

Fusing location and text features for sentiment classification

Wei Lun Lim, Chiung Ching Ho, Choo Yee Ting

Faculty of Computing and Informatics, Multimedia University, Cyberjaya, Malaysia

ABSTRACT

Geo-tagged Twitter data has been used recently to infer insights on the human aspects of social media. Insights related to demographics, spatial distribution of cultural activities, space-time travel trajectories for humans as well as happiness has been mined from geo-tagged twitter data in recent studies. To date, not much study has been done on the impact of the geo-location features of a Tweet on its sentiment. This observation has inspired us to propose the usage of geo-location features as a method to perform sentiment classification. In this method, the sentiment classification of geo-tagged tweets is performed by concatenating geo-location features and one-hot encoded word vectors as inputs for convolutional neural networks (CNN) and long short-term memory (LSTM) networks. The addition of language-independent features in the form of geo-location features has helped to enrich the tweet representation in order to combat the sparse nature of short tweet message. The results achieved has demonstrated that concatenating geo-location features to one-hot encoded word vectors can achieve higher accuracy as compared to the usage of word vectors alone for the purpose of sentiment classification.

Keywords: Location intelligence, sentiment analysis, social media analytics, text mining.

I. INTRODUCTION

The inferencing of insights from human physiological [1] and behavioral traits [2] have lead interesting insights in areas as diverse as security [3], human learning models [4], retail site location [5] [6] and social media [7].

The continual improvements in communication infrastructure and increased access to personal mobile devices has led to the dramatic increase in social media content. In some instances, the social media data is encoded with spatial temporal features including global positioning system (GPS) coordinates and names and types of nearby locations – which results in what is known as geo-social media[8].

One important aspect of geo-social media research is to examine the complex relationship between the users of geo-social media and their environment. Understanding this relationship can yield insights concerning the users of geo-social media, as well as the environment in which the geo-social media content is generated.

Geo-social media data have been used to infer observable aspects of geo-social media users, especially in the movement and location of users. The location of the home of geo-social media users can be inferred by using the location shared in tweets, with an accuracy rate of 80% from a sample rate of as low as 1.5 tweets per day [9]. Another study focused on the movement of users, which was derived with 95% accuracy rate using data collected over a 12 month period [10]. The success in inferring the location and movement of users via their geo-social media content has led to increased applications in the tourism sector, with studies being performed on itinerary [11] and location recommendation [12], as well as the tracking of number of visitations to tourist sites [13] and tourist-resident segmentation [14].

Non-observable aspects of a geo-social media user have also been studied, with a focus on the user's motivation for using geo-social media, as well as for localized opinion mining and reactions towards an event in the users' locality. In terms of motivation, geo-social media users often participated in a local social network in order to foster social integration and build ties, in the same way as a local newsletter [15]. Aside from community engagement, localized opinion mining can be performed using geo-social media, and have been used to elicit a response for participatory urban planning [16]. Opinion mining is often linked to responses towards an event, and collective reactions to an event has been studied [17] to infer a collective reaction across multiple geo-social media users.

The literature review performed during our study have clearly established that there is a relationship between geo-social media users and their environment. Spatial temporal data contributed by geo-social media have resulted in a "quantified environment", whereby an environment is measurable across time and space with

regards to the attributes of its users, both intrinsic and extrinsic.

The establishment of this relationship has led to the research question of whether the relationship between geo-social media users and their environment is one-way. There has been studies which reported a strong relationship between one's physical environment and one's mood and cognitive performance [18], as well as one's tendency to indulge in impulse buying [19]. As this area of research is relatively unexplored, we were thus motivated to study whether the environment impacted the geo-social media user's sentiment, in as much as the social-media user impacts the environment.

The goal of this study is to examine the impact of geo-location features from an environment on the sentiment expressed by users of geo-social media in the same environment.

In our study, we focus on geo-tagged tweets, with the objective to determine whether the concatenation of geo-location features to word vectors can result in increased accuracy of the sentiment of geo-tagged tweets. We examined a geo-tagged twitter dataset which has been tagged with its sentiment value using our approach.

II. PROPOSED APPROACH

The challenge of inferencing the sentiment of geo-tagged tweets is many-fold. Firstly geo-tagged tweets by their nature of being limited to 140 characters suffers from sparsity of data [20]. Secondly, the dependence of sentiments on phrases [21] requires that sequence information be preserved in order to achieve higher accuracy, which hampers bag-of-words approaches for tweet representation. Thirdly, the limited number of archival geo-tagged tweet datasets results in a relatively small pool of candidate data to be used as training for any word embedding model.

In this context, the proposed approach used in our study is inspired by previous solutions to the previously stated challenges. For the first challenge, concatenated word embeddings have been used to the improve performance for a word analogy task [22] in order to address the sparsity of the tweet representation. For the second challenge, the usage of a long short term memory (LSTM) architecture [23] for the purpose of maintaining sequential information have increased the accuracy of a tweet polarity task. For the final challenge, we have leveraged on the idea of using pre-trained word embeddings which was used to achieve good results on a sentiment classification task [24] which suffered from limited opportunities for training of a custom word embedding model [25].

A. Problem definition

In this study, the dataset used was modified from the

Geo-tagged Microblog Corpus [26] which contains geotagged tweets. The original dataset contained 377616 tweets, from which a subset of 10000 tweets was created via random sampling.

B. Text preprocessing and sentiment labeling

In this study, the sentiment of each tweet was determined based on purely text content. Non-text information which includes Unicode Strings, Numbers, Website links and URLs, Retweet and Mentions symbols, Hashtag symbols, Punctuation, as well as White Spaces was removed.

Following the removal of non-text information, the sentiment score for each tweet was using pattern matching of phrases implemented in the TextBlob [27] library. Each tweet was then classified to either a positive or negative class using the algorithm shown in Fig.1.

Algorithm : Classification of tweets to positive or negative sentiment

Data : Tweets

Results: categorize tweets sentiment to 0 or 1 based on sentiment score

```

while not at end of this document do read
current

    if score > 0 then

        label sentiment score as
1;

    else label sentiment score as
0;
  
```

Fig. 1. The algorithm used to label each tweet as exhibiting positive (sentiment score =1) or negative sentiment (sentiment score = 0) is shown in pseudocode

C. Geo-location features elicitation

Each tweet is processed to elicit its geo-location features. This is achieved by using the Geonames Find Nearby Web Services [28] and the Google Places Nearby Search [29] API to extract the categories of locations that is nearby the location of each tweet, as each tweet has its GPS coordinates. The nearby locations were extracted within a 300 meters radius of each tweet. The categories of nearby locations information are then vectorized before being appended to each tweet. The extraction of the categories of nearby location can be time consuming, as the number of free API calls allowed per day are limited.

D. Geo-location features elicitation

All tweets were padded to the length of 25, as this is the maximum length of the tweets after the tweets were integer encoded during vectorization. The vectorized tweets are then concatenated with the vectorized nearest location categories, resulting in a feature vector of length of 76 and 125, depending on whether the nearby locations were extracted using the Geonames and Google Places web services respectively. The length of the feature vector reflects the total categories returned by the Geonames (51 categories) and Google Places (100 categories) web services. In this study, we used two ways to generate the vectorized categories of nearby locations, namely one-hot encoding and count. The one-hot nearby location categories encoding transformed the nearby location categories into a binary vector consisting of the occurrences of said location category. Using the count method, the nearby location categories were converted into a numeric vector of occurrences, which has the value ranging from zero (non-occurrence) to N (where N is the number of occurrences). The concatenation process of the one-hot encoded text and vectorized categories of nearby location is shown in Fig. 2.

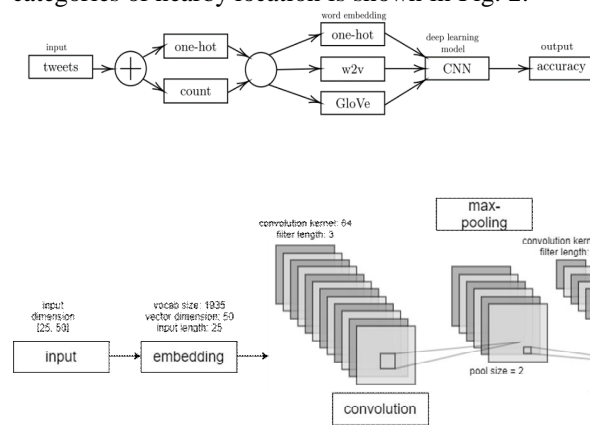


Fig.2. The concatenation of one-hot encoded text and vectorized categories of nearby location is used as an input to the embedding layer

E. Convolutional neural network architecture

In this study, a convolutional neural network (CNN) was used to perform sentiment classification using the concatenated text and geo-location features. The CNN model is composed of one embedding layer, 3 convolution layers followed by max-pooling layer, 1 flatten layer, and 2 dense layers with dropout. Each tweet is represented either as a one-hot encoded vector, or is a concatenation of the one-hot encoded text vector with vectorized geo-location features. For the embedding layer, three pre-trained word embedding models were used, namely word2vec which was trained on Google News, Glove 6B which was trained on Wikipedia, and Glove 27B which was trained on Twitter. Post embedding, three convolution layers with max pooling will generate the internal feature representation before being flattened into a one-dimensional vector. The resultant one-dimensional vector will be passed to two dense layers with dropout with the final layer utilizing a sigmoid activation function to perform the classification. Fig.3 shows the CNN architecture which was used in this study.

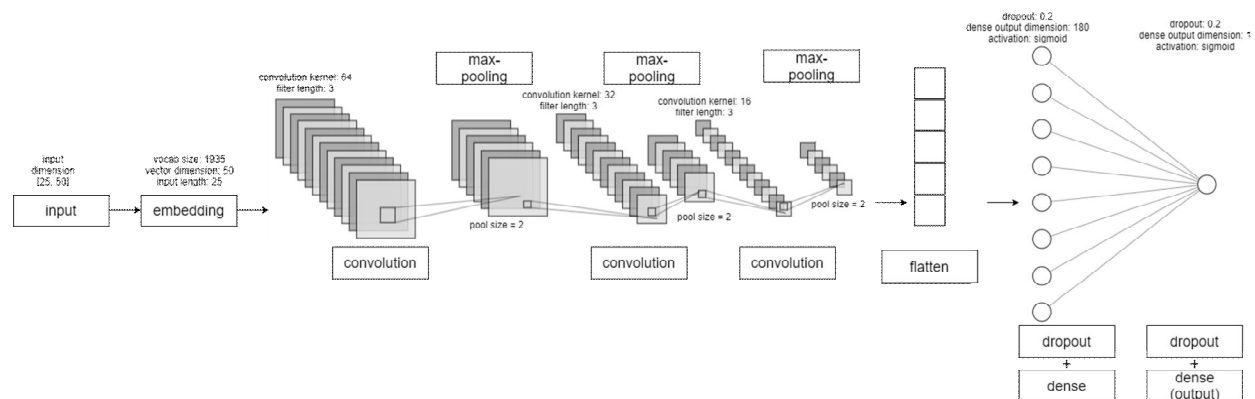


Fig. 3. The CNN architecture for performing tweet sentiment classification using a feature vector of a 50-length concatenated one-hot encoded text and vectorized categories of nearby locations

F. Long short-term memory architecture

In this study, a long short-term memory (LSTM) was used to perform sentiment classification using the concatenated text and geo-location features. The LSTM model is composed of one embedding layer, one bidirectional LSTM layer, and one dense layer with output. In this model, a bidirectional LSTMs train two, and not just one LSTMs on the concatenated feature post embedding. The first LSTM model was trained on the

input sequence as-is and the second LSTM model was trained on a reversed copy of the input sequence. For the embedding layer, three pre-trained word embedding models were used, namely word2vec which was trained on Google News, Glove 6B which was trained on Wikipedia, and Glove 27B which was trained on Twitter. Fig.4 shows the LSTM architecture which was used in this study.

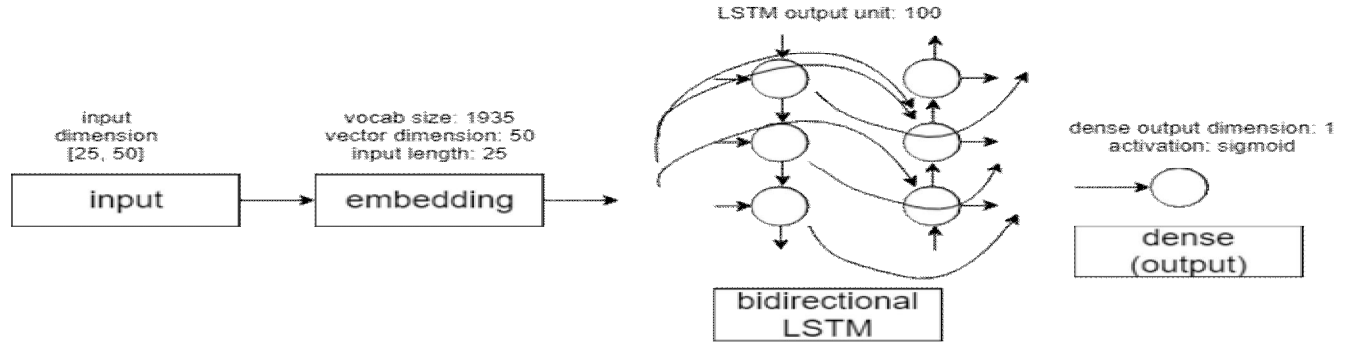


Fig. 4. The CNN architecture for tweet sentiment classification using a feature vector of a 50-length concatenated one-hot encoded text and vectorized categories of nearby locations

III. EXPERIMENTAL RESULTS

The proposed approach of concatenating one-hot encoded text to vectorized categories of nearby location was evaluated using a subset of the 10000 tweets from the Geo-tagged Microblog Corpus. From this subset, 500 geo-tagged tweets with sentiment labels were selected randomly, consisting of 250 positive and 250 negative tweets. This set of 500 labelled tweets were then split into a testing and training set using a 70:30 training to testing split.

The experimental results are grouped according to the deep learning architecture used. The loss (binary cross entropy) and activation function (Adam) was the same for both the CNN and LSTM experiments. Each experiment was repeated 10 times with 20 epochs.

A. Dataset

Five variants of dataset were used in this study : (i) tweets only, (ii) tweets concatenated with one-hot Geonames nearby location categories, (iii) tweets concatenated with one-hot Google Places nearby location categories, (iv) tweets concatenated with count of Geonames nearby location categories, and (v) tweets concatenated with count of Google Places nearby location categories.

B. Results using CNN

Tables I-IV shows the results of experiments performed, with the best performing experiments for using solely text features underlined, while the best performing experiments using the concatenated text and location features are bolded. In Table 1, the results are shown for experiments conducted with randomized weights for the embedding layer, while Tables II-IV shows the results for experiments performed using weights from the pre-trained embedding models.

Table I Accuracy of CNN model using one-hot embedding

Embedding vector dimension	Text only feature	Text & One-hot Geonames features	Text & One-hot Google Places features	Text & Count of Geonames features	Text & Count of Google Places features
200	<u>69.93</u>	70.93	66.39	68.33	71.66
300	69.26	69.33	66.66	69.33	70.93

Table II Accuracy of CNN model using pre-trained word2vec Wikipedia 100B embedding

Embedding vector dimension	Text only feature	Text & One-hot Geonames features	Text & One-hot Google Places features	Text & Count of Geonames features	Text & Count of Google Places features
300	<u>71.26</u>	73.20	73.33	73.60	72.26

Table III Accuracy of CNN model using pre-trained GloVe Twitter 6B embedding

Embedding vector dimension	Text only feature	Text & One-hot Geonames features	Text & One-hot Google Places features	Text & Count of Geonames features	Text & Count of Google Places features
200	69.86	70.00	69.99	70.93	69.46
300	<u>70.60</u>	71.79	72.00	71.26	70.60

Table IV Accuracy of CNN model using pre-trained GloVe Twitter 27B embedding

Embedding vector dimension	Text only feature	Text & One-hot Geonames features	Text & One-hot Google Places features	Text & Count of Geonames features	Text & Count of Google Places features
200	<u>76.40</u>	76.46	76.40	75.96	75.46

C. Results using LSTM networks

Tables V-VIII shows the results of experiments performed, with the best performing experiments for using solely text features underlined, while the best performing experiments using the concatenated text and location features are bolded. In Table V, the results are shown for experiments conducted with randomized weights for the embedding layer, while Tables VI-VIII shows the results for experiments

performed using weights from the pre-trained embedding models.

Table V Accuracy of LSTM model using one-hot embedding

Embedding vector dimension	Text only feature	Text & One-hot Geonames features	Text & One-hot Google Places features	Text & Count of Geonames features	Text & Count of Google Places features
200	68.99	71.93	66.73	70.99	68.40
300	<u>69.13</u>	69.86	69.19	70.93	69.80

Table VI Accuracy of LSTM model using pre-trained word2vec Wikipedia 100B embedding

Embedding vector dimension	Text only feature	Text & One-hot Geonames features	Text & One-hot Google Places features	Text & Count of Geonames features	Text & Count of Google Places features
300	<u>72.60</u>	74.20	74.53	74.53	74.86

Table VII Accuracy of LSTM model using pre-trained GloVe Twitter 6B embedding

Embedding vector dimension	Text only feature	Text & One-hot Geonames features	Text & One-hot Google Places features	Text & Count of Geonames features	Text & Count of Google Places features
200	76.73	71.26	71.53	70.86	70.26
300	<u>79.26</u>	74.33	73.86	74.26	73.93

Table VIII Accuracy of LSTM model using pre-trained GloVe Twitter 27B embedding

Embedding vector dimension	Text only feature	Text & One-hot Geonames features	Text & One-hot Google Places features	Text & Count of Geonames features	Text & Count of Google Places features
200	<u>76.93</u>	73.59	74.46	77.99	73.06

IV. DISCUSSION

The experiments performed have shown that the proposed approach of using concatenated feature vector consisting of one-hot integer encoded tweets and vectorized nearby categories of nearby locations have resulted in an increase in the accuracy of tweet sentiment classification.

The CNN experimental results showed that on average (the average of accuracy across different combinations of embedding vector dimension and concatenated vectors), the proposed concatenated feature vector consisting of one-hot integer encoded tweets and vectorized nearby categories of nearby locations is higher than that of using purely text features for all concatenated vectors used, except for the concatenation of one-hot text and one-hot Google Places nearby locations. The addition of geo-location information has enriched the initial one-hot vector representation – allowing for the convolution filters to detect additional useful local features.

The LSTM experimental results showed the opposite results, with the pure text features on average (the average of accuracy across different combinations of embedding vector dimension and concatenated vectors) outperforming the proposed concatenated feature vector. This outcome is heavily impacted by model which used the pre-trained Glove 6B model which fitted badly to the tweet data used. This outcome suggests that sequential information is of less importance to the geo-tagged tweet sentiment classification task.

In terms of API performances, it seems that there is no differences in using either the Google Places or Geonames web-services, as using both APIs resulted in the best results for 50% of the architectures used in the experiments.

For the vectorization approach for the category of the nearby locations, the count method yielded better results than the one-hot approach, with 5 out of 8 deep learning architectures used showing better results using the count approach.

We also observed that the size of the pre-trained word embedding models has a positive effect on the accuracy of the sentiment classification task, with the 100B and 27B word embedding models outperforming the 6B word embedding model.

V. CONCLUSION

In this study, we have proposed a new approach towards classification of geo-tagged tweets, using a concatenated feature vector consisting of one-hot integer encoded tweets and vectorized nearby categories of nearby locations. A significant challenge is in the interpretation of the effects of adding geo-location information to the text representation.

The addition of location features has resulted in an increase in accuracy rates in the experiments performed. For the experiments involving CNN, a total of 5.53% improvement was achieved when the concatenated word and location features were used as opposed to solely using one-hot encoded text vectors. RNN showed a similar net improvement when all experiments were considered with an improvement of 0.12% percent.

For future works, we will attempt to use a larger dataset with mixed language content, as well as develop a mechanism which will help to better explain and interpret the models developed using our approach.

ACKNOWLEDGEMENT

This study is funded by Telekom Malaysia Research & Development.

REFERENCES

- [1] C. C. Ho et al., "MMU GASPFA: A COTS multimodal biometric database," *Pattern Recognition Letters*, vol. 34, no. 15, pp. 2043–2050, Nov. 2013.
- [2] C. C. Ho, C. Eswaran, K.-W. Ng, and J.-Y. Leow, "An Unobtrusive Android Person Verification Using Accelerometer Based Gait," in *Proceedings of the 10th International Conference on Advances in Mobile Computing & Multimedia (MoMM2012)*, Bali, 2012, pp. 271–274.
- [3] C. C. Ho and C. Eswaran, "Consolidation of fingerprint databases: A Malaysian case study," in *In the Proceedings of the 2011 11th International Conference on Hybrid Intelligent Systems (HIS)*, 2011, pp. 455–462.
- [4] C.-Y. Ting, W.-N. Cheah, and C. C. Ho, "Student Engagement Modeling using Bayesian Networks," in *In Proc. IEEE International Conference on Systems, Man, and Cybernetics*, Manchester, UK, 2013.
- [5] C.-Y. Ting, C. C. Ho, H. J. Yee, and W. R. Matsah, "Geospatial Analytics in Retail Site Selection and Sales Prediction," *Big Data*, vol. 6, no. 1, pp. 42–52, Mar. 2018.
- [6] J. Y. L. Yap, C. C. Ho, and C.-Y. Ting, "A systematic review of the applications of multi-criteria decision-making methods in site selection problems," *Built Environment Project and Asset Management*, Sep. 2019.
- [7] K.-C. Ong, C. C. Ho, and K.-H. Poo, "A comparative study of HITS vs PageRank algorithms for Twitter users analysis," in *2014 International Conference on Computational Science and Technology (ICCST)*, 2014, pp. 1–6.
- [8] F. Chen, A. Boedihardjo, and C.-T. Lu, "Preface: special issue on geo-social media analytics," *Geoinformatica*, vol. 22, no. 3, pp. 503–506, Jul. 2018.
- [9] H. Kavak, D. Vernon-Bido, and J. J. Padilla, "Fine-Scale Prediction of People's Home Location Using Social Media Footprints," in *Social, Cultural, and Behavioral Modeling*, 2018, pp. 183–189.
- [10] Y. Yuan and X. Wang, "Exploring the effectiveness of location-based social media in modeling user activity space: A case study of Weibo," *Transactions in GIS*, vol. 22, no. 4, pp. 930–957, 2018.
- [11] H. Q. Vu, G. Li, R. Law, and Y. Zhang, "Tourist Activity Analysis by Leveraging Mobile Social Media Data," *Journal of Travel Research*, vol. 57, no. 7, pp. 883–898, Sep. 2018.
- [12] C. Bothorel, N. Lathia, R. Picot-Clemente, and A. Noulas, "Location Recommendation with Social Media Data," in *Social Information Access: Systems and Technologies*, P. Brusilovsky and D. He, Eds. Cham: Springer International Publishing, 2018, pp. 624–653.
- [13] Z. A. Hamstead, D. Fisher, R. T. Ilieva, S. A. Wood, T. McPhearson, and P. Kremer, "Geolocated social media as a rapid indicator of park visitation and equitable park access," *Computers, Environment and Urban Systems*, vol. 72, pp. 38–50, Nov. 2018.
- [14] M. M. Hasnat and S. Hasan, "Identifying tourists and analyzing spatial patterns of their destinations from location-based social media data," *Transportation Research Part C: Emerging Technologies*, vol. 96, pp. 38–54, Nov. 2018.
- [15] J. Swart, C. Peters, and M. Broersma, "Sharing and Discussing News in Private Social Media Groups," *Digital Journalism*, vol. 7, no. 2, pp. 187–205, Feb. 2019.
- [16] P. Nummi, "Social Media Data Analysis in Urban E-Planning," *Smart Cities and Smart Spaces: Concepts, Methodologies, Tools, and Applications*, pp. 636–651, 2019.
- [17] A. Dunkel, G. Andrienko, N. Andrienko, D. Burghardt, E. Hauthal, and R. Purves, "A conceptual framework for studying collective reactions to events in location-based social media," *International Journal of Geographical Information Science*, vol. 33, no. 4, pp. 780–804, Apr. 2019.
- [18] S. K. Das et al., "High altitude with monotonous environment has significant impact on mood and cognitive performance of acclimatized lowlanders: Possible role of altered serum BDNF and plasma homocysteine level," *Journal of Affective Disorders*, vol. 237, pp. 94–103, Sep. 2018.
- [19] B. Sivakumaran, G. Mohan, and P. Sharma, "Impact of store environment on impulse buying behavior," *European Journal of Marketing*, vol. 47, no. 10, pp. 1711–1732, Sep. 2013.
- [20] H. Saif, Y. He, and H. Alani, "Alleviating data sparsity for Twitter sentiment analysis," presented at the 2nd Workshop on Making Sense of Microposts (#MSM2012): Big things come in small packages at the 21st International Conference on the World Wide Web (WWW'12), Lyon, France, 2012, pp. 2–9.
- [21] R. Socher et al., "Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank," presented at the Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, 2013, pp. 1631–1642.
- [22] J. Gaten, K. Sagae, V. Ustun, and M. Dehghani, "Combining Distributed Vector Representations for Words," in *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, Denver, Colorado, 2015, pp. 95–101.
- [23] X. Wang, Y. Liu, C. Sun, B. Wang, and X. Wang, "Predicting Polarities of Tweets by Composing Word Embeddings with Long Short-Term Memory," 2015, pp. 1343–1353.
- [24] Y. Kim, "Convolutional Neural Networks for Sentence Classification," *EMNLP 2014*, 2014.
- [25] Y. Qi, D. Sachan, M. Felix, S. Padmanabhan, and G. Neubig, "When and Why Are Pre-Trained Word Embeddings Useful for Neural Machine Translation?," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, New Orleans, Louisiana, 2018, pp. 529–535.
- [26] 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*, Stroudsburg, PA, USA, 2010, pp. 1277–1287.
- [27] S. Loria, "Advanced Usage: Overriding Models and the Blobber Class — TextBlob 0.15.2 documentation," 2018. [Online]. Available: https://textblob.readthedocs.io/en/dev/advanced_usage.html#sentiment-analyzers. [Accessed: 01-May-2019].
- [28] Geonames.org, "GeoNames Web Service Documentation," 29-Apr-2019. [Online]. Available: <http://www.geonames.org/export/web-services.html>. [Accessed: 01-May-2019].
- [29] Google.com, "Place Search / Places API," Google Developers, 2019. [Online]. Available: <https://developers.google.com/places/web-service/search>. [Accessed: 01-May-2019]