EasyChair Preprint

№ 6333

Named Entity Recognition (NER) for Social Media Tamil Posts Using Deep Learning with Singular Value Decomposition

Panner Selvam Kathiravan and Rajiakodi Saranya

August 21, 2021

NAMED ENTITY RECOGNITION (NER) FOR SOCIAL MEDIA TAMIL POSTS USING DEEP LEARNING WITH SINGULAR VALUE DECOMPOSITION

P. Kathiravan[1], R. Saranya[2]
[1]*Research Scholar, Department of Computer Science, Central University of Tamil Nadu, India,*
E-mail: kathiravan.pa@gmail.com
[2] *Assistant Professor, Department of Computer Science, Central University of Tamil Nadu, India,*
E-mail: saranya@cutn.ac.in

Introduction

The Named Entity Recognition (NER) is a part of Information Extraction (IE) which is an emerging area of research and application that explores how to discover the knowledge (information) from a huge amount of text. There are so many data resources are available on the internet like social media, e-commerce sites, blogs, news portals, personal websites, and so on, where people share their thoughts and opinions in their native language. NER is a process of identifying named entities such as a person, organizations, locations, time, and amount from a given text data (Srinivasan, R., et al., 2019). It was introduced in 1996 at the sixth Message Understanding Conference (MUC-6). The various NER systems were proposed at that time to enhance the information extraction tasks such as Rule-based NER, Machine Learning-based NER, and Hybrid NER (Mansouri, A., et al., 2008). It can be used in many applications such as business intelligence, crime prediction, fraud detection, and recommendation systems.

The proper identification and extraction of named entities from Tamil text documents are very difficult and challenging tasks because of the semantic ambiguity i.e., the same word gives different meaning and its free word order. The following example describes the semantic ambiguity in the following Tamil sentence. "எங்கள் பல்கலைக்கழகம் முதல் இடத்தை வென்றது" describes the achievement of the university. Here the university is an organization. But in this case, "கவர்னர் எங்கள் பல்கலைக்கழகத்திற்கு வந்தார்" here the term University referred to as location. In the above example, the term "பல்கலைக்கழகம்" gives different meaning according to the context of the sentence.

The next challenge of the Tamil text is free word order ie., A sentence that gives the same meaning even though the word order is rearranged (Sankaravelayuthan, R., et al., 2019). It is described by the following example.

1. அவர் கல்லூரிக்கு செல்கிறார் (SOV).
2. அவர் செல்கிறார் கல்லூரிக்கு (SVO).
3. செல்கிறார் அவர் கல்லூரிக்கு (VSO).
4. செல்கிறார் கல்லூரிக்கு அவர் (VOS).
5. கல்லூரிக்கு செல்கிறார் அவர் (OVS).
6. கல்லூரிக்கு அவர் செல்கிறார் (OSV).

In the above example, all six sentences give the same meaning even though the word order is changed. It can be correctly identified by a human being but not by the NER system. To overcome the above said challenges we proposed an enhanced NER system by integrating feature extracting techniques such as REGEXP, Morphological analyzer, content feature

extraction, and Singular Value Decomposition (SVM) with the various Deep Learning Architectures such as Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), Long Short Term Memory (LSTM), and Gated Recurrent Unit (GRU).

Related works

Different methods such as Rule-based NER, Machine Learning-based NER, and Hybrid NER were used so far to develop a NER system to identify the named entities from given texts.

Srinivasan, R., et al., (2019) used supervised learning algorithms to extracts the named entities from the 1028 Tamil documents which were collected from Forum for Information Retrieval Evaluation (FIRE). After extracting the features by using REGEX, the morphological and context features were extracted using POS tagger. Then they applied the naive Bayes algorithm to build a classifier with 83.54% F-measure.

Theivendiram, P., et al.,(2016) developed the novel NER system Margin-Infused Relaxed Algorithm (MIRA) and compared it with Conditional Random Fields (CRF) for Tamil BBC news text data and the MIRA gave 81.38% F1-measure whereas CRF has given 79.13%.

Hariharan, V., et al., (2019) used Long-short Term Memory (LSTM) and the fastText word embedding technique to develop a NER system. The data set was collected from Tamil Wikipedia and FIRE-18. The result is compared with various word embedding techniques such as GloVe. The LSTM with fastText has given prominent results than LSTM with GloVe.

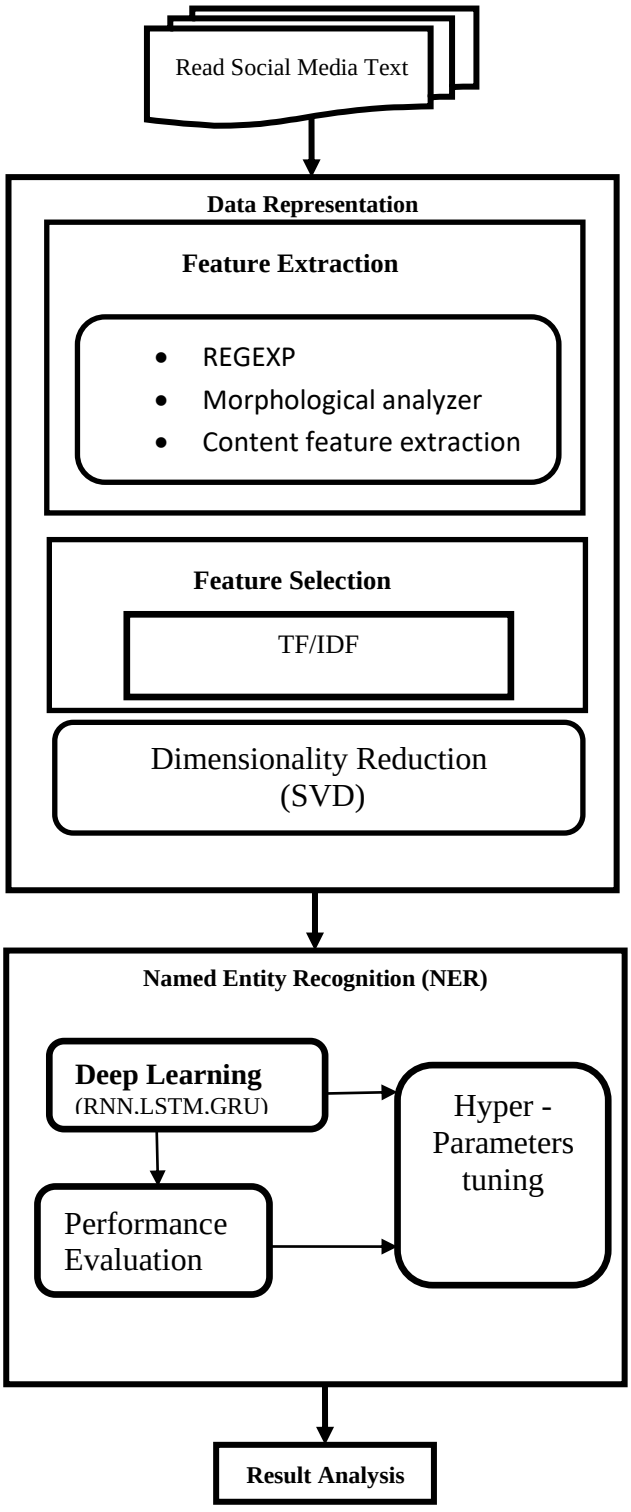
Abinaya, N., et al., (2015) proposed a statistical and supervised technique namely Random Kitchen Sink algorithm for Tamil NER and the result has compared with SVM and CRF algorithms. The RKS has given 87.88% accuracy which is higher than SVM and CRF.

Proposed Methodology

In this section, our proposed model and the various deep learning architectures are discussed in detail. The following Fig. 1 describes the detailed process of the proposed Named Entity Recognition (NER) systems for Tamil social media posts. The overall proposed methodology is segregated into four sections. The first section for data collection, the second section for Data preparation, the third section of this methodology is to implement deep learning with hyperparameter tuning and the final section is to extract the result.

In the first section, the Tamil social media posts are extracted from various social media such as Facebook, Twitter, and Instagram by using their APIs. The extracted input data will be sent to the data preparation section, which includes various activities like feature extraction, feature selection, and dimensionality reduction. The Feature extraction process specifies the different feature extracting techniques such as REGEXP and Morphological analyzer.

Fig 1: Detailed process of NER system.



In a given data set, the data might include numerical features like amount, date and time, and natural language texts (example: சிவா 3.4.2021 க்கு முன் தேர்வு கட்டணம் ரூ .3000 செலுத்த வேண்டும்). The numerical values are extracted using the REGEXP technique, which is used to extract the regular expression patterns and the morphological analyzer helps to tags the word with correct Part-of-speech. For example, Morphology analyzer analyzes the term பல்கலைக்கழகத்திற்கு as பல்கலைக்கழக <Entity>, த் <Santhi> and திற்கு <Locative case> then it extracts க்கு from the data and fixed as <Location NE>. The second part of the data preparation section is feature selection, which handles the series of tokens from the feature extraction phase to form a term-document vector-matrix using the *Term Frequency* (TF), *Document Frequency,* and *Inverse Document Frequency* (IDF). The TF/IDF is used to find the importance of a particular word in a given input data. Here, the data set is normalized to unit length and represented by term-document matrix otherwise called as vector space model of the entire data set (Arivoli et al., 2017).

After receiving the term-document matrix, The Singular Value Decomposition (SVD) will be applied to reduce its dimensionality. Then the various Deep Learning Architectures like Recurrent Neural Networks (RNN), Long Short Term Memory (LSTM), and Gated Recurrent Unit (GRU) are proposed to develop a NER system, which will enhance the better entity recognition from the Social Media Tamil Text data set.

In general, the Recurrent Neural Networks (RNN) is used in real-time streaming data like Social Media posts and Time series analysis. It effectively deals with the sequential data but not very long sequences (Subramani et al., 2019). So, the improved versions of RNN i.e., LSTM and GRU will be used in this work. Then the accuracy will be calculated by using various parameters like kappa statistic, specificity, sensitivity, F-measure, etc., and finally, the most suitable deep learning architecture with the hyperparameter will be recommended for Tamil NER systems.

Reference

[1] Abinaya, N., Kumar, M. A., & Soman, K. P. (2015). Randomized kernel approach for named entity recognition in Tamil. Indian Journal of Science and Technology, 8(24).

[2] Arivoli, P. V., & Chakravarthy, T. (2017). Document classification using machine learning algorithms-a review. *International Journal of Scientific Engineering and Research (IJSER)*, *5*(2), 48-54.

[3] Hariharan, V., Kumar, M. A., & Soman, K. P. (2019). Named Entity Recognition in Tamil Language Using Recurrent Based Sequence Model. In Innovations in Computer Science and Engineering (pp. 91-99). Springer, Singapore.

[4] Mansouri, A., Affendey, L. S., & Mamat, A. (2008). Named entity recognition approaches. International Journal of Computer Science and Network Security, 8(2), 339-344.

[5] Sankaravelayuthan, R., & Vasuki, C. D. G. (2019). English to Tamil Machine Translation System Using Parallel Corpus. LAP LAMBERT Academic Publishing.

[6] Srinivasan, R., & Subalalitha, C. N. (2019, July). Automated Named Entity Recognition from Tamil Documents. In *2019 IEEE 1st International Conference on Energy, Systems and Information Processing (ICESIP)* (pp. 1-5). IEEE.

[7] Subramani, S., Michalska, S., Wang, H., Du, J., Zhang, Y., & Shakeel, H. (2019). Deep Learning for Multi-Class Identification from Domestic Violence Online Posts. IEEE Access, 7, 46210–46224.

[8] Theivendiram, P., Uthayakumar, M., Nadarasamoorthy, N., Thayaparan, M., Jayasena, S., Dias, G., & Ranathunga, S. (2016, April). Named-entity-recognition (ner) for Tamil language using margin-infused relaxed algorithm (mira). In International Conference on Intelligent Text Processing and Computational Linguistics (pp. 465-476). Springer, Cham.