Twitter User Geolocation Using Deep Multiview Learning

Tien Huu Do, Duc Minh Nguyen, Evaggelia Tsiligiani, Bruno Cornelis, Nikos Deligiannis

Vrije Universiteit Brussel – imec, Belgium
Social Networks and Location of Users

- Location of users enable many applications
- User location profile information might be missed or ambiguous: e.g. “Small town”, “Everywhere”
- ~3% of tweets are geo-tagged [3]

The Tasks of Twitter User Geolocation

- Region classification: *Northeast, Midwest, West,* and *South*
- State classification: 50 *states*
- Geo-coordinates prediction: (latitude, longitude)

Region and state boundaries are from the US census shape files
Our Approaches

Content-based: Tweets are used for location prediction

Network-based: Online relationships (e.g. following, mentioning) are used for location prediction

“Congratulations to San Francisco’s Andrew Sean Greer and Compton’s Kendrick Lamar on earning Pulitzer Prizes for fiction and music”
Learning from Multiple Views

- Processing: Tweets from the same user are concatenated making up a tweet document

- Feature extraction:
  - Individual word level: Term frequency-inverse document frequency (TF-IDF)
  - Semantic level: Doc2vec
  - User connection structure: Node2vec
  - Metadata: Posting timestamps of tweets
User Representation as Node Embedding

- Sequences of node indices are sampled using Random Walk [7]
- Node sequences are the input to a simple neural network similar to word2vec [8]
- Node embeddings are trained using SGD

Congrats to @USER_2 and Sister Jean for a last-second upset - I had faith in my pick!

Coming to #tryswiftnc all the way from US... please give a hand to @User_X

@User_1, @User_3 😊 lol. Saying ok to both

@User_X How are you?

User_X
MENET: Proposed Architecture

TF-IDF → $h_{11}$ → Concatenate

Doc2Vec → $h_{12}$

Node2Vec → $h_{13}$

Timestamp → $h_{14}$

Fully connected layer

Region/State probabilities

ReLU

$\text{max}(0,c)$

$\frac{\sum_{k=1}^{K} e^{z_k}}{\sum_{k=1}^{K} e^{z_k}}$
From Classification to Regression

1. Predict the state label
2. Predict geographical coordinates using the centroid of the state
3. State centroid = median \{[\text{latitude, longitude}]\}
4. The centroid coordinates are calculated from the geographical coordinates available in the training set
Performance criteria

- Region and state classification: **Accuracy (%)**
- Geographical coordinates prediction:
  - Mean distance error (km)
  - Median distance error (km)
  - Accuracy within 161 km (~100 miles) or @161 (%)
- The distance between two locations is computed using the **Haversine** formula

\[
\begin{align*}
a &= \sin^2(\Delta\phi/2) + \cos \phi_1 \cdot \cos \phi_2 \cdot \sin^2(\Delta\lambda/2) \\
c &= 2 \cdot \text{atan2}(\sqrt{a}, \sqrt{1-a}) \\
d &= R \cdot c
\end{align*}
\]

φ: Latitude
λ: Longitude
R: The Earth’s radius
Experimental Results

Table 1. Region and state classification result on GeoText\textsuperscript{[1]} and UTGeo2011\textsuperscript{[4]}

<table>
<thead>
<tr>
<th></th>
<th>GeoText Region (%)</th>
<th>GeoText State (%)</th>
<th>UTGeo2011 Region (%)</th>
<th>UTGeo2011 State (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eisenstein \textit{et al.}\textsuperscript{[1]}</td>
<td>58</td>
<td>27</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Liu &amp; Inkpen \textsuperscript{[2]}</td>
<td>61.1</td>
<td>34.8</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Cha \textit{et al.}\textsuperscript{[3]}</td>
<td>67</td>
<td>41</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>MENET</td>
<td>76</td>
<td>64.8</td>
<td>83.7</td>
<td>69</td>
</tr>
</tbody>
</table>

- 9\% improvement for region classification
- 23.8\% improvement for state classification
## Experimental Results

Table 2. **Geo-coordinates prediction** on GeoText\(^1\) and UTGeo2011\(^4\)

<table>
<thead>
<tr>
<th>Method</th>
<th>GeoText mean (km)</th>
<th>GeoText median (km)</th>
<th>GeoText @161 (%)</th>
<th>UTGeo2011 mean (km)</th>
<th>UTGeo2011 median (km)</th>
<th>UTGeo2011 @161 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eisenstein <em>et al.</em>  (^1)</td>
<td>900</td>
<td>494</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Roller <em>et al.</em> (^4)</td>
<td>897</td>
<td>432</td>
<td>35.9</td>
<td>860</td>
<td>463</td>
<td>34.6</td>
</tr>
<tr>
<td>Liu and Inkpen (^2)</td>
<td>855.9</td>
<td>N/A</td>
<td>N/A</td>
<td>733</td>
<td>377</td>
<td>24.2</td>
</tr>
<tr>
<td>Cha <em>et al.</em> (^3)</td>
<td>581</td>
<td>425</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Rahimi <em>et al.</em> (2015) (^5)</td>
<td>581</td>
<td>57</td>
<td>59</td>
<td>529</td>
<td>78</td>
<td>60</td>
</tr>
<tr>
<td>Rahimi <em>et al.</em> (2017) (^6)</td>
<td>578</td>
<td>61</td>
<td>59</td>
<td>515</td>
<td>77</td>
<td>61</td>
</tr>
<tr>
<td><strong>MENET</strong></td>
<td><strong>570</strong></td>
<td><strong>58</strong></td>
<td><strong>59.1</strong></td>
<td><strong>474</strong></td>
<td><strong>157</strong></td>
<td><strong>50.5</strong></td>
</tr>
</tbody>
</table>
Conclusion

- Twitter user geo-location is challenging due to noisy data.
- Combine the content and network features can improve the geo-location accuracy.
- Multi-view learning can exploit different views of Twitter data for location prediction.
- The proposed architecture can be extended with different types of features or by adding more hidden layers.
- The distribution of Twitter users will be considered in the future work.
References


References


Thank you for your attention!

{ thdo, mdnguyen, etsiligi, bcorneli, ndeligia }@etrovub.be