2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining

TF-MF: Improving Multiview Representation for Twitter User Geolocation Prediction

Parham Hamouni
Graph Analytics Labs Inc.
Toronto, Canada
parham@graphanalytics.ai

Taraneh Khazaei
Aggregate Intellect Inc.
Toronto, Canada
tara@a-i.science

Ehsan Amjadian
Aggregate Intellect Inc.
Toronto, Canada
ehsan@a-i.science

***Abstract*—Twitter user geolocation detection can inform and benefit a range of downstream geospatial tasks such as event and venue recommendation, local search, and crisis planning and response. In this paper, we take into account user shared tweets as well as their social network, and run extensive comparative studies to systematically analyze the impact of a variety of language-based, network-based, and hybrid methods in predicting user geolocation. In particular, we evaluate different text representation methods to construct text views that capture the linguistic signals available in tweets that are specific to and indicative of geographical locations. In addition, we investigate a range of network-based methods, such as embedding approaches and graph neural networks, in predicting user geolocation based on user interaction network. Our findings provide valuable insights into the design of effective and efficient geolocation identification engines. Finally, our best model, called TF-MF, substantially outperforms state-of-the-art approaches under minimal supervision.**

***Index Terms*—Twitter, social computing, neural networks, graph, geolocation**

I. Introduction

Identification of user geolocation on social media can enable and support a wide range of downstream tasks [1] including public health surveillance [2], targeted ads and recommendations [3], [4]. Therefore, identifying user primary geolocation in online social platforms has gained significant attention from the research community in the past decade.

To model geolocation using social media platforms such as Twitter, since social media is an unstructrued data [5], learning useful representations for any downstream machine learning algorithm can be challenging. Previous work on Twitter geolocation prediction try to structure the signal by three directions of approaches, namely content-based, network-based, and multi-view techniques [6]. Content-based studies focus on user tweets and rely on linguistic signals and attributes that are specific to geographical locations. Assuming that people who live near each other are more likely to interact, network-based approaches study and analyze the social context signal of a user to predict their geolocation. This assumption is called *location homophily*. Multi-view approaches attempt to boost the predictive force of the model by taking into account both the content and the network views simultaneously. In fact, the best performance in Twitter geolocation prediction is obtained by such hybrid methods [6], [7]. In this study, we run extensive comparative experiments to capture the effectiveness of all of the possible views, that is; only content-based representation, only network-based representation and multiview representation. Since only a small portion (about 1% of Twitter data) is geotagged in real life [3], we focus our study to improve the performance of the task under real life scenario.

ASONAM '19, August 27-30, 2019, Vancouver, Canada

ACM ISBN 978-1-4503-6868-1/19/08
https://doi.org/10.1145/3341161.3342961

II. Methodology

We used one of the benchmark datasets for Twitter user geolocation identification called GEO-TEXT [6], [8], This data set consists of 9475 users that are tagged with their geolocation via latitude/longitude coordinates. These geographic coordinates are assigned to users based on their first geo-tagged tweet. The user tweet timelines[1] are collected by retrieving user tweets up to the data collection point. We pre-processed user timelines to create an adjacency matrix of user-user interactions based the mentions in their tweets. To generate discrete geo-labels, we applied k-d tree [9] to group nearby coordinates into buckets. This process resulted in 129 unique geo-labels. These two pre-processing steps are the same as the steps taken by Rahimi et al. [6], allowing us to directly compare the results between the two studies. Let us formalize the multi-view geolocation prediction task in mathematical terms:

- $d \in R^{|U|\times|V|}$ corresponds to the linguistic content (document) representation of each user U based on a vector of size $|V|$.
- $A \in 1^{|U|\times|U|}$ corresponds to the adjacency matrix of users U created based on user-user interaction network. We build this matrix based on the @-mention network between users, where two users are connected ($A_{ij} = 1$) if one mentions the other.

[1]A user timeline is comprised of a chronological concatenation of the user's tweets.

- We used k-d tree to discretize the geographical coordinates. We then assigned a geo-label to each location using a one-hot encoded vector $y_i \in 1^c$ where c is the number of target buckets.

To represent user content view (X), we employ TF-IDF [10] and doc2vec [11] and apply it to the Twitter timeline d. The goal of the model is to find a mapping from each user content and network views (X and A) to their location bucket (y). To capture network information, we experiment with node2vec [12], NetMF [13], and GraphSAGE [14].

III. Experiments, Results, and Analysis

We experimented with ten different multi-view models and investigate the effect of different techniques to capture the language and network views. A subset of these models utilize content and network embedding approaches to create vector representations of these views and pass them to a feed-forward neural network for geolocation prediction. By controlling the layer size and the depth of the feed-forward neural network, we systematically investigate the contributions of each these view (i.e., content and network) to the task of geolocation detection. Additionally, we passed both content and network representations simultaneously to the same feed-forward network to investigate the interaction of these two views. Furthermore, we compare these embedding-based models with an end-to-end graph convolutional network. A summary of the models we studied can be found in Table I.

Similar to earlier geolocation prediction work [6]–[8], we evaluated the models by calculating the distance between the center of the predicted label and the true coordinates as error, median error, and mean error. We also report accuracy 161, denoted by Acc@161, which is the accuracy of predicting a user within 161km or 100 miles of their location.

Table II provides a detailed comparison of the results of these models with the previous work on the geolocation prediction task under minimal and sufficient supervision. As the results suggest, our best model (TF-MF) obtains comparable results under sufficient supervision. However, TF-MF drastically outperforms the previous state-of-the-art method under minimal supervision, which is a more accurate reflection of the real-world scenario. We obtained this result by systematically analyzing content and network signals in the geolocation task and by exploiting the best model according to the analysis.

IV. Conclusion

We provided an extensive exploration of Twitter user geolocation prediction using language and network-based attributes. When identifying Twitter user geolocation, our results suggest that user network information carries stronger signals about their geolocation compared to their linguistic content. We also show that a traditional relative frequency based text representation method (i.e., TF-IDF) can better capture geographic references in tweets compared to document embedding methods (i.e., doc2vec). Our comparison of network processing methods shows the superiority of a factorization-based network embedding method. Furthermore, it was shown that that *location homophily* is not only a valid assumption, but it also provides a more consistent representation than content-based signals. Finally, our best prediction model, TF-MF, outperformed the previous techniques in the literature under minimal supervision and obtained competitive results with sufficient supervision. These results expand our reach for practical real-life scenarios where sufficient supervision is often not feasible or even possible.

Our experiments are based on a dataset of 9K US-based Twitter users. To generalize these findings, conducting a similar analysis with larger datasets is warranted. We will also investigate the generalizability of the results by running the same experiments on non-English Twitter networks. Finally, we plan to explore other recent network analysis frameworks in our future work.

References

[1] X. Zheng, J. Han, and A. Sun, "A survey of location prediction on twitter," *IEEE Transactions on Knowledge and Data Engineering*, vol. 30, no. 9, pp. 1652–1671, 2018.

[2] S. Ribeiro and G. L. Pappa, "Strategies for combining twitter users geolocation methods," *GeoInformatica*, vol. 22, no. 3, pp. 563–587, 2018.

[3] Z. Cheng, J. Caverlee, and K. Lee, "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*. ACM, 2010, pp. 759–768.

[4] A. Noulas, S. Scellato, N. Lathia, and C. Mascolo, "Mining user mobility features for next place prediction in location-based services," in *2012 IEEE 12th international conference on data mining*. IEEE, 2012, pp. 1038–1043.

[5] G. Gautam and D. Yadav, "Sentiment analysis of twitter data using machine learning approaches and semantic analysis," in *2014 Seventh International Conference on Contemporary Computing (IC3)*. IEEE, 2014, pp. 437–442.

[6] A. Rahimi, T. Cohn, and T. Baldwin, "Semi-supervised user geolocation via graph convolutional networks," *arXiv preprint arXiv:1804.08049*, 2018.

[7] T. H. Do, D. M. Nguyen, E. Tsiligianni, B. Cornelis, and N. Deligiannis, "Multiview deep learning for predicting twitter users' location," *arXiv preprint arXiv:1712.08091*, 2017.

[8] J. Eisenstein, B. O'Connor, N. A. Smith, and E. P. Xing, "A latent variable model for geographic lexical variation," in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2010, pp. 1277–1287.

[9] S. Roller, M. Speriosu, S. Rallapalli, B. Wing, and J. Baldridge, "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*. Association for Computational Linguistics, 2012, pp. 1500–1510.

[10] K. Sparck Jones, "A statistical interpretation of term specificity and its application in retrieval," *Journal of documentation*, vol. 28, no. 1, pp. 11–21, 1972.

[11] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *International conference on machine learning*, 2014, pp. 1188–1196.

[12] A. Grover and J. Leskovec, "node2vec: Scalable feature learning for networks," in *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2016, pp. 855–864.

[13] 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 *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. ACM, 2018, pp. 459–467.

[14] W. Hamilton, Z. Ying, and J. Leskovec, "Inductive representation learning on large graphs," in *Advances in Neural Information Processing Systems*, 2017, pp. 1024–1034.

Model Name	Language Representation	Network Representation	Classifier	View
tf-nn	TF-IDF	-	Feed-Forward NN	Language
d2v-nn	doc2vec	-	Feed-Forward NN	Language
n2v-nn	-	node2vec	Feed-Forward NN	Network
mf-nn	-	NetMF	Feed-Forward NN	Network
tf-n2v	TF-IDF	node2vec	Feed-Forward NN	Hybrid
d2v-n2v	doc2vec	node2vec	Feed-Forward NN	Hybrid
TF-MF	TF-IDF	NetMF	Feed-Forward NN	Hybrid
d2v-mf	doc2vec	NetMF	Feed-Forward NN	Hybrid
GSM	-	GraphSAGE-Mean		Network
tf-GSM	TF-IDF	GraphSAGE-Mean		Hybrid
d2v-GSM	doc2vec	GraphSAGE-Mean		Hybrid

TABLE I

DIFFERENT ARCHITECTURES WHICH INVESTIGATED LANGUAGE REPRESENTATION, NETWORK REPRESENTATION, AND CLASSIFICATION MODELS.

Model	Acc@161	Mean	Median
GCN 1% [6]	6	1103	609
tf-GSM 1%	39	857	472
tf-n2v 1%	40	854	372
TF-MF 1%	**43**	**779**	**295**
GCN [6]	**60**	**546**	**45**
NN [15]	59	578	61
tf-GSM	55	857	83
tf-n2v	58	596	61
TF-MF	55	686	98

TABLE II

THE RESULTS OF OUR MODELS AS WELL AS THE PREVIOUS STATE-OF-THE-ART ON THE GEOTEXT DATASET UNDER MINIMAL (1% OF THE TRAINING DATA) AND SUFFICIENT SUPERVISION. AS THE RESULTS SHOW, OUR NOVEL METHOD OUTPERFORMS STATE-OF-THE-ART UNDER MINIMAL SUPERVISION.

[15] A. Rahimi, T. Cohn, and T. Baldwin, "A neural model for user geolocation and lexical dialectology," *arXiv preprint arXiv:1704.04008*, 2017.