TWITTER USER GEOLOCATION USING DEEP MULTIVIEW LEARNING

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

Vrije Universiteit Brussel, Pleinlaan 2, B-1050 Brussels, Belgium
imec, Kapeldreef 75, B-3001 Leuven, Belgium
{thdo, mdnguyen, etsiligi, bcorneli, ndeligia}@etrovub.be

ABSTRACT

Predicting the geographical location of users on social networks like Twitter is an active research topic with plenty of methods proposed so far. Most of the existing work follows either a content-based or a network-based approach. The former is based on user-generated content while the latter exploits the structure of the network of users. In this paper, we propose a more generic approach, which incorporates not only both content-based and network-based features, but also other available information into a unified model. Our approach, named Multi-Entry Neural Network (MENET), leverages the latest advances in deep learning and multiview learning. A realization of MENET with textual, network and metadata features results in an effective method for Twitter user geolocation, achieving the state of the art on two well-known datasets.

***Index Terms*—** Twitter user geolocation, multiview learning, deep learning, feature learning.

1. INTRODUCTION

Social networks have become more and more popular, with billions of active users on a daily basis. Among the most widely-used social networks, Twitter stands out as an attractive option, with a unique mechanism of publishing short messages, termed *tweets* and re-posting messages, termed *retweets*. This way information can be broadcasted widely and quickly through the Twitter network. As one of the most popular social networks, a lot of useful, yet unstructured, information is available on Twitter. User location is essential for a wide range of applications such as social unrest forecasting [1], event detection [2] and location-based service recommendation [3]. Nevertheless, the availability of geo-tagged tweets and geolocation-enabled user profiles on Twitter is highly limited [4]. As a result, automatically analysing and predicting user's location from Twitter data is of great significance, and has received a lot of attention from both industry and academia.

The task of predicting users' locations on Twitter is often referred to as the *Twitter User Geolocation* problem. Several algorithms have been proposed so far to solve this problem. Existing algorithms can be categorized into two broad groups, namely content-based and network-based approaches. While content-based algorithms [4, 5] exploit textual contents from tweets, network-based algorithms [6, 7] make use of the connections and interactions between users for the task of predicting user's location. Both approaches have achieved good location accuracy until now [4, 8].

This paper focuses on a more generic approach for the Twitter user geolocation problem by leveraging recent advances in deep neural networks [9] and multiview learning [10]. Deep neural networks have been proven to be very effective in many domains including image classification [11], image super-resolution [12], speech recognition [13] and compressive sensing [14]. On the other hand, multiview learning, which considers learning with multiple feature sets to improve the generalization performance, has made a great progress recently [15, 16]. Based on these techniques, our model effectively predicts users' locations from Twitter data and achieves state-of-the-art results in well-known benchmarks. Our contribution in this paper is three-fold:

- We propose a neural network architecture, named multi-entry neural network (MENET), for Twitter user geolocation. MENET is capable of combining multiview features into a unified model to infer users' locations.
- We propose a realization of MENET in a Twitter user geolocation method with four specific types of features.
- We present an extensive experimental evaluation on popular benchmarks. The experiments show that our method achieves state-of-the-art results.

The remainder of this paper is organized as follows. Section 2 briefly describes related work. Section 3 explains our method in detail, including the model architecture, feature extraction and learning. Section 4 presents our experimental settings and results. Finally, we conclude our paper in Section 5.

2. RELATED WORK

Two main approaches were proposed in the literature for the Twitter user geolocation problem. The first approach, which has been investigated thoroughly, uses textual features from tweets for building location predictive models. The second approach, on the other hand, arises from the observation that a user often interacts with people in the vicinity, and exploits the network connections of users. This section brings a closer look on recent works in both approaches.

Plenty of *content-based* methods have been proposed for Twitter user geolocation using geographical topic models [17] and Gaussian Mixture Models (GMM) [18]. More recently, Liu and Inkpen [5] trained stacked denoising autoencoders for predicting location of Twitter users. Char *et al.* [4] estimated the location by exploiting the expressiveness of sparse coding to obtain state-of-the-art results on a benchmark dataset named GeoText [17]. These methods, however, do not take into account the distribution of users' locations over the regions of interest. Addressing this problem, grid-based geolocation is introduced in [19, 20], where an adaptive or uniform grid is created to partition the datasets into appropriate cells. The prediction of geographic coordinates then is converted to a classification problem using cells as classes.

The key idea behind the *network-based* approach is that there is a correlation between the likelihood of friendship of two social

network users and their geographical distance [21]. Using this correlation, the location of a user can be revealed via his or her friends' location. By leveraging social interactions like bi-directional following [22] and bi-directional mentioning [7], one can establish a graph where label propagation [23] or its variants are used to identify location of unlabeled users. The weakness of this method is that it can not propagate labels (locations) to users who are not connected to the graph. To address this problem, methods combining textual information and graph topology knowledge are proposed in [24, 8]. Furthermore, these works build densely undirected graphs based on mentioning of users, which helps improve significantly the results. A similar mention graph is utilized in this paper. However, instead of using label propagation directly on the graph like in [24, 8], we rely on an efficient embedding to capture the graph structure.

3. MULTIVIEW LEARNING ARCHITECTURE

In Twitter user geolocation, we consider a Twitter user of interest and collect multiple tweets over a period of time. Each tweet contains not only textual content, but also metadata information such as the posting timestamp. Semantic analysis of tweets can reveal information about the location of the user. Textual data can also be used to retrieve information concerning the user's interaction with other users, i.e., to create a network of users. Together with metadata, these types of information consist the multiple views of the model proposed in the present work.

The task of Twitter geolocation is to predict the location of a user in terms of geographical region or exact geocoordinates (longitude and lattitude). Predicting the geographical region is a classification problem. Exact prediction of geocoordinates is a regression problem; however, here, we also address this problem as a classification problem. Every region is assigned a pair of geocoordinates corresponding to the median value (centroid) of the geocoordinates of all the sample users belonging to that region. After classifying a new user to a region, we use the region's centroid as an estimation of the user's location.

We propose a generic neural network model, referred to as Multi-Entry Neural Network model (MENET), to leverage multiple views of the Twitter data for this task. We realize our generic model into an effective Twitter user geolocation system, with four different types of features. Next, we present the different feature types employed in our realization, followed by the proposed MENET architecture.

3.1. Multiview Features

A common way to employ unstructured information into machine learning models is the use of embeddings to bring the information into a structured form. We build representations of textual information using word and paragraph embeddings such as Term Frequency - Inverse Document Frequency (TF-IDF) [25] and doc2vec [26]. The user network information is represented using an algorithm known as node2vec [27]. In our realization, we have also employed a timestamp feature to leverage information concerning the posting time of tweets which is often related with user's location. Next, we describe the four feature types employed in the proposed MENET architecture.

TF-IDF Feature

Term Frequency - Inverse Document Frequency (TF-IDF) [25] is a statistical measure used to evaluate how important a term is to a document in a corpus. The importance of a term increases proportionally to the number of times a term appears in the document (TF) but is offset by the frequency of the term across the corpus (IDF). TF-IDF is then defined as a product of TF and IDF values. In this work, we consider a document a concatenation of tweets posted from the same user. We employ the well-implemented library *scikit-learn* [28] to calculate TF-IDF feature from the document

Context Feature

The context feature is a mapping from a variable-length document, to a fixed-sized continuous valued vector. This vector provides a numerical representation, capturing the context of the document. Originally proposed in [26], the context feature is also referred to as doc2vec or *Distributed Representation of Sentences*, and it is an extension of the broadly used *word2vec* model [29].

In this work, we employ the Distributed Bag of Words of Paragraph Vector algorithm (PV-DBOW) [26] for extracting the doc2vec feature from the Twitter documents. The PV-DBOW algorithm delivers robust performance if trained on large datasets [26]. Our implementation utilizes the Gensim library [30] for both training and extracting features.

Node2vec Feature

Whereas the TF-IDF and doc2vec features capture textual content of the tweets, with the third feature, we aim to capture the information of a Twitter user's network. In particular, from the given tweets, we build a user network graph with each node corresponding to a user, and employ the node2vec algorithm [27] to extract continuous feature representations for each node. Given a set V of nodes, the basic idea of node2vec is to learn a function $f : V \rightarrow \mathbb{R}^d$ that maps each node into a d-dimensional feature space which preserves the connectivity patterns of the whole network.

Our user graph is formed in a way similar as in [8], [24] but instead of predicting users' locations directly on the graph, we extract node2vec feature for later use in our model. First, a unique set of nodes, V, is created for all the users of interest. If a user mentions directly another user and both of them belong to V, an edge is created reflecting this interaction. The weight of an edge is equal to the number of mentions between the two corresponding users. Moreover, if two users of interest mention a third user, who may or may not belong to V, we create an edge between these two users, with a weight equal to the sum of the mentioning times. In addition, we define celebrity users as users with a high number of unique connections with regard to a pre-defined threshold. We remove all connections to these celebrities since celebrities often have a huge number of active connections, thus mentioning a celebrity is much less likely to reveal geographical relation.

Timestamp Feature

In many Twitter databases like GeoText [17] and UTGeo2011 [19], the posting time (timestamp) of all tweets is available in terms of the Coordinated Universal Time (UTC). In [31], it was shown that there exists a relation between the time and location of a Twitter stream. In fact, it is less likely that people tweet late at night than at any other time, which implies a drift in longitudes. Therefore, the variation in timestamp could be an indication for longitude. We obtain the timestamp feature for a given user as follows. First, we extract the timestamps from all the tweets of that user and convert them to the standard format to extract the hour. After that, a 24-dimensional

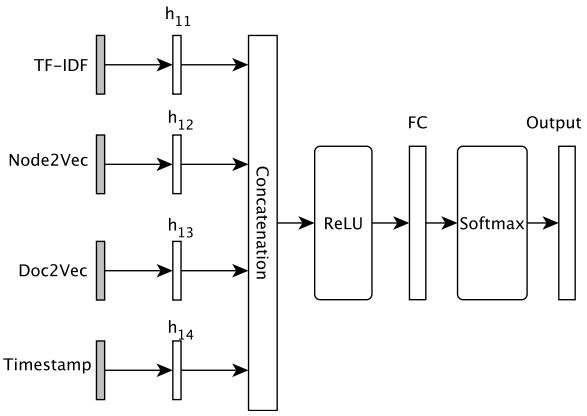

Fig. 1: The proposed multi-entry neural network architecture

vector is created corresponding to 24 hours in a day; the i-th element of this vector equals the number of messages posted by the user at the i-th hour. Finally, this feature vector is ℓ_2 normalized.

3.2. Model Architecture

The proposed model is illustrated in Fig. 1. The model takes as input different types of feature vectors. Each feature vector corresponds to one view, capturing specific information of the Twitter data. Using different views of the available Twitter data, the model classifies the respective user into one of the predefined classes corresponding to geographical regions. With four feature types employed, our realization of MENET in this work has four views, as shown in Fig. 1.

Each view is the input to one branch in MENET, which is one fully connected hidden layer. The hidden layer is followed by a Rectified Linear unit (ReLU) [32] activation function. Each branch realizes a non-linear function, mapping the original feature vectors into a unified feature space. In the learned feature space, the outputs from all views are concatenated to form a compact representation of each user. This representation serves as the input to a classifier with one fully-connected (FC) layer. This classifier employs an m-way softmax to transform scores into class probabilities, with m the number of classes.

It is worth mentioning that a more straight-forward approach to combine multiple features is to concatenate them before inputting into the network. Nevertheless, we argue that our architecture is more effective. The simple concatenation of the original feature vectors results in input vectors of high dimensions, which significantly increase the number of parameters in the model and make the model more prone to overfitting. Although the number of hidden units in each layer is adjustable, we opt to set the output of each branch, except for the timestamp, to be of lower dimension than their corresponding inputs. As a result, each branch can be seen as a dimensionality reduction function. This way, we can mitigate the overfitting problem during training.

We formulate the Twitter user geolocation task as a classification problem, and employ the cross-entropy loss as the objective function to train our model. Considering m classes of users, the cross-entropy loss over n training samples is given by

$$L = -\sum_{i=1}^{n}\sum_{j=1}^{m} y_i^j \log(\tilde{y}_i^j), \quad (1)$$

where y_i, $i = 1, \ldots, n$, is the ground-truth vector for sample i, $\tilde{y}_i$ is the vector holding the predicted probabilities of sample i for each class, and y_i^j denotes the j-th element of the respective vector.

We train our MENET model using the Stochastic Gradient Descent (SGD) algorithm. In order to control overfitting, we employ a weight decay regularizer and early stopping strategy. Particularly, during training, the model performance in a seperated validation set is monitored. If this performance decreases for a pre-defined number of epochs, the training is stopped. We also anneal the learning rate as the training proceeds.

At the testing stage, we compare the predicted location classes with the ground truth labels to measure the model's performance in terms of accuracy.

Table 1: Hyperparameter setting for MENET. $n_{h_{11}}, n_{h_{12}}, n_{h_{13}}, n_{h_{14}}$ denote the number of neurons in the hidden layers $h_{11}, h_{12}, h_{13}, h_{14}$ for the features TF-IDF, node2vec, doc2vec and timestamp, respectively.

	Number of hidden units
$n_{h_{11}}$	150
$n_{h_{12}}$	150
$n_{h_{13}}$	30
$n_{h_{14}}$	30

Table 2: Regional and state classification accuracy results on the GeoText and UTGeo2011 datasets. N/A stands for not available.

	GeoText		UTGeo2011	
	Region (%)	State (%)	Region (%)	State (%)
Eisenstein *et al.* [17]	58	27	N/A	N/A
Liu & Inkpen [5]	61.1	34.8	N/A	N/A
Cha *et al.* [4]	67	41	N/A	N/A
MENET	**76**	**64.8**	83.7	69

4. EXPERIMENTS

4.1. Datasets

In order to evaluate our proposed model, we employ two datasets, namely, the GeoText [17] and UTGeo2011 [19] datasets. The GeoText dataset contains approximately 370K tweets from 9475 users in the US, collected during the first week of March, 2010. This dataset is splitted into non-overlapping subsets of 7580, 1895 and 1895 users, for training, validation and testing, respectively. Compared to the GeoText dataset, the UTGeo2011 dataset is larger, with 38M tweets collected from 449K users in the US. In this dataset, 429K users, approximately, are reserved for training, while each one of the validation and testing sets consists of 10K users. In both datases, all tweets from a specific user are concatenated to form a single document. Following [4, 24, 8], the location prediction is performed at user level, with the ground-truth location of each user defined as the geocoordinates of the their first tweet. The location is characterized by two numbers, the longitude and latitude values.

4.2. Performance Criteria

We evaluate our model in three different tasks: (i) four-way classification of US regions including Northeast, Midwest, West and South; (ii) fifty-way classification at US state level; (iii) estimation of the real-valued user coordinates. For the first two tasks, we report the classification accuracy, whereas, for the latter task, we report

Table 3: Performance comparison on geographical coordinates prediction. The results include the mean and median distance errors and the accuracy within 161 kilometers. N/A stands for not available.

	GeoText			UTGeo2011		
	mean (km)	median (km)	@161 (%)	mean (km)	median (km)	@161 (%)
Eisenstein *et al.* [17]	900	494	N/A	N/A	N/A	N/A
Roller *et al.* [19]	897	432	35.9	860	463	34.6
Liu and Inkpen [5]	855.9	N/A	N/A	733	377	24.2
Cha *et al.* [4]	581	425	N/A	N/A	N/A	N/A
Rahimi *et al.* (2015) [24]	581	**57**	59	529	78	60
Rahimi *et al.* (2017) [8]	578	61	59	515	**77**	**61**
MENET	**570**	58	**59.1**	**474**	157	50.5

the mean and median distance errors. We also compare our model against reference methods in terms of the accuracy measure @161, which is defined as the percentage of predictions with a distance error less than 161 km[1]. All distance measures between coordinates are computed using the Haversine formula [33]. We compare the results of our models to those of recent reference methods in [17], [4], [19], [24] and [8].

4.3. Implementation Details

We implement our model using Tensorflow[2]. We first pre-process the Twitter data by performing tokenization, removing stop words, URLs and punctuation, and finally stemming the words. These preprocessing steps are implemented using the NLTK library [34].

Concerning the features, TF-IDF features are extracted using the *scikit-learn* library [28], with minimum term frequency across documents set to 40 and 500 for the GeoText and UTGeo2011 datasets, respectively. We utilize the original source codes provided by the authors in [27], [26] to extract the node2vec and doc2vec features and set both feature types to have 300 dimensions.

In our experiments, we empirically configure our model for each dataset. The model's configuration is shown in Table 1. During training, we use a small learning rate, $\alpha = 0,0001$ and regularize the weight of the output layer, with the regularization parameter λ set to 0.1. The training procedure is done using the ADAM optimization algorithm [35].

4.4. Results

The regional and state classification results are shown in Table 2. As can be seen from this table, our model significantly outperforms the state of the art in both region and state levels on the GeoText dataset. By leveraging the classification strength of multiple features, the improvement in regional accuracy is 9% compared to the state-of-the-art results presented in [4]. Concerning the state classification, the achieved accuracy is 64.8% compared to 41% in [4]. It should be noted that the classification results of the reference methods on the UTGeo2011 dataset are not available.

The comparison between all methods on the geocoordinate prediction task is presented in Table 3. Our MENET model performs overall the best on the GeoText dataset. Compared to [8], our model has lower mean and median distance errors, and marginally higher accuracy measure @161. On the UTGeo2011 dataset, our MENET model also outperforms all reference methods in terms of mean distance error. Nevertheless, [24] and [8] achieve better results for other metrics. It should be noted that these two methods employ map partitioning strategies to create new classes optimized for each dataset taking into account the geographical distribution of users. Requiring each class to have approximately the same number of users, the partitioning algorithm yields more balanced classes: high density regions (classes) are divided into smaller areas resulting in better accuracy. Large areas results in lower prediction accuracy. On the other hand, our method relies on the administrative boundaries of regions and states, ignoring the users's distribution. This brings an adverse effect on our method. However, the map partitioning strategies are independent from the network architecture and can be applied to the proposed MENET model. We leave this exploration for our future work.

[1] 161 km $\sim$ 100 mile

[2] https://www.tensorflow.org/

5. CONCLUSION

Twitter user geolocation is a challenging task because of insufficient labelled training data. The linguistically noisy nature of the Twitter data and the excessive size of the Twitter network make the task even harder. While there exist several approaches in the literature, the problem of attaining a high accuracy still remains open. In this paper, we follow the multiview learning paradigm by combining knowledge from both user-generated content and network interaction. In particular, we propose a neural network model, referred to as MENET, that uses word frequency, paragraph semantics, network topology and timestamp information, to infer users' locations. Our model achieves state-of-the-art results and can be easily extended to leverage other types of information, besides the considered types of data.

6. REFERENCES

[1] R. Compton, C. Lee, T. Lu, L. D. Silva, and M. Macy, "Detecting future social unrest in unprocessed twitter data: emerging phenomena and big data," in *IEEE International Conference on Intelligence and Security Informatics*, 2013, pp. 56–60.

[2] A. Guille and C. Favre, "Mention-anomaly-based event detection and tracking in twitter," in *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 2014, pp. 375–382.

[3] J. Bao, Y. Zheng, D. Wilkie, and M. Mokbel, "Recommendations in location-based social networks: a survey," *GeoInformatica*, vol. 19, no. 3, pp. 525–565, 2015.

[4] M. Cha, Y. Gwon, and H. T. Kung, "Twitter geolocation and regional classification via sparse coding," in *International AAAI Conference on Web and Social Media*, 2015, pp. 582–585.

[5] 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.

[6] D. Jurgens, "That's what friends are for: Inferring location in online social media platforms based on social relationships," in *The International AAAI Conference on Weblogs and Social Media*, 2013, vol. 13, pp. 273–282.

[7] R. Compton, D. Jurgens, and D. Allen, "Geotagging one hundred million twitter accounts with total variation minimization," in *IEEE International Conference on Big Data*, 2014, pp. 393–401.

[8] 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.

[9] Y. Bengio, I. J. Goodfellow, and A. Courville, "Deep learning," *Nature*, vol. 521, pp. 436–444, 2015.

[10] J. Zhao, X. Xie, X. Xu, and S. Sun, "Multi-view learning overview: Recent progress and new challenges," *Information Fusion*, vol. 38, pp. 43–54, 2017.

[11] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2012, pp. 1097–1105.

[12] W. Zhou, X. Li, and D. Reynolds, "Guided deep network for depth map super-resolution: How much can color help?," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2017, pp. 1457–1461.

[13] A. Graves, A. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 6645–6649.

[14] D. M. Nguyen, E. Tsiligianni, and N. Deligiannis, "Deep learning sparse ternary projections for compressed sensing of images," in *IEEE Global Conference on Signal and Information Processing [Available: arXiv:1708.08311]*, 2017.

[15] L. Zhang, L. Zhang, D. Tao, and X. Huang, "On combining multiple features for hyperspectral remote sensing image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 50, no. 3, pp. 879–893, 2012.

[16] J. Yu, D. Liu, D. Tao, and H. S. Seah, "On combining multiple features for cartoon character retrieval and clip synthesis," *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, vol. 42, no. 5, pp. 1413–1427, 2012.

[17] 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 Processing*, 2010, pp. 1277–1287.

[18] R. Priedhorsky, A. Culotta, and S. Y. D. Valle, "Inferring the origin locations of tweets with quantitative confidence," in *Conference on Computer supported Cooperative Work & Social Computing*, 2014, pp. 1523–1536.

[19] 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.

[20] B. Wing and J. Baldridge, "Hierarchical discriminative classification for text-based geolocation.," in *Conference on Empirical Methods in Natural Language Processing*, 2014, pp. 336–348.

[21] L. Backstrom, E. Sun, and C. Marlow, "Find me if you can: improving geographical prediction with social and spatial proximity," in *International Conference on World Wide Web*, 2010, pp. 61–70.

[22] C. A. Davis Jr, G. L. Pappa, D. R. R. Oliveira, and F. L. Arcanjo, "Inferring the location of twitter messages based on user relationships," *Transactions in GIS*, vol. 15, no. 6, pp. 735–751, 2011.

[23] U. N. Raghavan, R. Albert, and S. Kumara, "Near linear time algorithm to detect community structures in large-scale networks," *Physical review E*, vol. 76, no. 3, pp. 036106, 2007.

[24] 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.

[25] J. Leskovec, A. Rajaraman, and J. D. Ullman, *Mining of Massive Datasets*, Cambridge University Press, 2011.

[26] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *International Conference on Machine Learning*, 2014, pp. 1188–1196.

[27] 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.

[28] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[29] 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.

[30] R. Rehurek and P. Sojka, "Software framework for topic modelling with large corpora," in *LREC 2010 Workshop on New Challenges for NLP Frameworks*, 2010.

[31] M. Dredze, M. Osborne, and P. Kambadur, "Geolocation for twitter: Timing matters.," in *Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016, pp. 1064–1069.

[32] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *International Conference on Artificial Intelligence and Statistics*, 2011, pp. 315–323.

[33] R. W. Sinnott, "Virtues of the haversine," *skytel*, vol. 68, pp. 158, Dec. 1984.

[34] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python*, O'Reilly Media, Inc., 1st edition, 2009.

[35] D.P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *The International Conference on Learning Representations*, 2015.