Twitter User Geolocation Using Deep Multiview Learning

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

Vrije Universiteit Brussel – imec , Belgium

VRIJE UNIVERSITEIT BRUSSEL

imec

Social Networks and Location of Users

- 2.2B active users
- 330M active users
- 255M active users

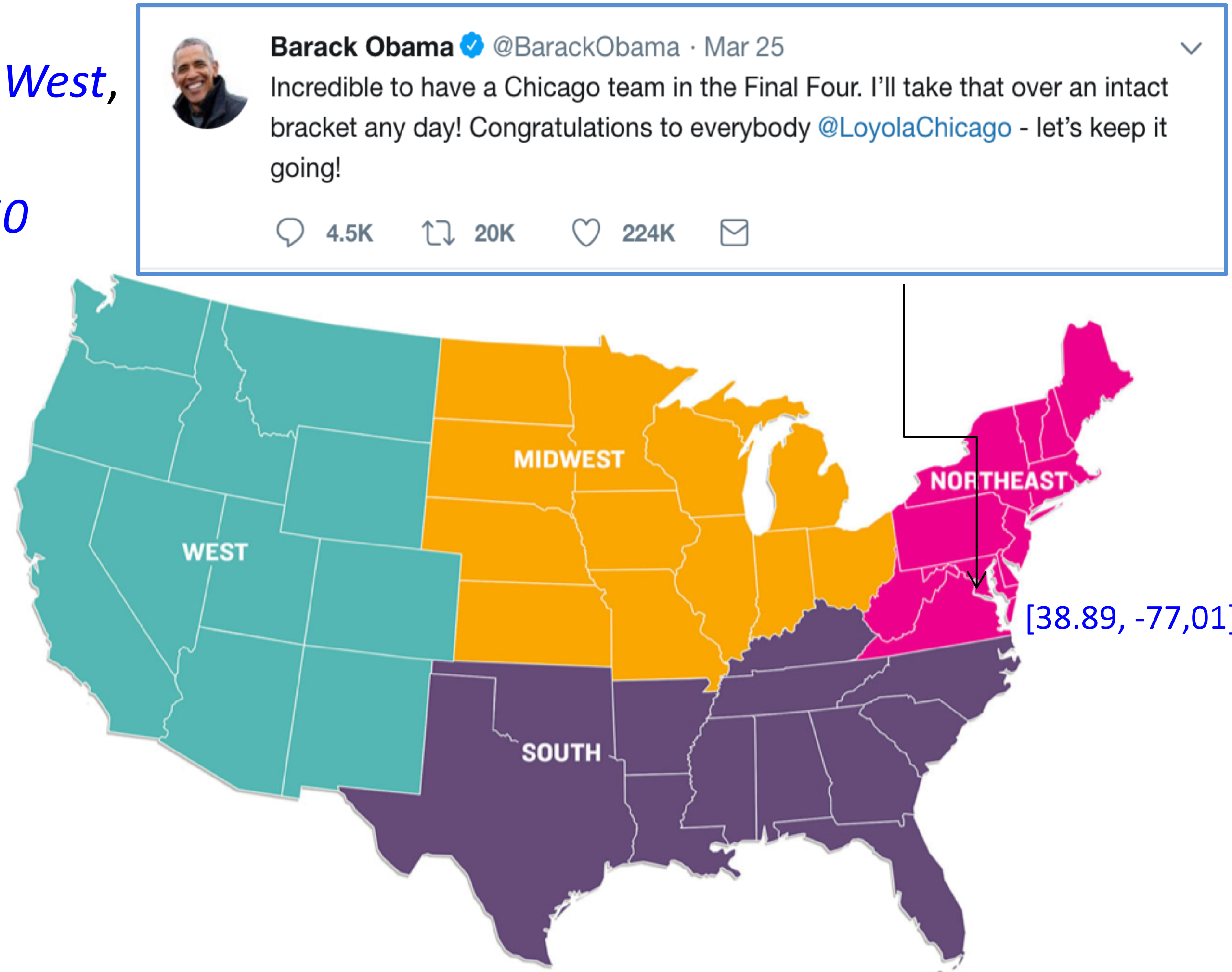
- Location of users enable many applications
- User location profile information might be missing or ambiguous: e.g. "Small town", "Everywhere"
- ~3% of tweets are geo-tagged [3]

Reference: https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/

VRIJE UNIVERSITEIT BRUSSEL

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

VRIJE UNIVERSITEIT BRUSSEL
imec

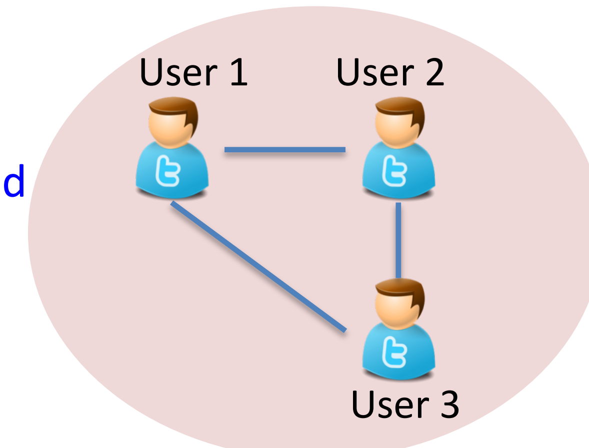
Our Approaches

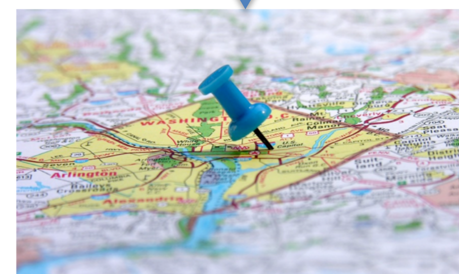
Our approach

Content-based

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

Network-based





Metadata

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

VRIJE UNIVERSITEIT BRUSSEL

imec

Data Processing and Feature Extraction

- Processing:
  - Tokenization
  - Stop-word removal
  - Stemming
  - 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*[8]
  - User connection structure: *Node2vec*[7]
  - Metadata: Posting *timestamps* of tweets

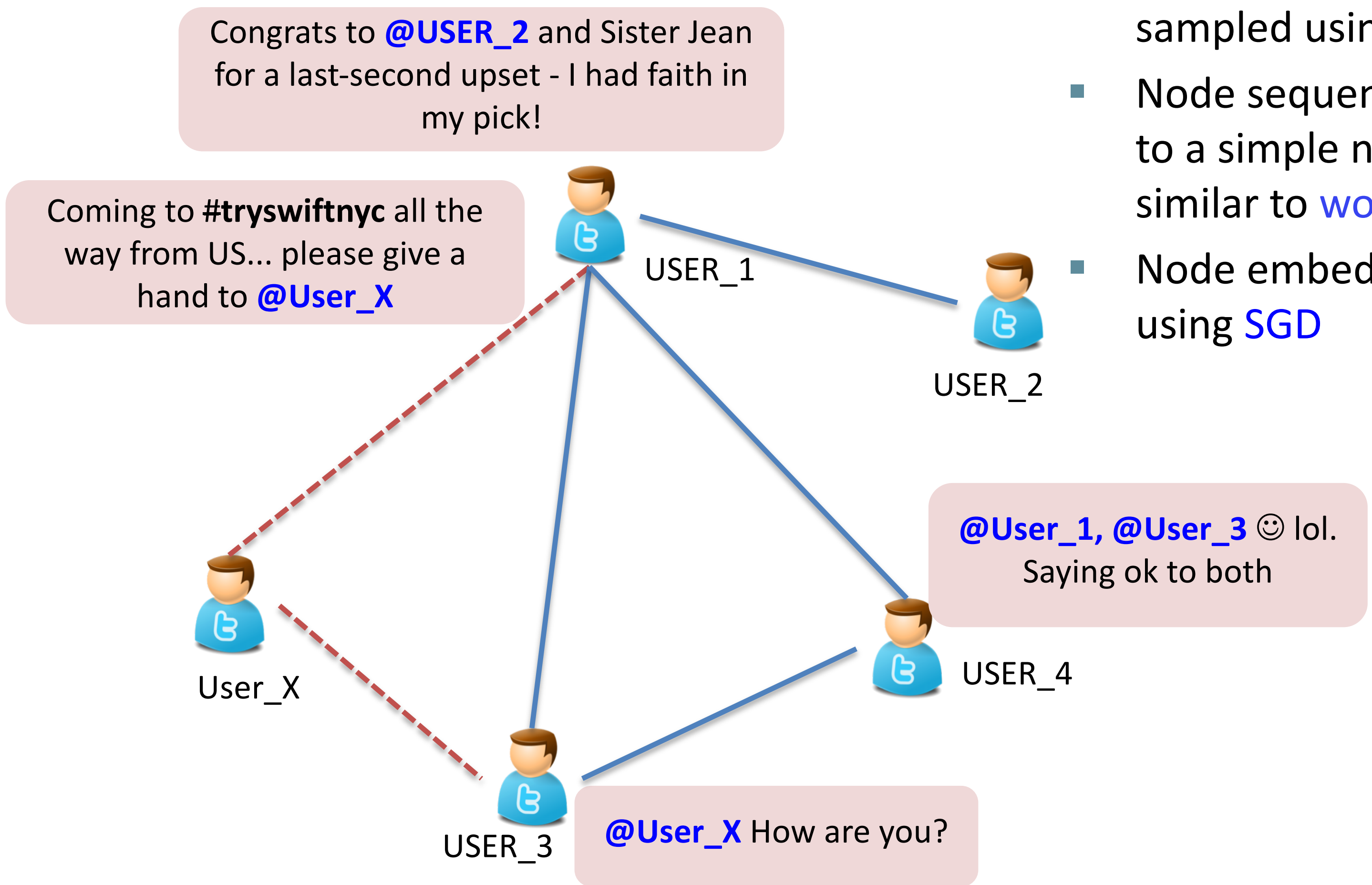
Content features

Network feature

VRIJE UNIVERSITEIT BRUSSEL

imec

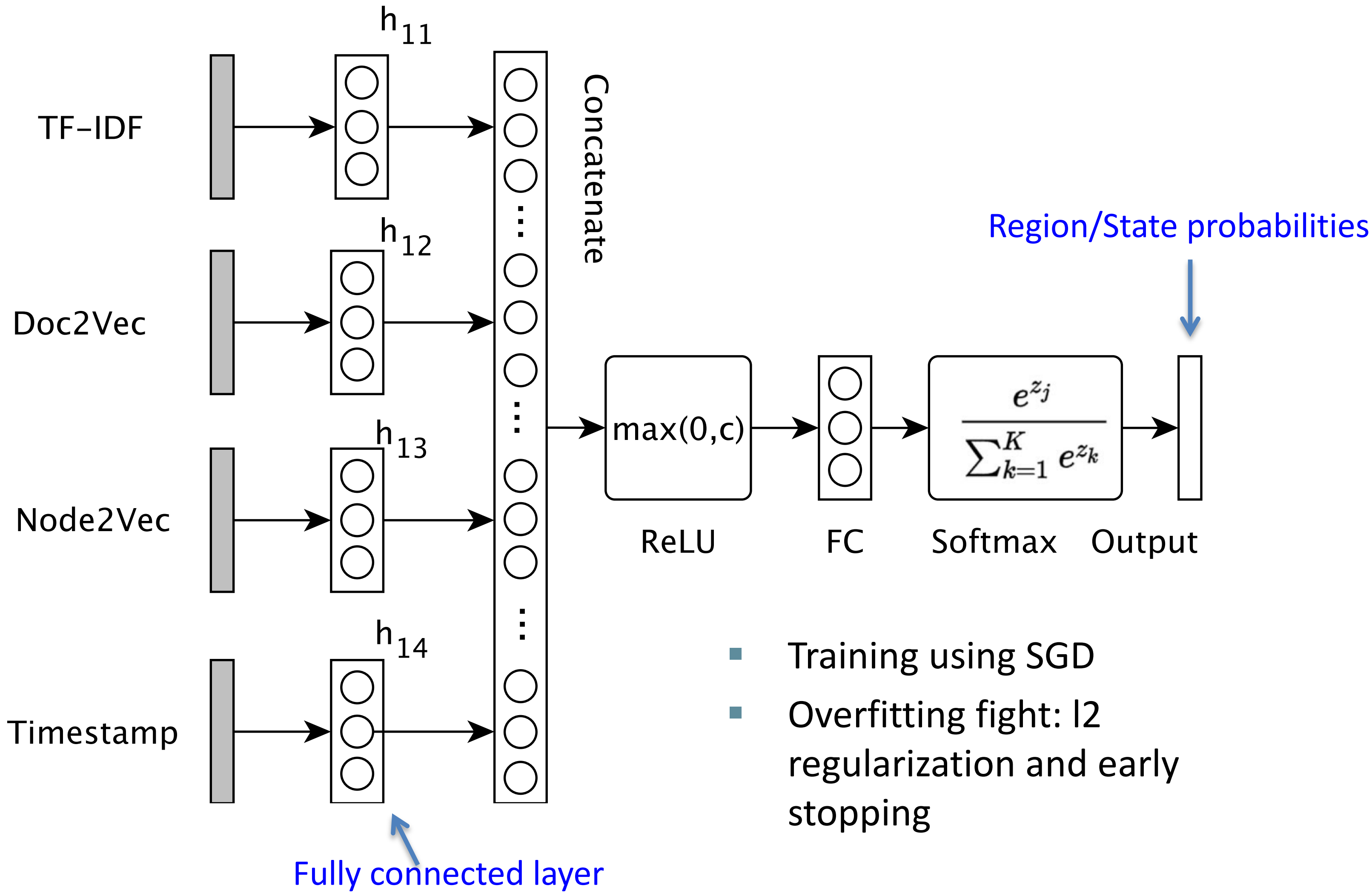
User Representation as Node Embedding



- 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

VRIJE UNIVERSITEIT BRUSSEL · imec

MENET: Proposed Architecture

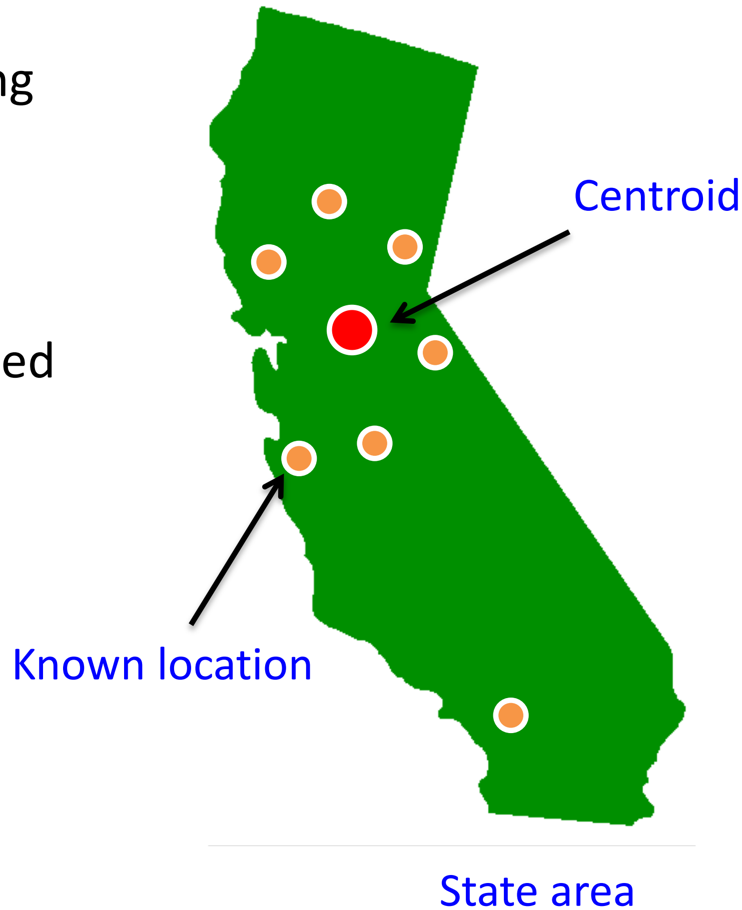


- Training using SGD
- Overfitting fight: l2 regularization and early stopping

VRIJE UNIVERSITEIT BRUSSEL imec

From Classification to Regression

1. Predict the state label
2. Predict geographical coordinates using the centroid of the state
3. State centroid = median {[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

$$a = \sin^2(\Delta\varphi/2) + \cos\varphi_1 \cdot \cos\varphi_2 \cdot \sin^2(\Delta\lambda/2)$$

$$c = 2 \cdot \operatorname{atan2}(\sqrt{a}, \sqrt{(1-a)})$$

$$d = R \cdot c$$

ϕ: Latitude
λ: Longitude
R: The Earth's radius

Experimental Results

Table 1. Region and state classification result on GeoText[1] and UTGeo2011[4]

	GeoText Region (%)	GeoText State (%)	UTGeo2011 Region (%)	UTGeo2011 State (%)
Eisenstein *et al.* [1]	58	27	N/A	N/A
Liu & Inkpen [2]	61.1	34.8	N/A	N/A
Cha *et al.* [3]	67	41	N/A	N/A
MENET	**76**	**64.8**	83.7	69

- 9% improvement for region classification
- 23.8% improvement for state classification

VRIJE UNIVERSITEIT BRUSSEL

Experimental Results

Table 2. Geo-coordinates prediction on GeoText[1] and UTGeo2011[4]

	GeoText			UTGeo2011		
	mean (km)	median (km)	@161 (%)	mean (km)	median (km)	@161 (%)
Eisenstein *et al.* [1]	900	494	N/A	N/A	N/A	N/A
Roller *et al.* [4]	897	432	35.9	860	463	34.6
Liu and Inkpen [2]	855.9	N/A	N/A	733	377	24.2
Cha *et al.* [3]	581	425	N/A	N/A	N/A	N/A
Rahimi *et al.* (2015) [5]	581	**57**	59	529	78	60
Rahimi *et al.* (2017) [6]	578	61	59	515	**77**	**61**
MENET	**570**	58	**59.1**	**474**	157	50.5

VRIJE UNIVERSITEIT BRUSSEL

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.
- Considering distribution of Twitter users can improve the geolocation accuracy.

References

1. J. Eisenstein, B. O'Connor, N. A. Smith, and E. P. Xing, "A latent variable model for geographic lexical variation", in *Conference on Empirical Methods in Natural Language Process ing*, 2010, pp. 1277–1287.
2. J. Liu and D. Inkpen, "Estimating user location in social media with stacked denoising auto-encoders", in *Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies*, 2015, pp. 201– 210.
3. M.Cha, Y.Gwon, and H.T.Kung, "Twitter geolocationandre- gional classification via sparse coding", in *International AAAI Conference on Web and Social Media*, 2015, pp. 582–585.
4. S. Roller, M. Speriosu, S. Rallapalli, B. Wing, and J. Baldridge, "Supervised text-based geolocation using language models on an adaptive grid", in *Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 2012, pp. 1500–1510.

VRIJE UNIVERSITEIT BRUSSEL

References

5. A.Rahimi, T.Cohn, and T.Baldwin, "Twitter user geolocation using a unified text and network prediction model", in *Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing*, 2015, pp. 630–636.
6. A. Rahimi, T. Cohn, and T. Baldwin, "A neural model for user geolocation and lexical dialectology", in *Annual Meeting of the Association for Computational Linguistics*, 2017, pp. 209–216.
7. A. Grover and J. Leskovec, "node2vec: Scalable feature learning for networks", in *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 855-864
8. T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality", in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

VRIJE UNIVERSITEIT BRUSSEL

Thank you for your attention !

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

VRIJE
UNIVERSITEIT
BRUSSEL