DEEP LEARNING MODEL TO RETRIEVES THE LOCATION ENTITY FROM USER TEXT POSTS

Mehveen Mehdi Khatoon[1]*, Tasneem Rahath[2], Ruhiat Sultana[3]

Abstract

Background/objectives: The position must be predicted before any analysis can be done. One of the analyses used to process geographic information is location analysis (GI).

Methodology: The Global Positioning System (GPS) on a user's device, for instance, can be used to add latitude and longitude to posts or tweets on social media. Regrettably, not every user or content producer includes latitude and longitude in their tweets. Using the user's text posts is one method for determining where they are in the world. This method takes the user's text post and retrieves the location entity from it. The predicted location candidate is the extracted location entity. With the use of user data, we hope to include another strategy in this study.

Findings: The user who tweeted the tweet then adds more details to the classification of the retrieved location entities. Later, this classification will determine if a user's post is relevant or not. For the model being utilised, deep learning would be combined.

Novelty/improvements: The pre-trained BERT model is used with the extracted location entities. Using BiLSTM-CNN, the final classifications will determine how to forecast a position. Since the research is still in its early stages, this paper cannot yet reveal the precise findings and evaluation.

Keywords: Location Prediction, Social Media, BERT, CNN, BiLSTM

[1*]Associate Prof, Bhoj Reddy Engineering College for Women, Dept of IT, JNTUH, INDIA, Email:mkmehveen@gmail.com
[2]Associate Prof, Bhoj Reddy Engineering College for Women, Dept of IT, JNTUH, INDIA, Email: tas.rahath@gmail.com
[3]Associate Prof, Bhoj Reddy Engineering College for Women, Dept of IT, JNTUH, INDIA, Email: ruhiatsultana@gmail.com

***Corresponding Author:** Mehveen Mehdi Khatoon
*Associate Prof, Bhoj Reddy Engineering College for Women, Dept of IT, JNTUH, INDIA, Email:mkmehveen@gmail.com

DOI: 10.48047/ecb/2023.12.si5a.0607

1. Introduction

Location analysis keeps growing as many applications using and producing location information. The analysis is not only used in the public sector but also used in the private sector. Location analysis is important for the public sector to solve the public's problems, for example analysis early warning system for disaster [[1],[2],[3]], surveillance [4], event detection [[5],[6]], and traffic analysis [[7],[8]]. In the private sector, they need to know where to place facilities that would be successful to sustain their business, such as (1) targeted advertising, (2) market segmentation [9], and (3) sentiment analysis [[10],[11],[12],[13]].

Social media data provide both clear and implied location information. Some platforms included check-in features to display the specific user's location at that moment. The linked text, image, or video contains the location that is associated with the coordination of latitude and longitude. The latitude and longitude are not, however, provided by all users or content producers. The motives could range from discomfort to concerns about privacy [14]. Since social media data may be used to predict location, numerous studies have done so [[15], [16], [17], [18], [19], [20], [21], [22], [23], [24]].

There is any approach to exploit the location in social media data. (1) User text post[[25],[26]], (2) User Information[[27],[28]]. User text post is the text content that has been sent to social media. It's including posting/tweet text, time of posting, and related tag of user. The text content is going to analyze to extract the location entity. One of the methods to extract is Name Entity Recognition (NER). User information is the corelated information of the user in social media. It's including user profile and user correlation, such as user following and user followers. Trajectory is the positioning history of the users. It related with user check-in position. Usually, this approach to get the next location user based on the history pattern.

In this study, we combine two approaches to predict location using user text posts and user information. User text post is used to get location extraction. If any location entity is found in the user text post, it will be a candidate for classification whether the location is related to the user's posting location or not. There are any researchers which already used the combination method to solve location prediction[[29],[30]]. Some of them reported improved performance in their research.

For the model used will perform some combination of deep learning. For the approach using user text post will use location extraction technique. Deep learning is used using the Bidirectional Encoder from Transformers (BERT) model. Meanwhile, to get the relationship between the user's post location or not, it will combine the two neural networks Convolutional Neural Network (CNN) and Bidirectional Long Short-Term Memory (BiLSTM). The combination of these two techniques is used to take advantage of the architectural advantages of these two deep learning techniques. CNN is used to extract high-level features in text and absorb complex nonlinear mapping relationships from text [3]. Meanwhile, BiLSTM helps extract word semantics in the context of before and after information. The advantage of BiLSTM will be to use forward and backward learning. The proposed combinations allow extracting deeper and maximum information from the user's post data whether it is related to the user's location or not. Putra et al [31] in his research carried out location extraction using the Named Entity Recognition (NER) method by adopting the Neuro NER method with BiLSTM and Conditional Random Field (CRF).

The format of this paper is as follows. Following the introduction in section 1, section 2 explains the research methodology, section 3 is about results and discussion, and section 4 discusses its conclusions. In this paper we are still using one model i.e., BERT to solve NER problem.

2. Methodology

This study went through numerous stages. Data collection and analysis come first, then data administration, layer configuration, and training stages. This paper will implement the information extraction process' training step. The processes of developing the classification model and testing currently cannot be completed in this study. This is depicted in Figure. 1. The majority of the reporting in this article is still in the planning stage. One such is the yet-to-be-implemented combo of bert and bilstm-CNN.

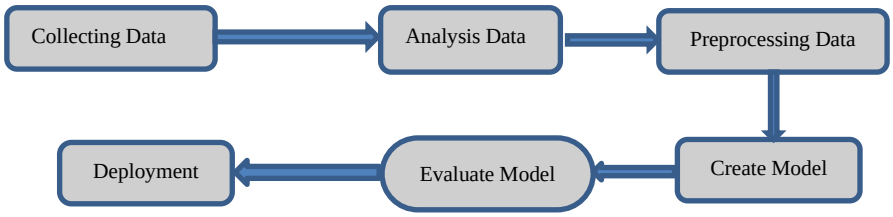

Fig 1 Method Research

2.1 Collecting and analysing data

Twitter is used to crawl for Twitter data. Users can crawl based on keywords, user ids, times/dates, and/or locations by utilising the Twitter API. Data collection was scheduled to take place for one month. CSV storage is used to store crawler results. Figure 2 displays the data crawling flow on Twitter.

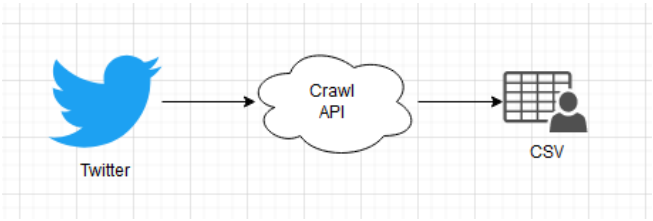

Fig 2 Process Crawling

2.2 Pre-processing Data

Data preparation is done across the system at various levels. The following list of pre-processing methods is what the system uses:

- ❖ Select a Twitter post: The location is already present in the Twitter post that was chosen, which makes it easier to label the dataset for location predictions (see Figure. 3).
- ❖ Case folding Lowercase all fonts
- ❖ Removed the phrase "RT" from the start of the tweet text;
- ❖ Removed URLs and mentions;
- ❖ Removed substrings from links and mentions in tweet data;
- ❖ Removed all punctuation, save for the question mark “?" point "." comma ",", and dash "-"

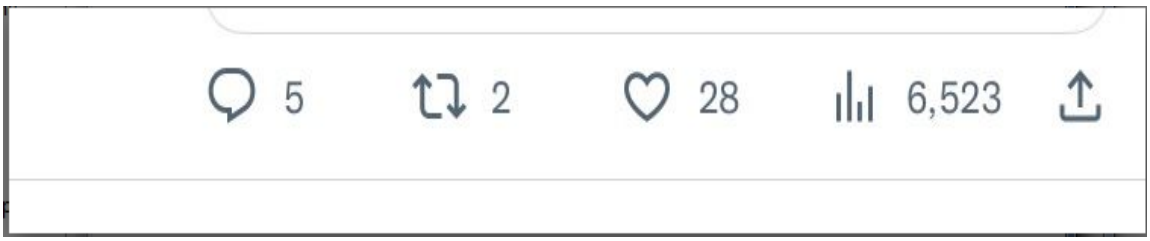

Fig 3 Select Twitter Post included location

2.3 Name Entity Recognition (NER)

NER is used in this study's information extraction technique to extract entities. Six entity notation labels, such as: divide the different sorts of entities employed in this study. PERSON (B-Per, I-Per), PLACE (B-Loc, I-Loc), and ORGANIZATION (B-Per, I-Per) (B-Org,I-Org). A sufficient amount of training data and accurate labelling are necessary for the algorithm to extract information correctly. The training set that was utilised was the same set that was used to manually label the various tweet kinds. The IOB or BIO format is used for data labelling. Labeled B (Begin) if the word is the initial word of the entity, I (Inside) if it isn't, and O (Outside / Other) if it doesn't fall into any of the entity categories. Table 1 will show the example of labelling results for the training and test data sets.

Table 1 Example Dataset

Tweet Text	Token	Label
Carol went to Bengaluru	Carol	B-Person
	Went	O
	To	O
	Bangaluru	B-Loc

3.Discussion and Results

By evaluating the ratio of the number of entities that are correctly identified to the total number of entities recognised, precision is employed in this process to gauge how well the named entity recognition system can identify the type of entity. Equation 1 serves as the precision calculation formula. TP stands for true positive, FP for false positive, where TP is the total number of correctly detected entities. FP, on the other hand, is the total number of entities in a given entity that are properly and wrongly recognised.

$$Precision = \frac{TP}{TP+FP}\ x\ 100\% \qquad (1)$$

Measure the system's capacity to identify entities that are pertinent to the type of entity. The recall formula is applied as in Equation 2, where FN = false negative, or the quantity of entity identification mistakes that are identified in the kind of entity, is used.

$$Recall = \frac{TP}{TP+FN}\ x\ 100 \qquad (2)$$

F-Measure (F) is the harmonic mean of precision and recall. The f-measure formula is as in Equation 3.

$$F = 2\ x\ \frac{Precision\ x\ Recall}{Precision +\ Recall}\ x\ 100 \qquad (3)$$

The location entity extraction process uses the bert pre-training use case. In this study, training was conducted using a dataset from NLU. The results show the fine tuning obtained as follows in table 2.

Table 2 Matriks Evaluation

Precision	Recall	F1	Accuracy
73%,	89%	70%	96%

Table 3 shows the results of the NER process result. The model consumes the text input. Next, the model output text output result with each word to classified to each entity.

Table 3 NER Result

Text Input	Text Output
John comes from Chennai	B-Person I-Person comes from B-LOC

4.Conclusion

Despite the limitation of geographical information in social media data. Location Prediction (LP) is another solution to get location information. Combined multiple approach is one of the solutions to improve the accuracy of location prediction. In this research, we purpose using more than one approach and algorithm. The approach is used are using the user post text and user information. The algorithm using BERT, BiLSTM and CNN, we hope the system capable to predict user post location. The process implementing location entity recognition before processing classification as the main requirement. Based on first process to extract location entity, the calculations precision, recall, and f-1 measure obtained 71%, 80%, 75%. The accuracy is 95% in epoch eight. We plan to continue the research to get the location classify. Evaluate the whole model and comparing with another algorithm is another plan to do in the research. Hopefully, the purposing method is successful and bring some contributions.

5. References

1. A. Kumar and J. P. Singh, 'Location reference identification from tweets during emergencies : A deep learning approach', *Int. J. Disaster Risk Reduct.*, vol. 33, pp. 365–375, Feb. 2019, doi: 10.1016/j.ijdrr.2018. 10. 021.
2. J. P. Singh, Y. K. Dwivedi, N. P. Rana, A. Kumar, and K. K. Kapoor, 'Event classification and location prediction from tweets during disasters', *Ann. Oper. Res.*, vol. 283, no. 1–2, pp. 737–757, Dec. 2019, doi: 10.1007/s10479-017-2522-3.
3. C. Fan, F. Wu, and A. Mostafavi, 'A Hybrid Machine Learning Pipeline for Automated Mapping of Events and Locations from social media in Disasters', *IEEE Access*, vol. 8, pp. 10478–10490, 2020, doi: 10.1109/ACCESS. 2020.2965550.
4. T. Wen Long, A. Zainal, and M. Nizam Kassim, 'News Event Prediction using Causality Approach on South China Sea Conflict', in *2021 3rd International Cyber Resilience Conference (CRC)*, Langkawi Island, Malaysia, Jan. 2021, pp.1–6. doi:10.1109/CRC50527.2021.9392431.
5. X. Wan, M. C. Lucic, H. Ghazzai, and Y. Massoud, 'Empowering Real-Time Traffic Reporting Systems With NLP-Processed Social Media Data', *IEEE Open J. Intell. Transp. Syst.*, vol. 1, pp. 159–175, 2020, doi: 10.1109/OJITS. 2020.3024245.
6. Y. Wang, Z. He, and J. Hu, 'Traffic Information Mining From Social Media Based on the MC-LSTM-Conv Model', *IEEE Trans. Intell. Transp. Syst.*, pp. 1–13, 2020, doi: 10.1109/TITS.2020.3021096.

7. L. E. Ferro-Díez, N. M. Villegas, J. Díaz-Cely, and S. G. Acosta, 'Geo-Spatial Market Segmentation amp; Characterization Exploiting User Generated Text Through Transformers amp; Density-Based Clustering', *IEEE Access*, vol. 9, pp. 55698–55713, 2021, doi: 10.1109/ACCESS.2021.3071620.
8. W. L. Lim, C. C. Ho, and C.-Y. Ting, 'Sentiment Analysis by Fusing Text and Location Features of Geo-Tagged Tweets', *IEEE Access*, vol. 8, pp. 181014–181027, 2020, doi: 10.1109/ACCESS.2020.3027845.
9. W. Yue and L. Li, 'Sentiment Analysis using Word2vec-CNN-BiLSTM Classification', in *2020 Seventh International Conference on Social Networks Analysis, Management and Security (SNAMS)*, Paris, France, Dec. 2020, pp. 1–5. doi: 10.1109/SNAMS52053.2020.9336549.
10. M. Xuanyuan, L. Xiao, and M. Duan, 'Sentiment Classification Algorithm Based on Multi-Modal Social Media Text Information', *IEEE Access*, vol. 9, pp. 33410–33418, 2021, doi: 10.1109/ACCESS.2021.3061450.
11. A. Mostafa, W. Gad, T. Abdelkader, and N. Badr, 'Predicting the Tweet Location Based on KNN-Sentimental Analysis', in *2020 15th International Conference on Computer Engineering and Systems (ICCES)*, Dec. 2020, pp. 1–6. doi: 10.1109/ICCES51560.2020.9334566.
12. C.-Y. Huang, H. Tong, J. He, and R. Maciejewski, 'Location Prediction for Tweets', *Front. Big Data*, vol. 2, p. 5, May 2019, doi: 10.3389/fdata.2019.00005.
13. K. Alharthi, K. E. Hindi, and S. M. Alzahrani, 'Venue-Popularity Prediction Using Social Data Participatory Sensing Systems and RNNs', *IEEE Access*, vol. 9, pp. 3140–3154, 2021, doi: 10.1109/ACCESS.2020.3047680.
14. Q. Gao, F. Zhou, G. Trajcevski, K. Zhang, T. Zhong, and F. Zhang, 'Predicting Human Mobility via Variational Attention', in *The World Wide Web Conference on - WWW '19*, San Francisco, CA, USA, 2019, pp. 2750–2756. doi: 10.1145/3308558.3313610.
15. J. Chen, J. Li, and Y. Li, 'Predicting Human Mobility via Long Short-Term Patterns', *Comput. Model. Eng. Sci.*, vol. 124, no. 3, pp. 847–864, 2020, doi: 10.32604/cmes.2020. 010240.
16. 'Next Location Prediction with a Graph Convolutional Network Based on a Seq2seq Framework', *KSII Trans. Internet Inf. Syst.*, vol. 14, no. 5, May 2020, doi: 10.3837/tiis.2020.05.003.
17. X. Wang, Y. Liu, X. Zhou, Z. Leng, and X. Wang, 'Long- and Short-Term Preference Modeling Based on Multi-Level Attention for Next POI Recommendation', *ISPRS Int. J. Geo-Inf.*, vol. 11, no. 6, p. 323, May 2022, doi: 10.3390/ijgi11060323.
18. A. Azhar*et al.*, 'Detection and prediction of traffic accidents using deep learning techniques', *Clust. Comput.*, Jan. 2022, doi: 10.1007/s10586-021-03502-1.
19. L. Huang, Y. Ma, Y. Liu, and K. He, 'DAN-SNR: A Deep Attentive Network for Social-aware Next Point-of-interest Recommendation', *ACM Trans. Internet Technol.*, vol. 21, no. 1, pp. 1–27, Feb. 2021, doi: 10.1145/3430504.
20. Y. Chen, C. Long, G. Cong, and C. Li, 'Context-aware Deep Model for Joint Mobility and Time Prediction', in *Proceedings of the 13th International Conference on Web Search and Data Mining*, Houston TX USA, Jan. 2020, pp. 106–114. doi: 10.1145/3336191.3371837
21. J. Diaz, B. Poblete, and F. Bravo-Marquez, 'An integrated model for textual social media data with spatio-temporal dimensions', *Inf. Process. Manag.*, vol. 57, no. 5, p. 102219, Sep. 2020, doi: 10.1016/j.ipm.2020.102219.
22. L. Huang, Y. Ma, S. Wang, and Y. Liu, 'An Attention-Based Spatiotemporal LSTM Network for Next POI Recommendation', *IEEE Trans. Serv. Comput.*, vol. 14, no. 6, pp. 1585–1597, Nov. 2021, doi: 10.1109/TSC.2019 .2918310.
23. J. H. Reelfs, M. Bergmann, O. Hohlfeld, and N. Henckell, 'Understanding & Predicting User Lifetime with Machine Learning in an Anonymous Location-Based Social Network', in *Companion Proceedings of the Web Conference 2021*, Apr. 2021, pp. 324–331. doi: 10.1145/3442442.3451887.
24. L. Alsudias and P. Rayson, 'Social Media Monitoring of the COVID-19 Pandemic and Influenza Epidemic with Adaptation for Informal Language in Arabic Twitter Data: Qualitative Study', *JMIR Med. Inform.*, vol. 9, no. 9, p. e27670, Sep. 2021, doi: 10.2196 /27670.
25. L. F. Simanjuntak, R. Mahendra, and E. Yulianti, 'We Know You Are Living in Bali: Location Prediction of Twitter Users Using BERT Language Model', *Big Data Cogn. Comput.*, vol. 6, no. 3, p. 77, Jul. 2022, doi: 10.3390/bdcc6030077.
26. H. C. M. Senefonte, M. R. Delgado, R. Lüders, and T. H. Silva, 'PredicTour: Predicting Mobility Patterns of Tourists Based on Social Media User's Profiles', *IEEE Access*, vol. 10, pp. 9257–9270, 2022,

doi: 10.1109/ACCESS.2022.3143503.
27. H. C. M. Senefonte, M. R. Delgado, R. Lüders, and T. H. Silva, 'PredicTour: Predicting Mobility Patterns of Tourists Based on Social Media User's Profiles', *IEEE Access*, vol. 10, pp. 9257–9270, 2022, doi: 10.1109/ACCESS.2022.3143503.
28. S. Safavi and M. Jalali, 'RecPOID: POI Recommendation with Friendship Aware and Deep CNN', *Future Internet*, vol. 13, no. 3, p. 79, Mar. 2021, doi: 10.3390/fi13030079.
29. D. Contractor, B. Patra, Mausam, and P. Singla, 'Constrained BERT BiLSTM CRF for understanding multi-sentence entity- seeking questions', *Nat. Lang. Eng.*, vol. 27, no. 1, pp. 65–87, Jan. 2021, doi: 10.1017/S1351324920000017.
30. Y. Bao, Z. Huang, L. Li, Y. Wang, and Y. Liu, 'A BiLSTM-CNN model for predicting users' next locations based on geotagged social media', *Int. J. Geogr. Inf. Sci.*, vol. 35, no. 4, pp. 639–660, Apr. 2021, doi: 10.1080/13658816.2020.1808896.
31. F. N. Putra and C. Fatichah, 'Klasifikasijeniske jadianmenggunakankombinasi Neuro NER dan Recurrent Convolutional Neural Network pada data Twitter', *Regist. J. Ilm. Teknol. Sist. Inf.*, vol. 4, no. 2, p. 81, Jul. 2018, doi: 10.26594/register. v4i2.1242.