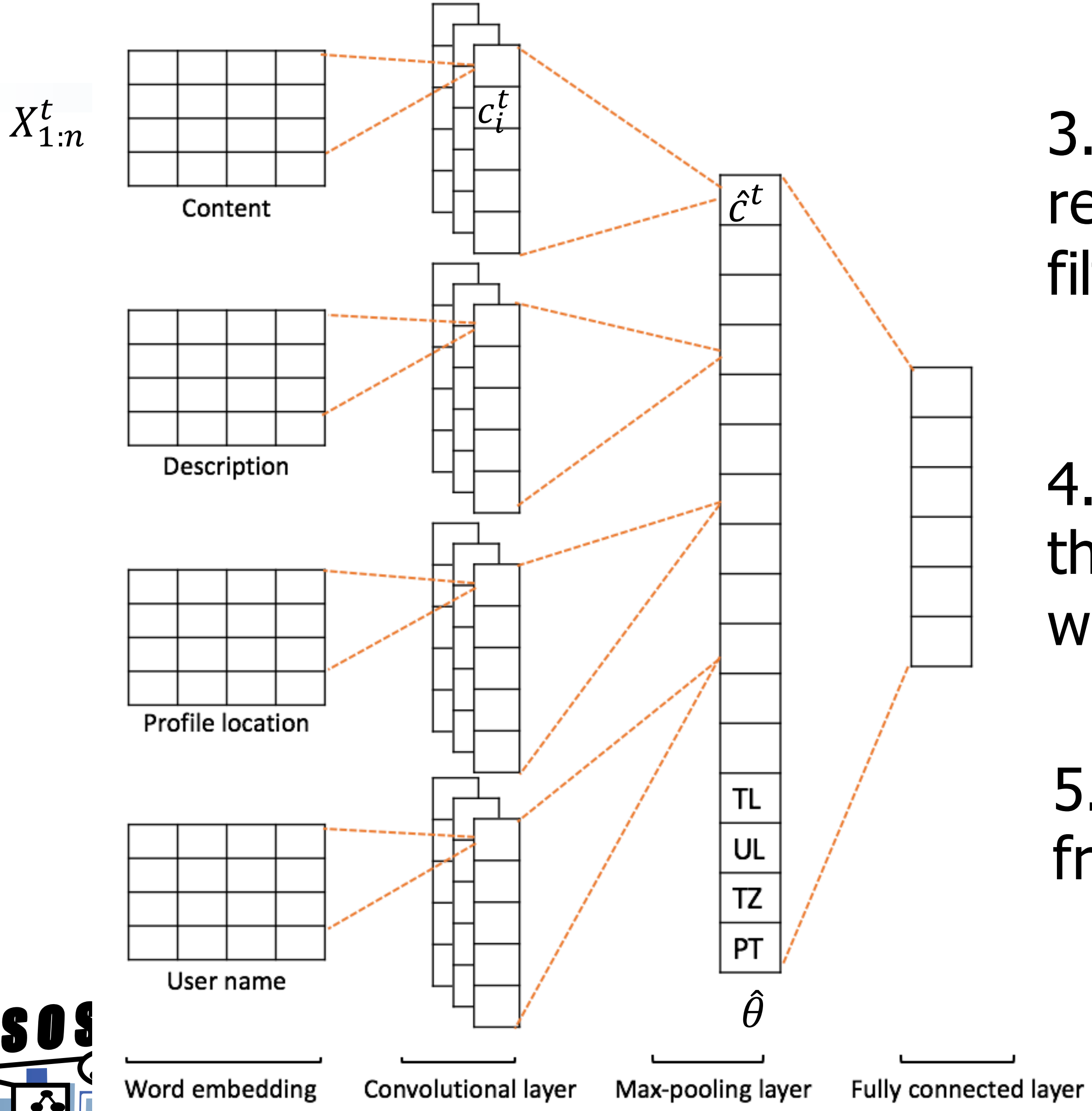
$$X_{1:n}^t = x_1^t \oplus x_2^t \oplus \dots \oplus x_n^t,$$
$$t \in \{\text{content}, \text{description}, \text{profile location}, \text{username}\}$$

3. In the conv. layer, there are *m* filters with size *h* that extracts useful features from texts.

$$c_i^t = f\left(w \cdot X_{i:i+h-1}^t + b\right), \text{ where } f(x) = \max(x, 0)$$

Architecture is based on [1]

Tweet Location Prediction Architecture



3. Max-pooling layer selects the most representative features generated by each filter in the convolutional layer.

$$\hat{c}^t = \max(c^t_1, c^t_2, \ldots, c^t_{n-h+1})$$

4. Assume there are *m* convolutional filters, then we can get a feature vector $\theta \in R^{4m}$ which is appended by TL, UL, TZ and PT.

5. The probability of one tweet coming from location l_i is

$$P\left(l_i \middle| \hat{\theta}\right) = \frac{\exp\left(\beta_i^T \hat{\theta}\right)}{\sum_{j=1}^{L} \exp\left(\beta_j^T \hat{\theta}\right)}$$

Tweet Location Prediction Architecture

- Cross entropy Loss

$$L = -\sum_i \sum_k I(l_i = k) log P(l_i = k | x_i)$$

- Using gradient descent minimize the loss function with respect to:
  - The word vector for all the words
  - Parameter w and b in the convolutional layer.
  - Parameter β in the fully connected layer.

Outline

- Introduction and problem description
- Our method:
  - Useful Features
  - Our neural network architecture
- Experiments:
  - Country-level prediction
  - City-level prediction
- A case study

Experiments

- Data collection: geo-tagged tweets from geo-bounding box [-180, -90, 180, 90] from Jan. 7, 2017 to Feb. 1, 2017.

# of tweets	# of users	# of timezones	# of lang.	# of countries (or regions)	Tweets per country	# of cities	Tweets per city
4645692	3321194	417	103	243	19118.0 (99697.1)	3709	1252.5(4184.5)

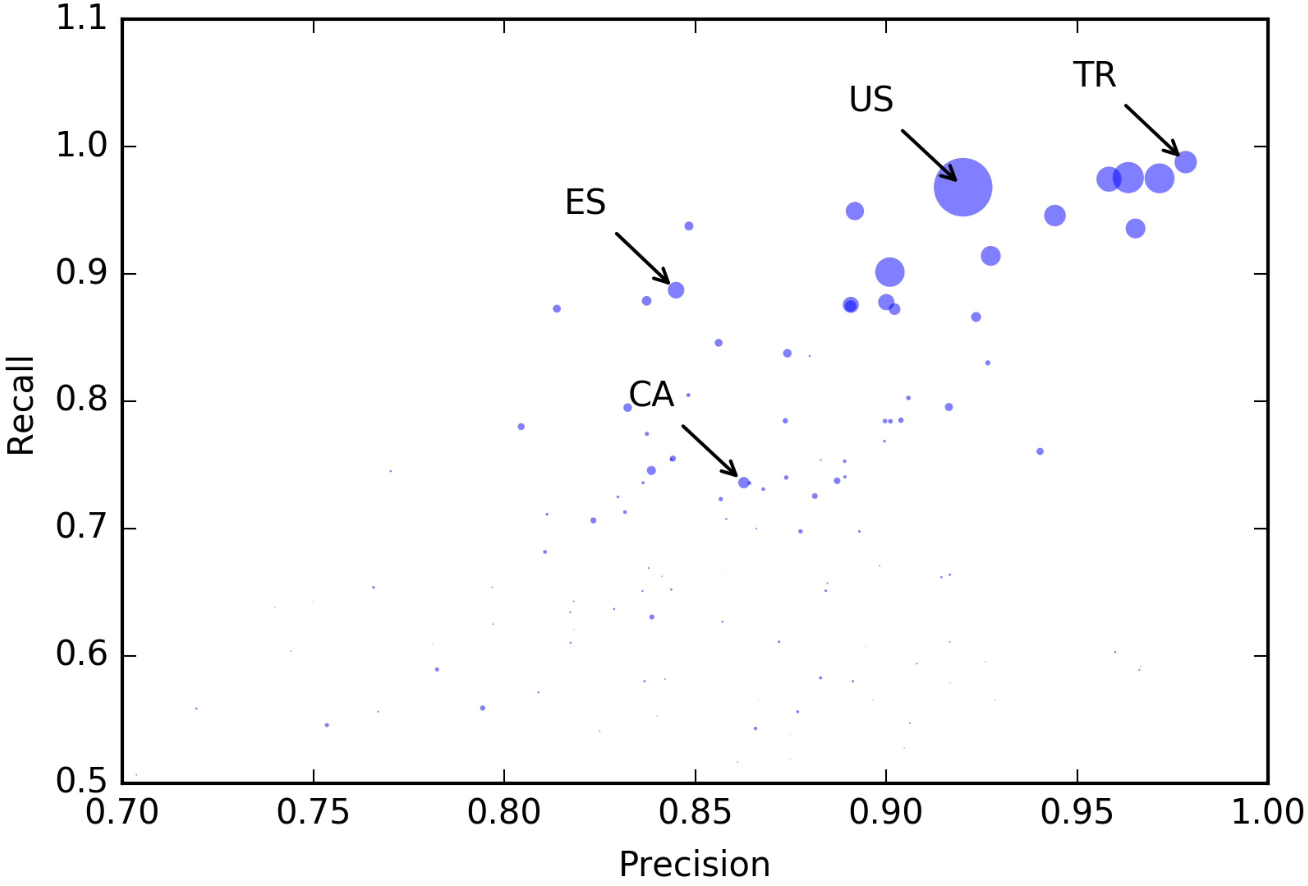
- We randomly selected one tweet for each user-city pair.
- We used 10% users as testing data, 90% users for training. We picked 50000 users in training data as development set to tune hyperparameter.

Country-level prediction

Each geo-tagged tweet has a "country_code" field.

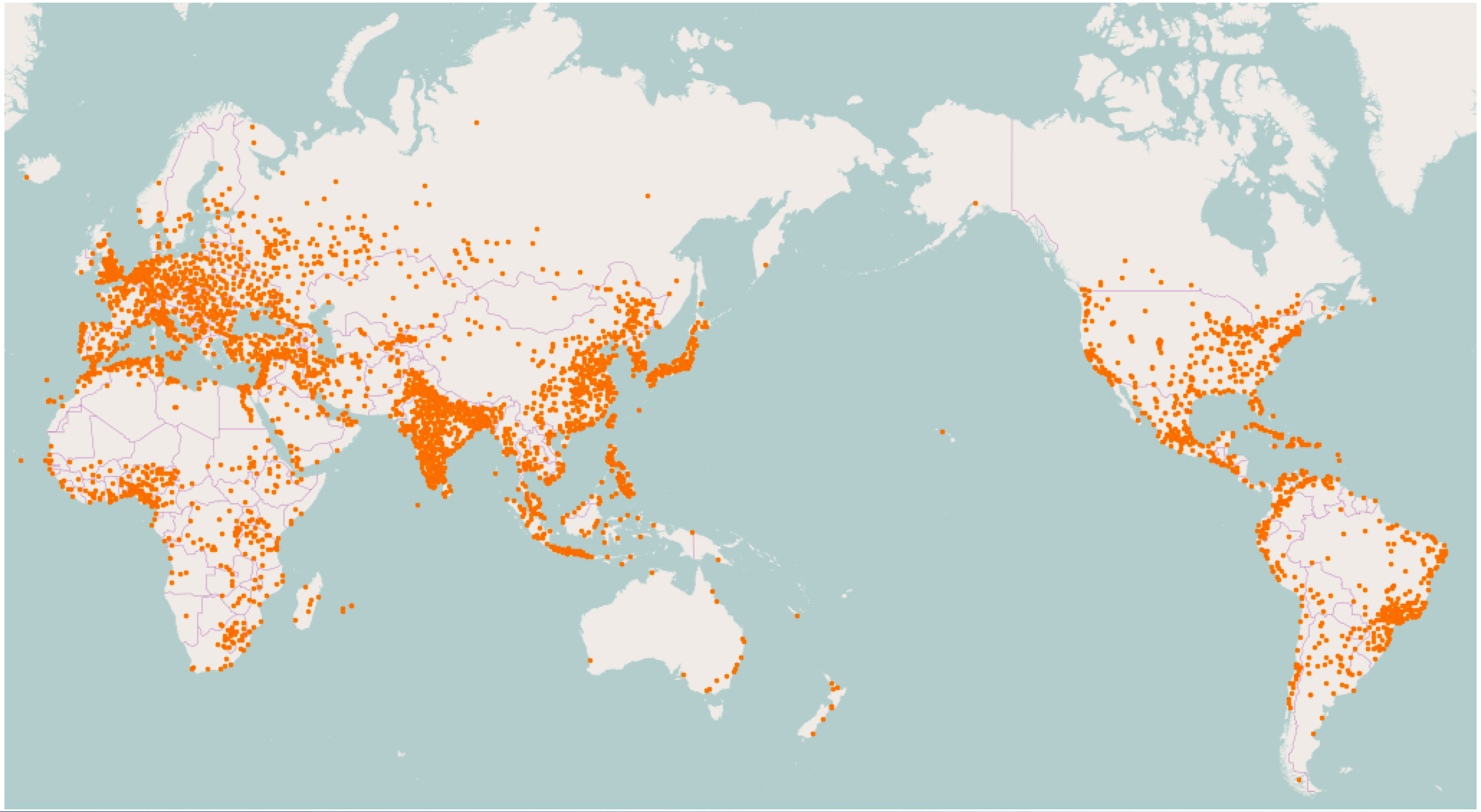
Table 3. Country prediction results.

	Acc	Acc@Top5
STACKING[2]	0.868	0.947
STACKING+	0.871	0.950
Our approach	**0.921**	**0.972**

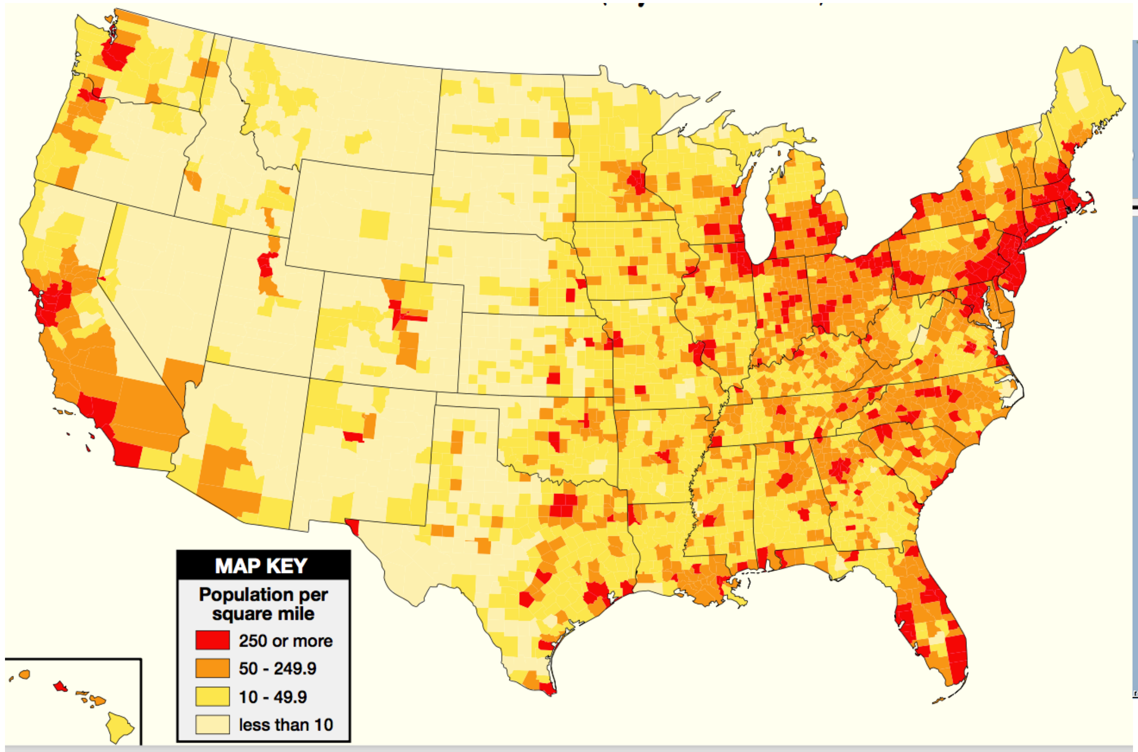


City-level prediction

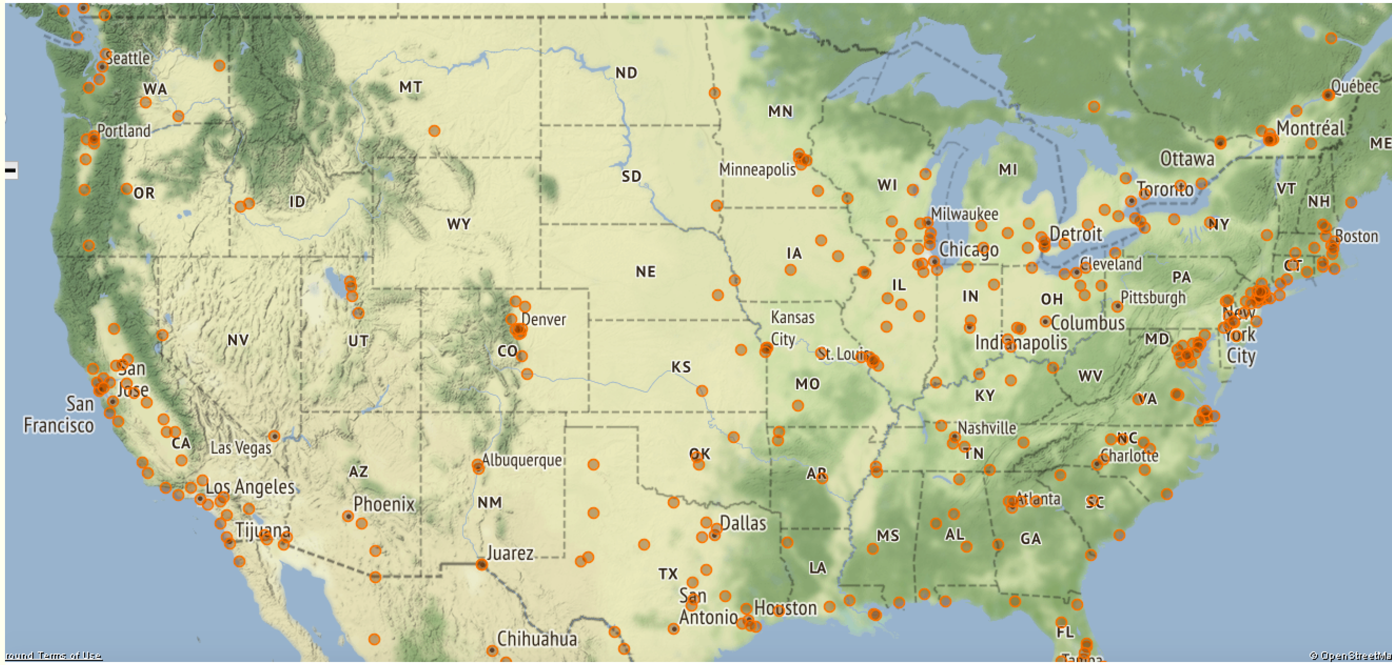
- Target cities: 3709 cities selected based on population[2]



City-level prediction



US Census



Target city distribution in US

City-level prediction

- Acc@161: The percentage of predicted city which are within a 161km(100 mile) radius of the true coordinates of original tweet
- Median: The median distance from the predicted city to the true coordinates

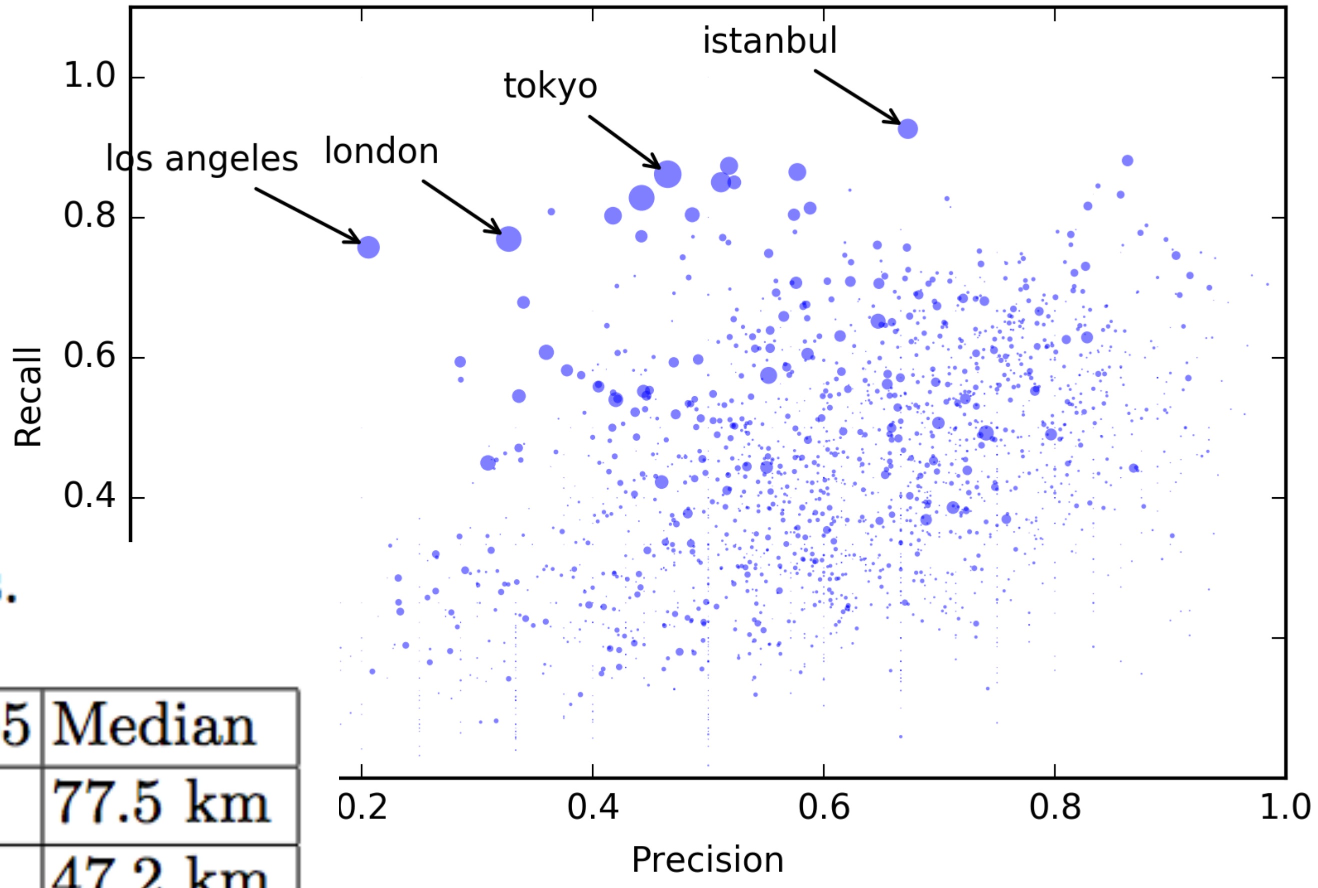
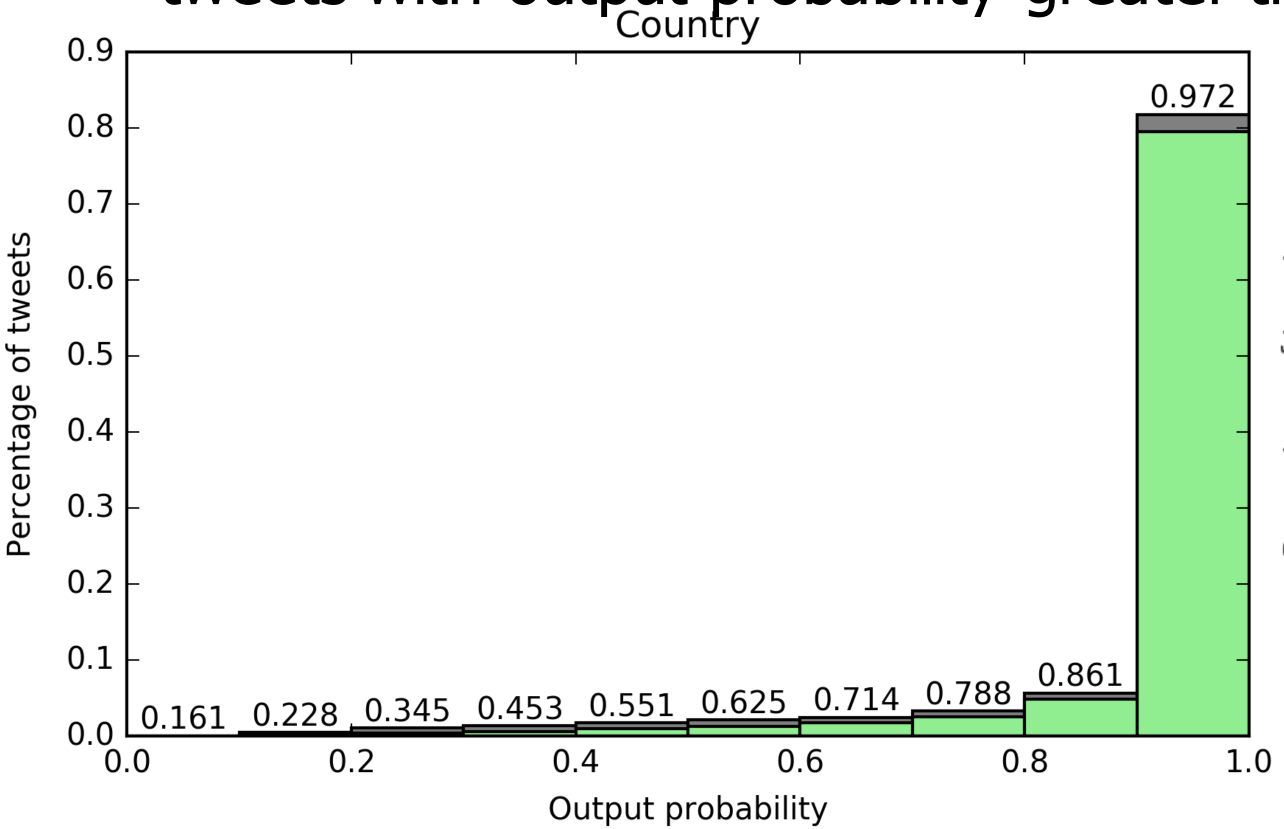


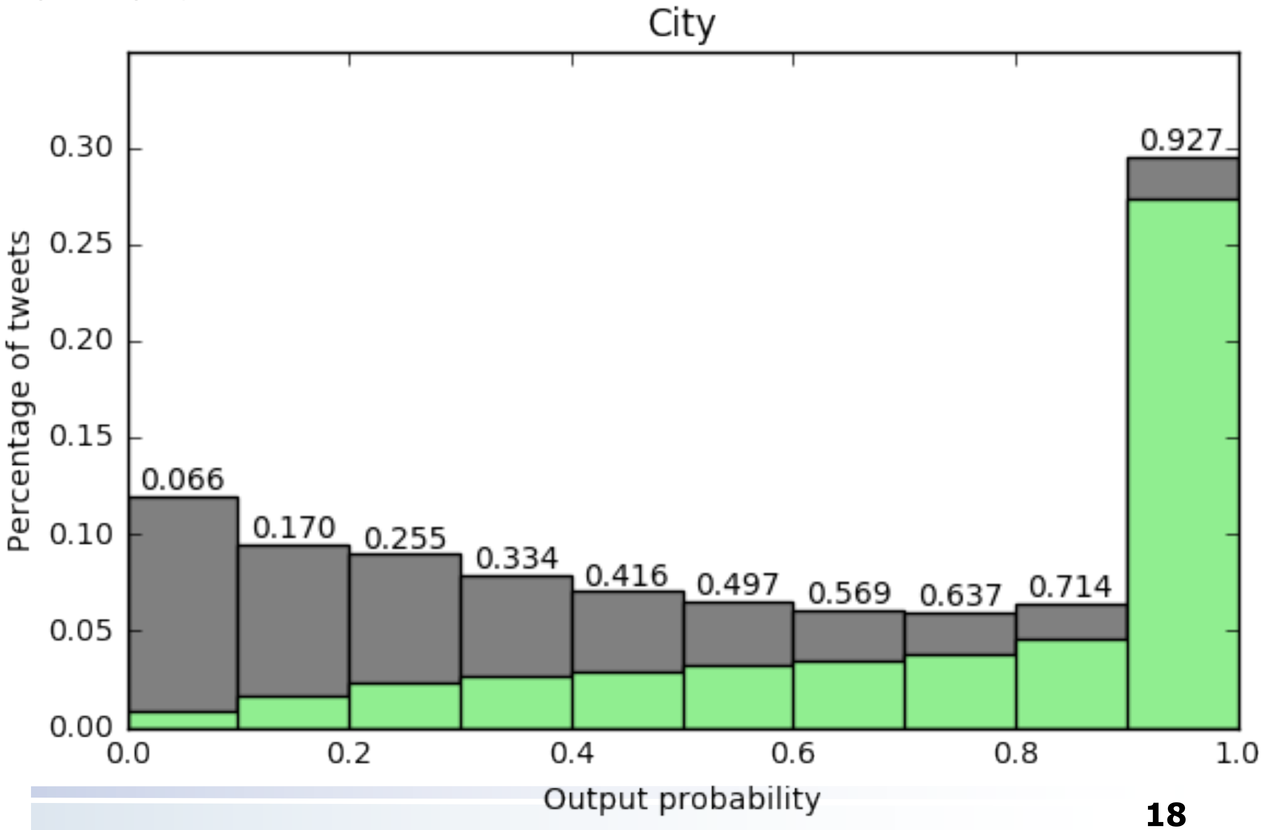
Table 4. City prediction results.

	Acc	Acc@161	Acc@Top5	Median
STACKING	0.389	0.573	0.595	77.5 km
STACKING+	0.439	0.616	0.629	47.2 km
Our approach	**0.528**	**0.692**	**0.711**	**28.0 km**

Application Scenario

1. We get 97.2% accuracy for country-level prediction with output probability larger than 0.9.
2. Surprisingly, the accuracy of city-level is as high as 92.7% for the 29.6% of the tweets with output probability greater than 0.9.



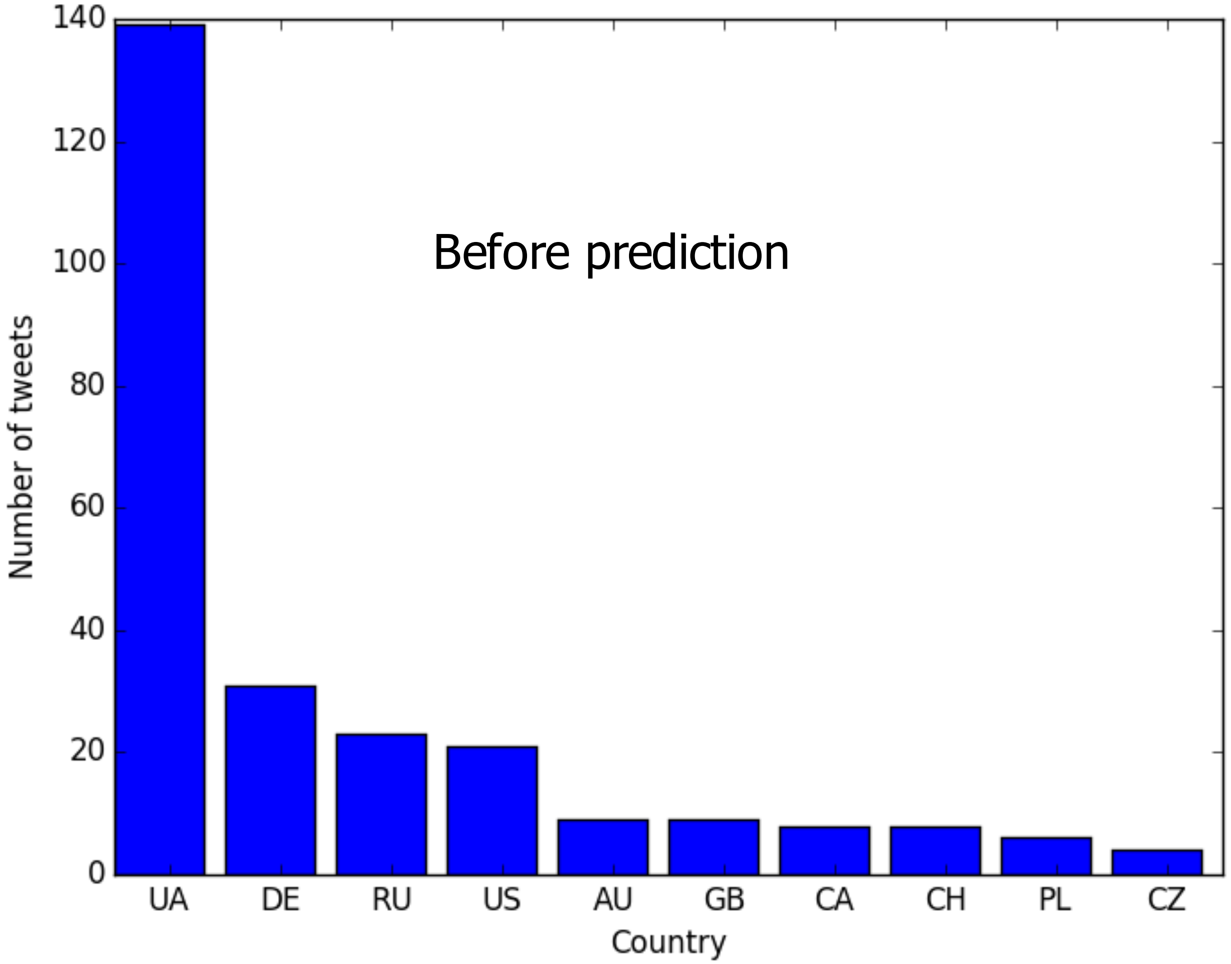


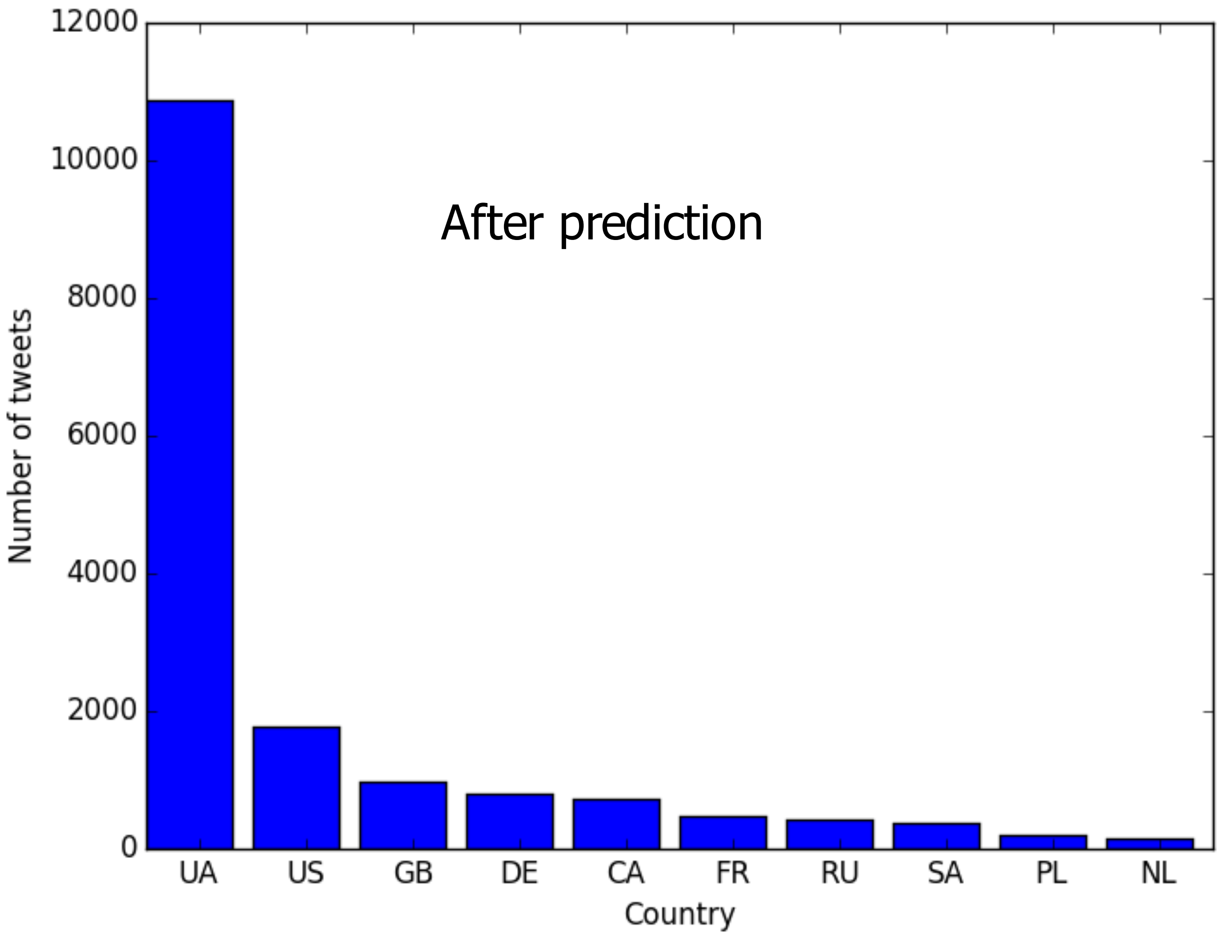
Outline

- Introduction and problem description
- Our method:
  - Useful Features
  - Our neural network architecture
- Experiments:
  - Country-level prediction
  - City-level prediction
- A case study

A case study on an Ukraine data

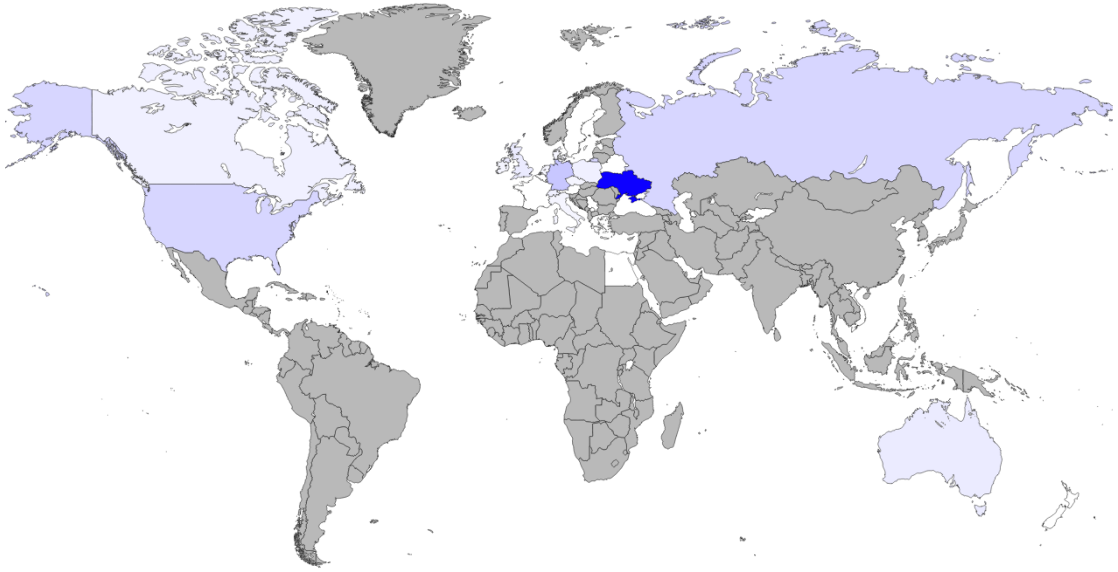
- Data is collected on Twitter by a keyword search. There are 18297 tweets in total and 292 of them are geo-tagged.



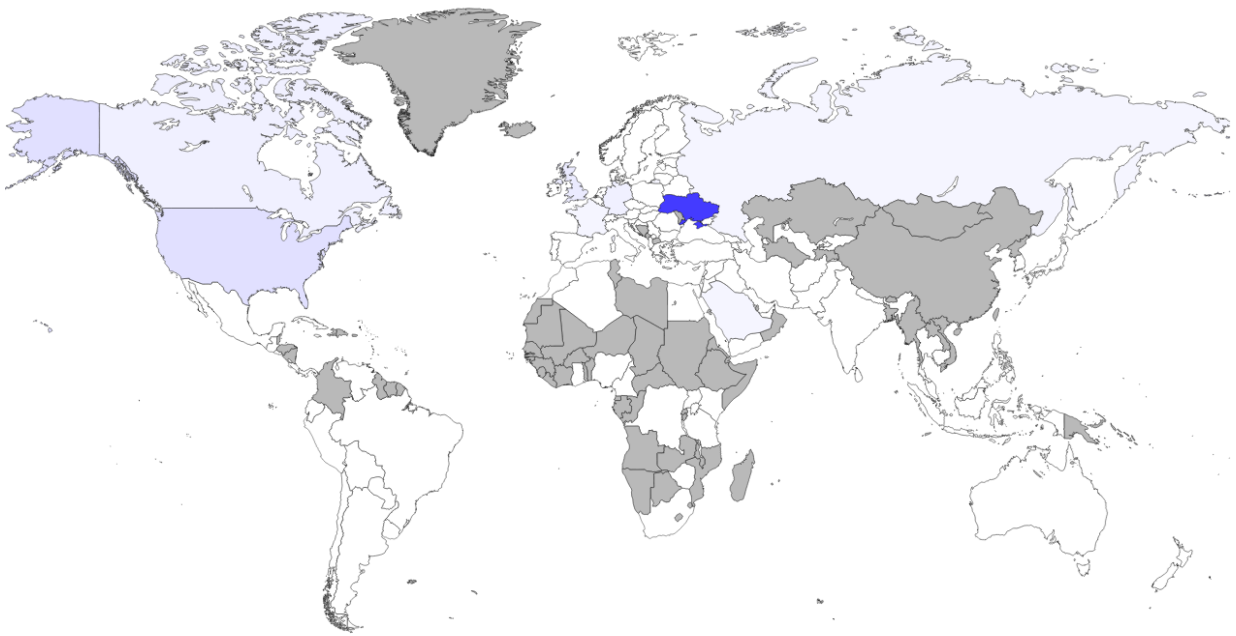


Heatmap of tweets

Before location prediction

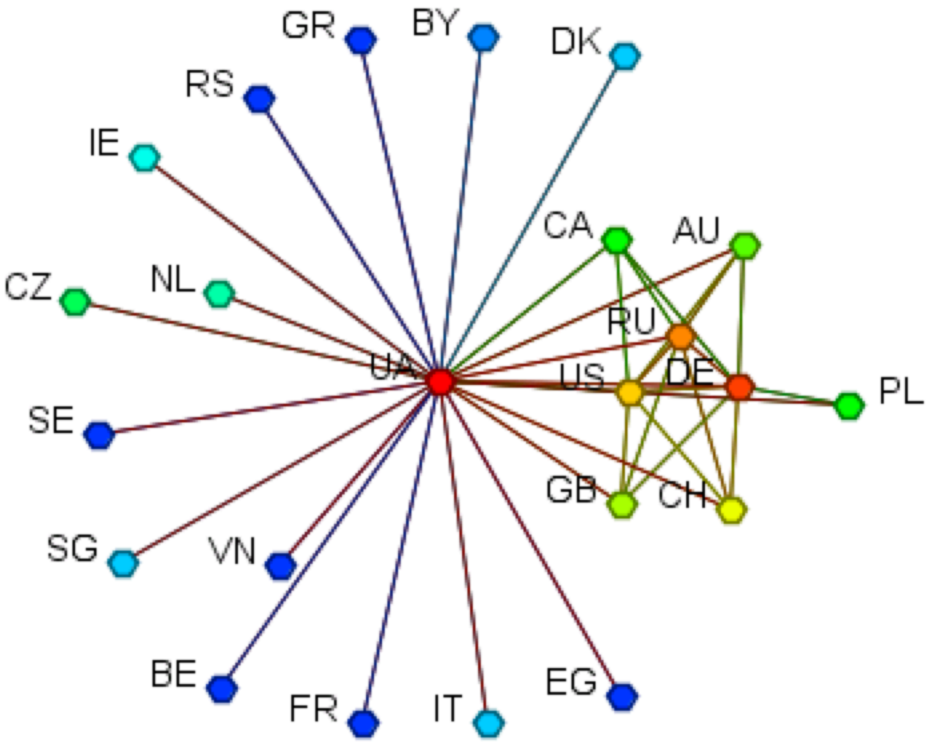


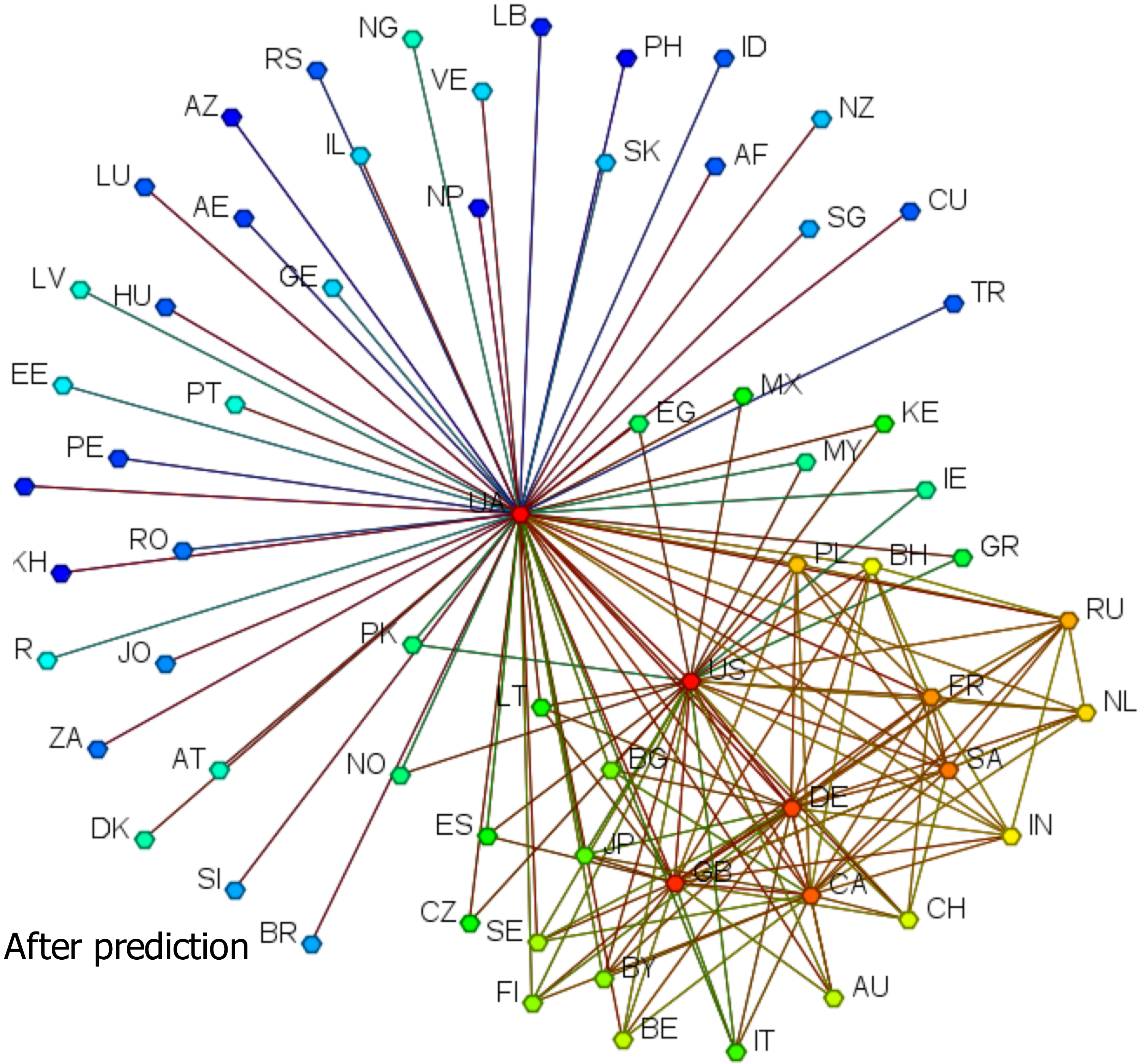
After location prediction



Country co-hashtag network

Before prediction





After prediction

Thank you!

- [1] Kim, Yoon. "Convolutional neural networks for sentence classification." *arXiv preprint arXiv:1408.5882* (2014).
- [2] Han, Bo, Paul Cook, and Timothy Baldwin. "A Stacking-based Approach to Twitter User Geolocation Prediction." *ACL (Conference System Demonstrations)*. 2013.
- [3] https://dev.twitter.com/streaming/overview/request-parameters#locations
- [4] Hale, S., Gaffney, D., Graham, M.: Where in the world are you? geolocation and language identification in twitter. Proceedings of ICWSM 12, 518–521 (2012)