

Extracting geo-references from social media text using bi-long short term memory networks

Dharmendra Mangal¹, Hemant Makwana²

¹Department of Computer Science and Engineering, Medi-Caps University, Indore, India

²Institute of Engineering and Technology, Institute of Engineering and Technology, Indore, India

Article Info

Article history:

Received Dec 21, 2023

Revised Mar 18, 2024

Accepted Apr 13, 2024

Keywords:

Bi-LSTM

Event detection

Geo-references

Location extraction

Named entity recognition

ABSTRACT

The social media data provides great source of information about global and local events, with millions of users. More precisely, the fact that brief messages are practical and are highly popular. Many recent studies have been motivated to estimate the location of the events identified by tracking posts in social media text messages. It might be difficult to extract location data and estimate the location of an event while maintaining a sufficient level of situation awareness, particularly in disaster situations like fires or traffic accidents. In this presented work we proposed an approach to identify geo-references in the text messages. We used bi-directional long short term memory (LSTM) neural networks to extract location information in the text messages. The results show that applying Bi-LSTM on dataset gives high level accuracy after fine-tuning (up to 10 epochs). The testing results show that accuracy achieved is 0.98 and 0.076 loss value. This proves that the proposed methodology is better than the previous conditional random field (CRF)-based approaches.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Dharmendra Mangal

Department of Computer Science and Engineering, Medi-Caps University

Indore, India

Email: managldharmendra83@gmail.com

1. INTRODUCTION

For rich and current source of information social media is best choice with their steadily growing user base. Particularly popular are social media platforms, which are heavily used by a huge number of users due to their well-liked features including multimedia messages and quick access on mobile devices [1]. Twitter (now-X) seems to be the most well-liked service among these platforms, with a large user base. Given that most users permit the public to view their posts and user profile data, which offers extensive data for text-based research sociological analysis, information extraction, text mining, or analysis [2].

Extraction of events and estimation of locations from text message data in social networks are essential for keeping up with current happening of various events [3]. These methods are useful for knowing what is happening on nearby and for getting emergency information quickly in a variety of contexts. One of the potential use cases may be to determine the location of an incident, open alternate roads, and administer first aid based on Twitter tweets from people who witnessed a serious collision involving multiple cars [2]. Likewise, in terms of handling natural calamities like earthquakes or floods, having access to quick access to such information is beneficial [4].

Finding evidence of a location, particularly by identifying geographical names (also called toponyms), in social network messages, is a crucial challenge for location extraction [5]. The extraction of location from text is also known as Geo-parsing. The Geo-parsing challenge is a particular sub-problem of

the well-known entity recognition (NER) job, a natural language processing (NLP) problem that aims to extract certain entities from text, such as persons, organizations, dates, and locations [6]. The majority of the lengthy, formal texts that are used by NER approaches as suggested in the literature are newspaper stories and social media text [5]. Once identified in social media postings, toponyms offer rich and practical input for location estimation tasks. But there are a few difficulties with this phase as well. A significant issue is that, besides well-known toponyms, there might be additional indicators of the event's location, including the geo-tagged messages. This additional information also can be utilized [7].

We can classify the previous work done in the area of toponym recognition into three categories i.e., gazetteer-based, rule based and machine learning based. The first method uses a gazetteer or predetermined list of location names [8]. Primarily the recognition procedure depends on searching if a specific token is listed in the gazetteer [9]. The location recognition applications context determines the lists content and level of granularity. The list may include the names of the nation, city or town for general-purpose solutions. Depending on the application's context, this list may include more specific information for a given geographic area, such as banks, pharmacies, hospitals [7]. The OpenStreetMap and Wikipedia are widely used resources for general-purpose gazetteers [10], [11]. The rule-based method defines specific patterns in the form of rules. For example, when a token comes after the word "city" it's regarded as a toponym that refers to a city name (like "Jaipur"). Part-of-speech (POS) tagging and morphological analysis also helps to find the patterns (sequencing of words) in a sentence. For example, when a pronoun comes before a token in English, it usually indicates that the token is a toponym (e.g., "New Delhi") [12].

The gazetteer based and rules-based approaches are work well on very small set of location information. But social media data scope is global one [8]. To handle such global data machine learning approaches works better. Further the more advanced approaches are based on deep learning concepts that improve the accuracy of the decision. One of the popular approaches used in the machine learning-based method for toponym recognition is supervised learning. The conditional random field (CRF) is the most widely accepted supervised learning method used for named entity recognition. One major benefit of CRF is that it allows contextual information to be included into the model, such as the presence of words close to a toponym. In this work we have used bi-directional long short term memory (Bi-LSTM) model. The results show that deep learning method we applied improves the accuracy compared to CRF-based approach which is a machine learning based approach [13].

2. METHOD

The detection of the events from social media is insufficient for event analysis as it requires correct location of the event also [14]. This motivates to propose novel methodology for location extraction. The Figure 1 shows the proposed methodology.

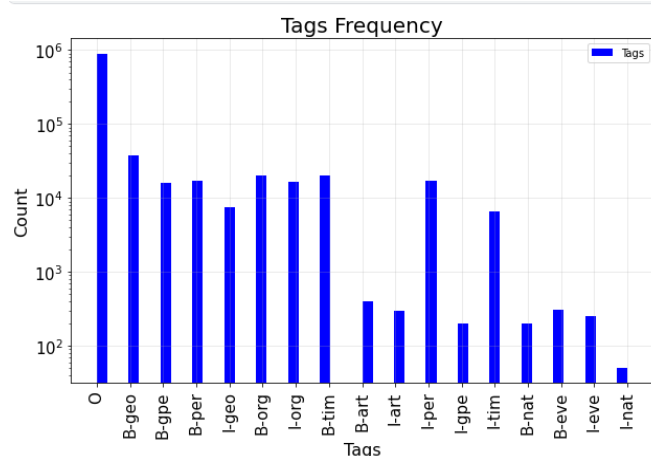


Figure 1. Different types of labels

2.1. Dataset

An annotated NER corpus is used which is available on Kaggle with 47,959 unique text rows. The numbers of unique words are 35,178 along with 17 unique NER tags. There are three chunks of words:

beginning chunk (B) for the words in the beginning of the sentence, inside words belong to I-chunk remaining words we consider for O-chunk i.e., outside the scope. These I-O-B chunks further classified into different tags like for geographical entity (Geo-tag), organization (org-tag), person (per-tag), geopolitical entity (gpe-tag), time indicator (tim-tag), event (eve-tag), natural phenomenon (nat-tag). The following table shows the frequency of words belongs to major chunks in the dataset used. The Table 1 can be visually analyzed using the following graph shown in Figure 1. The graph shows the relevancy of the dataset for this work as it has ample amount of geographical related information to let model learn and test. The following figure shows the methodology used.

Table 1. Distribution of tags in datasets

Tag	O	B-geo	B-gpe	B-per	I-geo	B-org	I-org	B-tim	I-per	I-gpe	I-tim	B-nat	B-eve	I-eve	I-nat
Count	887908	37644	15870	16990	7414	20143	16784	20333	17251	198	6528	201	308	253	51

2.2. Data preprocessing

The preprocessing of the data is mandatory before it is inputted to the learning model. It not only improves the overall performance of the system along with making learning faster. The first step in data preprocessing is tokenizing the sentences into words. Each of unique word is identified and assigned with an index number. By using word embedding the word sequence is converted into the number sequence of the unique word. After this sentence padding is done to make each sentence of equal length. After that word tagging is done with the 17 tags. The tags are indexed with values from 0 to 16. Completion of preprocessing led to next step i.e., model training.

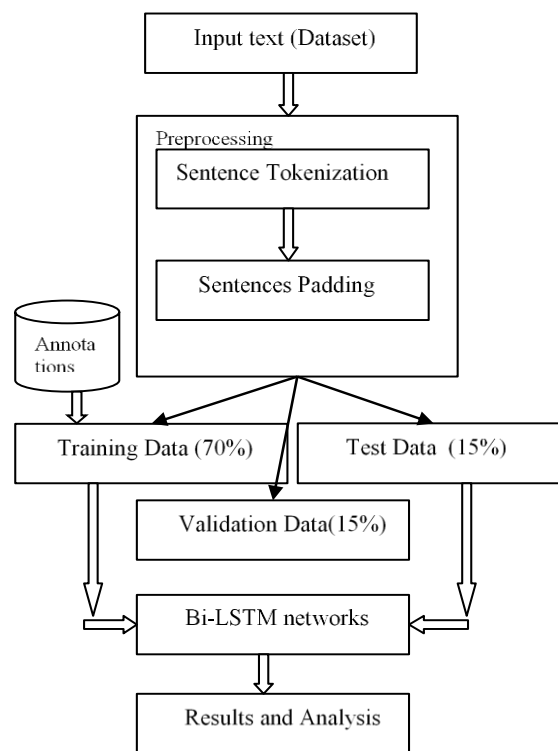


Figure 2. The methodology

2.3. Training the model

In this work we used Bi-LSTM model for training purpose. This is a sequence model with two LSTM layer first layer is used for processing input in the forward direction and the second layer is used for processing in the backward direction [15]. With having capability of working in two reverse directions this model is referred to as bidirectional LSTM, or Bi-LSTM [16]. It is widely applied to tasks involving NLP (natural language processing). The idea behind this method is that the model can better comprehend the relationship between sequences (e.g., knowing the words that proceeds and follow in a sentence)

by processing data in both directions [17], [18]. The Figure 3 shows the working of Bi-LSTM model. There are two directional activation function A_f and A_b which work in forward and backward directions respectively. The Figure 4 shows the design of single block of LSTM network.

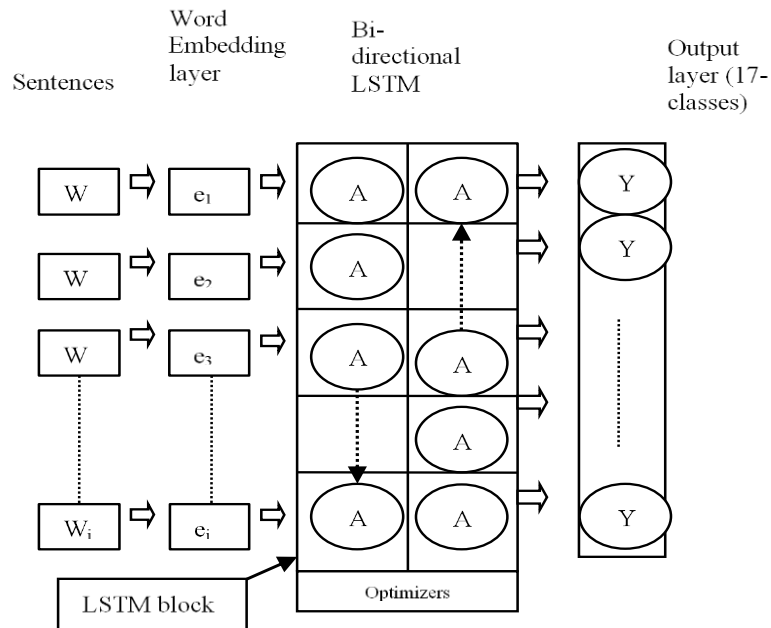


Figure 3. Single task Bi-LSTM based architecture

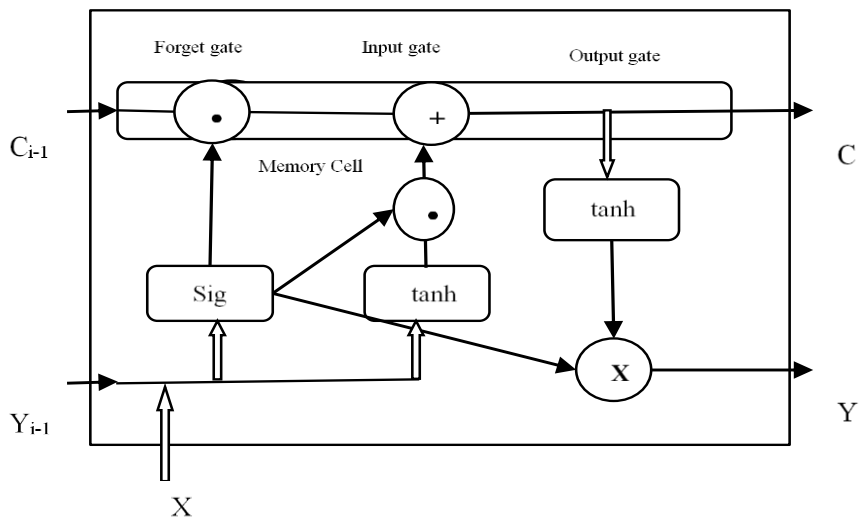


Figure 4. LSTM block architecture

There are three major components forget gate, input gate and output gate along with the cell memory which is a short term memory. The memory cell holds the current state of the block. The forget gate removes the noise from current state whereas input gate adds new information to it. Finally output gate updated the state value. C_{i-1} represents the previous cell value while X_i and Y_i represent the current input and output respectively. The sig (sigmoid) and tanh are activation functions while \bullet and $+$ are input based multiplication and addition operations respectively.

The impact of forget gate, input gate and output gate on cell state is defined as (1)-(3):

$$\text{Forget gate } F_i = \text{sig}(W_f X_i + U_f Y_{i-1} + b_f) \text{ updates } C_i = C_{i-1} \bullet F_i \tag{1}$$

$$\text{Input gate } I_i = \text{sig}(WIX_i + UIY_i - 1 + bI), a_i = \text{tanh}(WaXi + UaY_i - 1 + ba) \text{ updates } C_i = C_i - 1 \bullet I_i \bullet a_i \tag{2}$$

$$\text{Output gate } O_i = \text{sig}(WOX_i + UOY_i - 1 + bO) \text{ produces } Y_i = \text{tanh}(C_i) \bullet O_i \tag{3}$$

where W and U are weights and b is the bias. The activation functions are defined as in (4), (5).

$$\text{sig}(x) = 1/1 + e^{-x} \text{ ranges } [0,1] \tag{4}$$

$$\text{tanh}(x) = (e^x + e^{-x}) / (e^x - e^{-x}) \text{ ranges } [-1, +1] \tag{5}$$

A sentence can be defined as sequence of word [19], [20]. Thus, sentence S can be represented as $w_{1...i}$ where w represents a word. The task composed of various activities like sentence tokenization, word embedding, decision making and optimization. The input of the task is sentence whereas the output is $Y_{1...17}$ which represents the 17 tags of the tagsets. The sequence $w_{1...i}$ will be converted into numeral index $e_{1...i}$ to make it computational by activation functions. Further optimization functions are used to tune the estimated parameters of the learned model. An optimizer adjusts the values produced by activation functions so that loss value will be minimized [21]. Gradient descent is considered as popular choice among the class of optimizers. For larger datasets gradient descent approach is expensive one as gradient is to be calculated for huge numbers of the data values [22]. To tackle this challenge stochastic gradient descent approach is used for massive data-based problems such as NLP tasks [23]. In this approach randomly selected batches of data are processed in each iteration of learning instead of entire dataset. This makes stochastic gradient descent approach computationally inexpensive [24].

3. RESULTS AND DISCUSSION

The experiment employed a training set, test set and validation set with 70%, 15%, and 15% of preprocessed dataset respectively. Figure 5 shows the splitting of dataset into three sets. The training set is used for learning and validation set is used for fine tune the learning. The validation set is used for fine tuning purpose i.e., hyper parameter adjusting [25]. The test sets are used for checking the correctness the model [26].

The training set will feed into Bi-LSTM network to prepare the model to improve the learning with validation set 10 iterations had been performed. We have used ADAM optimization algorithm which is a first-order gradient-based optimization of stochastic objective functions. This algorithm is very cost effective with respect to computation time and memory space requirements.

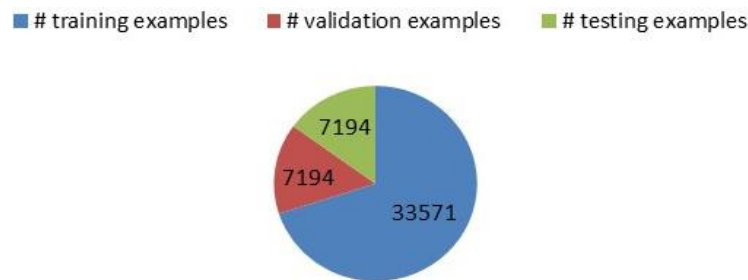


Figure 5. Splitting of dataset

The accuracy and loss values are help to evaluate system correctness [27], [28]. The stated parameter defined as in (6).

$$\text{Accuracy} = (X_{tp} + X_{tn}) / (X_{tp} + Y_{fp} + Y_{fn} + X_{tn}) \tag{6}$$

Where X_{tp} is number of true positives, X_{tn} is number of true negatives, Y_{fp} is number of false positives, and X_{fn} is number of false negatives.

The second parameter we consider for comparison is loss value. The loss value will be calculated using in (7).

$$\text{loss} = \sum nt = 1|Qt - Q't| \tag{7}$$

Where Q_t = actual data value and Q'_t = predicted data value. The loss function is simply the difference between Q_t and Q'_t [29], [30], whereas. N = number of samples used in the test set. Table 2 shows that in each iteration model improving its learning as accuracy value is increasing and loss value is decreasing.

Table 2. Accuracy and loss in 10 epochs

Epoch#	Accuracy	Loss
01	0.9661	0.1104
02	0.9733	0.0783
03	0.9788	0.0687
04	0.9816	0.0672
05	0.9842	0.0684
06	0.9861	0.0706
07	0.9873	0.0720
08	0.9887	0.0773
09	0.9900	0.0761
10	0.9911	0.0796

The Figure 6 depicts the comparison between the accuracy value and loss value in multiple epochs. The blue curved line shows the values for training set and green shows for validation set. This comparison is useful to show that validation is integral part of any machine learning based approach to improve the system performance.

The accuracy and loss value based on test dataset are 0.98 and 0.076 respectively. The previous approaches are based on random forest and CRF classifiers. Both the approaches were very popular for NLP task. The random forest is the decision tree-based classifier whereas CRF is a graph-based prediction model [20], [21]. In this work the both the models are also applied on the same dataset. The finding is shown in Table 3.

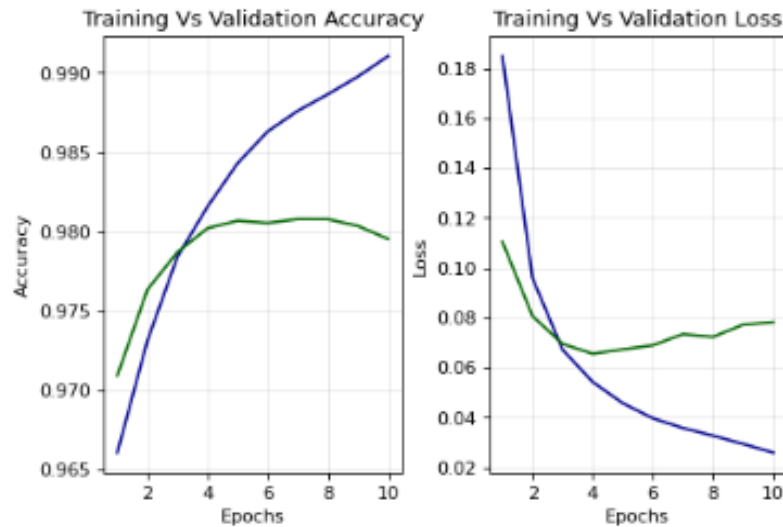


Figure 6. Training v/s validation loss in 10 epochs

Table 3. Comparison with CRF and random forest

Model	Random forest	CRF	Bi-LSTM
Accuracy	0.76	0.93	0.98

4. CONCLUSION

A technique for extracting location from social networks text data was provided in this paper. The methodology was divided into four activities: first the dataset preprocessing second is training the model, third is extracting the location in the text and finally testing the results. During the result analysis the proposed methodology is compared with CRF-based approach which was popularly used. Our

implementation of word indexing and Bi-LSTM classifier achieved 98% accuracy in 10 epochs as compared to CRF-base approaches which reaches to 96% in 10 epochs. Hence this approach is better for decision making on the locations of the events. This work will help the institutions such as disaster control management, police department and other administrative departments to get first-hand information about any event for which immediate action is required. Numerous geo-parsing potentials were found, providing avenues for future investigation. These include fixing typical spelling mistakes and clarifying ambiguous or inaccurate references to the event location on social media.





REFERENCES

- [1] K. D. Onal, P. Karagöz, and R. Çakici, "Türkçe twitter gönderilerinde lokasyon tanıma," in *2014 22nd Signal Processing and Communications Applications Conference, SIU 2014 - Proceedings*, Apr. 2014, pp. 1758–1761, doi: 10.1109/SIU.2014.6830590.
- [2] O. Ozdıkis, H. Oguztuzun, and P. Karagoz, "Evidential location estimation for events detected in Twitter," in *Proceedings of the 7th Workshop on Geographic Information Retrieval, GIR 2013*, Nov. 2013, pp. 9–16, doi: 10.1145/2533888.2533929.
- [3] R. Li, K. H. Lei, R. Khadiwala, and K. C. C. Chang, "TEDAS: a twitter-based event detection and analysis system," in *Proceedings - International Conference on Data Engineering*, Apr. 2012, pp. 1273–1276, doi: 10.1109/ICDE.2012.125.
- [4] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake shakes Twitter users: Real-time event detection by social sensors," in *Proceedings of the 19th International Conference on World Wide Web, WWW '10*, Apr. 2010, pp. 851–860, doi: 10.1145/1772690.1772777.
- [5] T. Sakaki, M. Okazaki, and Y. Matsuo, "Tweet analysis for real-time event detection and earthquake reporting system development," *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 4, pp. 919–931, Apr. 2013, doi: 10.1109/TKDE.2012.29.
- [6] J. Sankaranarayanan, H. Samet, B. E. Teitler, M. D. Lieberman, and J. Sperling, "TwitterStand: news in tweets," in *GIS: Proceedings of the ACM International Symposium on Advances in Geographic Information Systems*, Nov. 2009, pp. 42–51, doi: 10.1145/1653771.1653781.
- [7] G. A. Şeker and G. Eryiğit, "Initial explorations on using CRFs for Turkish named entity recognition," in *Proceedings of COLING 2012*, pp. 2459–2474, 2012.
- [8] Z. Xu *et al.*, "Crowdsourcing based description of urban emergency events using social media big data," *IEEE Transactions on Cloud Computing*, vol. 8, no. 2, pp. 387–397, Apr. 2020, doi: 10.1109/TCC.2016.2517638.
- [9] P. Kosmidis, C. Remoundou, K. Demestichas, I. Loumiotis, E. Adamopoulou, and M. Theologou, "A location recommender system for location-based social networks," in *Proceedings - 2014 International Conference on Mathematics and Computers in Sciences and in Industry, MCSI 2014*, Sep. 2014, pp. 277–280, doi: 10.1109/MCSI.2014.39.
- [10] Y. Liu, Q. Zhang, and L. Ni, "Opportunity-based topology control in wireless sensor networks," in *Proceedings - The 28th International Conference on Distributed Computing Systems, ICDCS 2008*, Jun. 2008, pp. 421–428, doi: 10.1109/ICDCS.2008.91.
- [11] P. R. Pooja and B. Hariharan, "An early warning system for traffic and road safety hazards using collaborative crowd sourcing," in *Proceedings of the 2017 IEEE International Conference on Communication and Signal Processing, ICCSP 2017*, Apr. 2017, vol. 2018-January, pp. 1203–1206, doi: 10.1109/ICCSP.2017.8286570.
- [12] Y. Liu, Y. Zhu, L. Ni, and G. Xue, "A reliability-oriented transmission service in wireless sensor networks," *IEEE Transactions on Parallel and Distributed Systems*, vol. 22, no. 12, pp. 2100–2107, Dec. 2011, doi: 10.1109/TPDS.2011.113.
- [13] V. Krishnamurthy and H. V. Poor, "A tutorial on interactive sensing in social networks," *IEEE Transactions on Computational Social Systems*, vol. 1, no. 1, pp. 3–21, Mar. 2014, doi: 10.1109/TCSS.2014.2307452.
- [14] J. Yuan *et al.*, "DMPPT control of photovoltaic microgrid based on improved sparrow search algorithm," *IEEE Access*, vol. 9, pp. 16623–16629, 2021, doi: 10.1109/ACCESS.2021.3052960.
- [15] K. Zahra, M. Imran, and F. O. Ostermann, "Automatic identification of eyewitness messages on twitter during disasters," *Information Processing and Management*, vol. 57, no. 1, p. 102107, Jan. 2020, doi: 10.1016/j.ipm.2019.102107.
- [16] W. Zhao, Z. Zhang, and L. Wang, "Manta ray foraging optimization: An effective bio-inspired optimizer for engineering applications," *Engineering Applications of Artificial Intelligence*, vol. 87, p. 103300, Jan. 2020, doi: 10.1016/j.engappai.2019.103300.
- [17] Y. Zhu and N. Yousefi, "Optimal parameter identification of PEMFC stacks using adaptive sparrow search algorithm," *International Journal of Hydrogen Energy*, vol. 46, no. 14, pp. 9541–9552, Feb. 2021, doi: 10.1016/j.ijhydene.2020.12.107.
- [18] L. Li, M. Bensi, Q. Cui, G. B. Baecher, and Y. Huang, "Social media crowdsourcing for rapid damage assessment following a sudden-onset natural hazard event," *International Journal of Information Management*, vol. 60, p. 102378, Oct. 2021, doi: 10.1016/j.ijinfomgt.2021.102378.
- [19] X. Li, H. Liu, W. Wang, Y. Zheng, H. Lv, and Z. Lv, "Big data analysis of the internet of things in the digital twins of smart city based on deep learning," *Future Generation Computer Systems*, vol. 128, pp. 167–177, Mar. 2022, doi: 10.1016/j.future.2021.10.006.
- [20] P. K. Diederik, "Adam: a method for stochastic optimization," *Conference Paper at ICLR*, 2015.
- [21] X. Liang, D. Cheng, F. Yang, Y. Luo, W. Qian, and A. Zhou, "F-HMTC: Detecting financial events for investment decisions based on neural hierarchical multi-label text classification," in *IJCAI International Joint Conference on Artificial Intelligence*, Jul. 2020, vol. 2021-January, pp. 4490–4496, doi: 10.24963/ijcai.2020/619.
- [22] Z. Deng, J. Ren, and L. B. Liu, "Short-term traffic flow prediction algorithm based on multiple CRF model," *Computer Engineering and Design*, vol. 38, no. 10, pp. 2887–2891, 2017.
- [23] J. O. Wallgrün, M. Karimzadeh, A. M. MacEachren, and S. Pezanowski, "GeoCorpora: building a corpus to test and train microblog geoparsers," *International Journal of Geographical Information Science*, vol. 32, no. 1, pp. 1–29, Sep. 2018, doi: 10.1080/13658816.2017.1368523.
- [24] M. Radke, P. Das, K. Stock, and C. B. Jones, "Detecting the geospatialness of prepositions from natural language text," in *14th International Conference on Spatial Information Theory (COSIT 2019)*, 2019.
- [25] M. Gritta, M. T. Pilehvar, N. Limsopatham, and N. Collier, "What's missing in geographical parsing?," *Language Resources and Evaluation*, vol. 52, no. 2, pp. 603–623, Mar. 2018, doi: 10.1007/s10579-017-9385-8.
- [26] H. Chen, M. Vasardani, and S. Winter, "Georeferencing places from collective human descriptions using place graphs," *Journal of Spatial Information Science*, vol. 17, no. 17, pp. 31–62, Dec. 2018, doi: 10.5311/JOSIS.2018.17.417.





- [27] M. Won, P. Murrieta-Flores, and B. Martins, "Ensemble named entity recognition (NER): Evaluating NER tools in the identification of place names in historical corpora," *Frontiers in Digital Humanities*, vol. 5, Mar. 2018, doi: 10.3389/fdigh.2018.00002.
- [28] S. Clematide *et al.*, "Crowdsourcing toponym annotation for natural features: how hard is it? corpus in GIScience: going beyond butterfly collecting," *Workshop at GIScience. Melbourne, Australia*, 2018.
- [29] F. Melo and B. Martins, "Automated geocoding of textual documents: a survey of current approaches," *Transactions in GIS*, vol. 21, no. 1, pp. 3–38, Jun. 2017, doi: 10.1111/tgis.12212.
- [30] R. Paul, J. Arkin, D. Aksaray, N. Roy, and T. M. Howard, "Efficient grounding of abstract spatial concepts for natural language interaction with robot platforms," *International Journal of Robotics Research*, vol. 37, no. 10, pp. 1269–1299, Jun. 2018, doi: 10.1177/0278364918777627.

BIOGRAPHIES OF AUTHORS



Dharmendra Mangal     is assistant professor in Medi-Caps University, Indore also Ph.D. scholar at Institute of Engineering and Technology, DAVV, Indore. The research interest includes NLP and machine learning. He can be contacted at email: mangaldharmendra83@gmail.com.



Dr. Hemant Makwana     is associate professor at Institute of Engineering and Technology DAVV, Indore. The research interest includes Computer Graphics and Computer Architecture. He is Ph.D. in computer engineering. He can be contacted at email: hmakwana@ietdavv.edu.in.