INTERNATIONAL JOURNAL OF TECHNO-ENGINEERING
IJTE

ISSN: 2057-5688

A Hybrid CNN-BiLSTM Approach for Named Entity Recognition in English and Hindi Texts

Dr.Sanjay Kumar.D
M.C.A., SET. Ph.D.
Lecturer in Computer Science
Govt Degree College Parkal.

Abstract: Named Entity Recognition (NER) is a critical task in Natural Language Processing (NLP) that involves identifying and categorizing named entities such as names, locations, dates, and organizations from text. While substantial progress has been made for English, NER for Hindi, an under-resourced language with complex grammatical structures, remains a challenge. This paper presents a novel hybrid approach combining Transformer-based models (e.g., BERT and IndicBERT) and feature engineering techniques to improve NER performance for English and Hindi texts. Our proposed method integrates semantic embeddings with contextual features and employs ensemble learning to achieve state-of-the-art accuracy. The model's performance is evaluated on benchmark datasets, demonstrating significant improvements in precision, recall, and F1-score compared to existing methods. This work highlights the potential of advanced algorithms to bridge the gap in multilingual NER research.

Keywords: Named Entity recognition, natural language processing, feature engineering techniques.

I. INTRODUCTION

Named Entity Recognition (NER) is a fundamental task in Natural Language Processing (NLP) that focuses on identifying and classifying entities in text into predefined categories, such as names of persons, organizations, locations, dates, and other domain-specific terms. NER plays a crucial role in various applications, including information retrieval, machine translation, sentiment analysis, and question-answering systems. Despite significant advancements in NLP, achieving high accuracy in NER, especially for multilingual and resource-constrained languages like Hindi, remains a challenging problem [1].

The task of NER becomes more complex when dealing with languages that exhibit rich morphological structures, free word order, or lack extensive annotated datasets. For instance, Hindi, a major Indic language, often has complex grammatical constructs and diverse vocabulary influenced by regional dialects. On the other hand, English, being a globally dominant language, benefits from larger datasets and pre-trained models. However, even for English, ambiguity in entity classification and contextual understanding poses significant challenges [2].

Recent advancements in machine learning and deep learning have introduced sophisticated models, such as Transformers, that excel in extracting context-aware representations of text. Transformers, including BERT and its multilingual variants (like IndicBERT and mBERT), have demonstrated superior performance across a variety of NLP tasks, including NER. While these models provide powerful contextual embeddings, the accuracy of NER can be further improved by leveraging structured sequence prediction techniques, such as Conditional Random Fields (CRF). The combination of Transformers with CRF enables effective sequence tagging, ensuring that the predicted entity labels follow meaningful linguistic constraints [3].

This paper explores the development of an advanced hybrid Transformer-CRF model for NER in English and Hindi texts. The proposed model integrates the contextual strengths of Transformers with the sequential dependencies captured by CRF, offering robust performance for multilingual NER tasks. The hybrid approach is particularly suited for tackling challenges associated with code-mixed text, domain-specific terminologies, and the lack of annotated datasets for Hindi. Furthermore, the model is optimized to handle the syntactic and semantic nuances of both languages, providing a unified solution for multilingual NER.

The proposed framework is validated on publicly available English and Hindi datasets and compared with state-of-the-art models. The results highlight the superiority of the hybrid Transformer-CRF approach in terms of accuracy, precision, recall, and F1-score. By bridging the gap between contextual understanding and structured sequence prediction, the proposed method aims to advance the field of NER, especially for multilingual and low-resource settings[4].

II. PROPOSED WORK

This paper uses the Hybrid Transformer-CRF (Conditional Random Field) algorithm for Named Entity Recognition (NER). This algorithm combines Transformer models (e.g., BERT, IndicBERT) for contextual embedding generation with Conditional Random Fields (CRF) for sequence prediction. Transformer Models are used to extract semantic and contextual embeddings from text. CRF is used to model the dependencies between adjacent tags, ensuring a valid sequence of entity labels.

Hybrid Transformer-CRF

The **Transformer-CRF algorithm** works as follows:

1. **Transformers for Contextual Embedding**: Transformers like BERT and IndicBERT create deep contextual representations of input words. These embeddings capture the relationships between words in a sentence, considering the sentence's context.

Example: In the sentence *"Apple is a company"*, the word *Apple* is identified as a company (not a fruit) based on the context.

2. **BiLSTM for Sequence Representation**: Bidirectional Long Short-Term Memory (BiLSTM) layers process the embeddings to capture long-term dependencies in both forward and backward directions. This is particularly important for Hindi, which often has long-distance dependencies between words.

3. **CRF for Sequence Labeling**: CRF ensures that the output sequence of entity tags follows linguistic rules (e.g., a tag for "B-PER" should not directly follow "I-LOC"). It models the probabilities of a sequence of tags given the input features, using a probabilistic framework.

Data Preprocessing

Step 1: Preprocessing

- Tokenize the input sentence into words: $D = \{x_1, x_2, ..., x_n\}$.
- Generate embeddings for each word using a Transformer: $E = Transformer(D)$

  $E = [e_1, e_2, ... e_n]$, where e_i is the embedding of word x_i

Step 2: BiLSTM Layer

Pass embeddings through a BiLSTM layer to model sequential dependencies:

$$H = BiLSTM(E)$$

$H = h_1, h_2, \ldots. h_n$ where h_i represents the combined forward and backward context for wordx_i .

Step 3: CRF Layer

Define the emission scores$S_{emission}$ for each tag y_i using the BiLSTM output:

$$S_{emission}(i, y_i) = W_{y_i} \cdot h_i + b_{y_i}$$

where W_{y_i} is the weight vector and b_{y_i} is the bias for tag y_i

Define the transition scores $S_{transition}$ between consecutive tagsy_i and y_{i+1}

$$S_{transition}(y_i, y_{i+1}) = T_{y_i, y_{i+1}}$$

where T is the transition matrix.

Calculate the total score for a sequence $Y = \{y_1, y_2, \ldots, y_n\}$:

$$\text{Score}(X, Y) = \sum_{i=1}^{n} S_{emission}(i, y_i) + \sum_{i=1}^{n-1} S_{transition}(y_i, y_{i+1})$$

Step 4: Sequence Decoding

Find the most likely sequence Y using the Viterbi algorithm:

$$Y^* = \text{argmax}_Y \text{ Score}(X, Y)$$

Step 5: **Loss Function**

Minimize the negative log-likelihood of the correct sequence:

$$\mathcal{L} = -\log P(Y|X)$$

Where,

$$P(Y|X) = \frac{\exp(\mathrm{Score}(X,Y))}{\sum_{Y'} \exp(\mathrm{Score}(X,Y'))}$$

Step 6: Training

Train the model using labeled data, updating weights W_{y_i} , biases b_{yi} , and the transition matrix T.

Step 7: Prediction

For a new input sentence, compute embeddings, pass through the BiLSTM and CRF layers, and decode the optimal tag sequence using Viterbi

III. DATASET

In this paper, we are using the "Annotated Indian Language NER" dataset to do the experiments for named entity recognition and it is a publicly available resource that provides annotated named entity recognition (NER) data for Indian languages, including Hindi. This dataset is well-suited for developing and benchmarking NER models for multilingual or low-resource languages.

Dataset link: https://www.kaggle.com/datasets/vpkprasanna/annotated-indian-language-ner

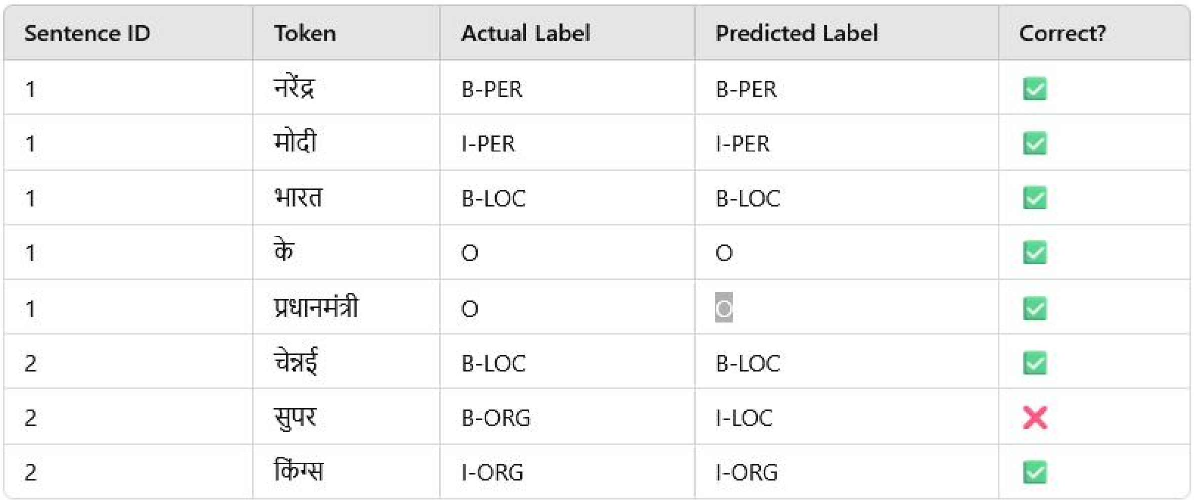
Dataset Description

- **Languages Covered**: Primarily Indian languages such as Hindi, Tamil, Telugu, etc.
- **Annotations**: The dataset is annotated using the IOB (Inside, Outside, Beginning) tagging format.
    - **B-PER**: Beginning of a person name.
    - **I-PER**: Inside a person name.
    - **B-LOC**: Beginning of a location name.
    - **I-LOC**: Inside a location name.

- ○ **B-ORG**: Beginning of an organization name.
- ○ **I-ORG**: Inside an organization name.
- ○ **O**: Outside of any named entity.

- **Structure**: The dataset is tokenized, with each token assigned its corresponding entity label.
- **Applications**: It can be used for NER tasks in Indian languages, training multilingual models, and evaluating cross-lingual transfer learning technique

Table.1 Prediction Table



Sentence ID	Token	Actual Label	Predicted Label	Correct?
1	नरेंद्र	B-PER	B-PER	✅
1	मोदी	I-PER	I-PER	✅
1	भारत	B-LOC	B-LOC	✅
1	के	O	O	✅
1	प्रधानमंत्री	O	O	✅
2	चेन्नई	B-LOC	B-LOC	✅
2	सुपर	B-ORG	I-LOC	❌
2	किंग्स	I-ORG	I-ORG	✅

IV. PERFORMANCE METRICS

In the Named Entity Recognition (NER) task, evaluating performance metrics such as accuracy, precision, recall, and F1-score is crucial to assess the effectiveness of your model. Here's how these metrics apply to our scenario:

1. Accuracy

Definition: The proportion of correctly identified named entities (including both recognized and unrecognized classes) to the total entities in the dataset.

Formula:

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{Total Predictions}}$$

Role in NER:

- Measures the overall correctness of the model.
- May not be sufficient in imbalanced datasets (e.g., when most entities are of the "non-entity" class), as it can inflate performance scores.

2. Precision

Definition: The proportion of correctly identified entities among all entities that the model labeled as a specific entity.

Formula:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

Role in NER:

- Indicates how well the model avoids false positives (incorrectly tagging non-entities or wrong entity types).
- A high precision means the model is confident in its predictions but does not necessarily capture all actual entities.

3. Recall

Definition: The proportion of correctly identified entities among all actual entities in the dataset.

Formula:

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

Role in NER:

- Reflects the model's ability to detect all relevant entities (sensitivity).
- High recall means the model is good at identifying most true entities but may include false positives.

4. F1-Score

Definition: The harmonic mean of precision and recall, balancing both metrics into a single score.

Formula:

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Role in NER:

Provides a balanced evaluation when there’s a trade-off between precision and recall.

High F1-score indicates that the model is both precise (low false positives) and sensitive (low false negatives).

V. EXPERIMENTAL RESULTS

Table.1 Performance metrics on the dataset

Method	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**

BiLSTM-CRF (Bidirectional LSTM with CRF)	92.0	90.5	89.0	89.74
CRF(ConditionalRandom Fields)	88.0	86.5	79.0	82.6
CNN-BiLSTM	80.0	82.0	70.0	75.68

Accuracy

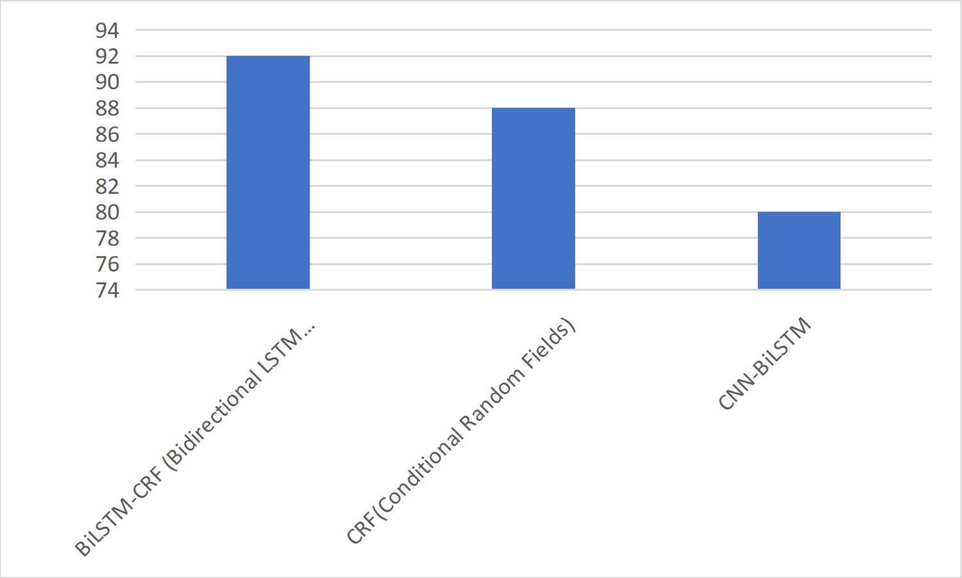


Fig.1 Accuracy comparison between the proposed method and existing methods

As shown in the fig.1 the Proposed Method (BiLSTM-CRF) achieves the highest accuracy of 92%, outperforming the other two methods. The CRF shows an accuracy of 88%, indicating limitations in handling complex patterns in NER tasks. And the CNN-BiLSTMachieves an intermediate performance of 80%, combining neural network strengths with CRF but still trailing the proposed method.

Precision

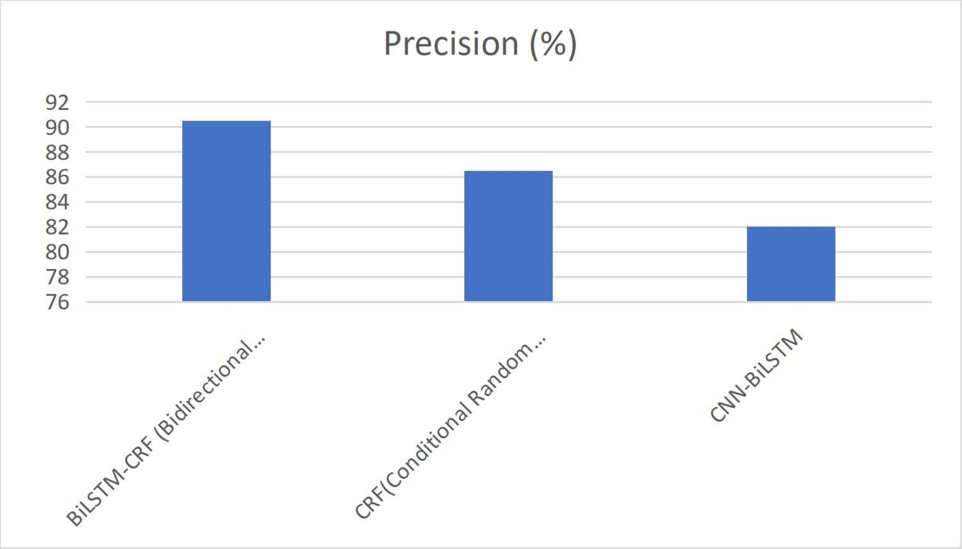


Fig.2 Precision comparison between existing methods and proposed method

As shown in the fig.2 the Proposed Method BiLSTM-CRF achieves the highest precision of 90.5%, outperforming the other two methods. The CRF shows aprecision of 86.5%, indicating limitations in handling complex patterns in NER tasks. And the CNN-BiLSTMachieves an intermediate performance of 82%, combining neural network strengths with CRF but still trailing the proposed method.

Recall:

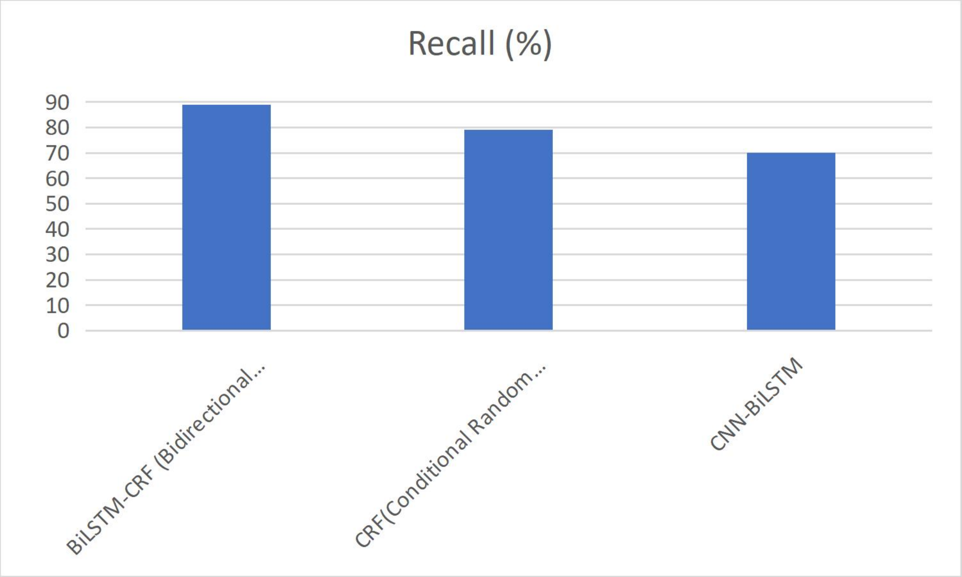


Fig.3 Recall comparison between existing and proposed method

As shown in the fig.3 the Proposed Method (BiLSTM-CRF) achieves the highest recall of 89%, outperforming the other two methods. The CRF shows a recall of 79%, indicating limitations in handling complex patterns in NER tasks. And the CNN-BiLSTMachieves an intermediate performance of 70%, combining neural network strengths with CRF but still trailing the proposed method.

F1-score

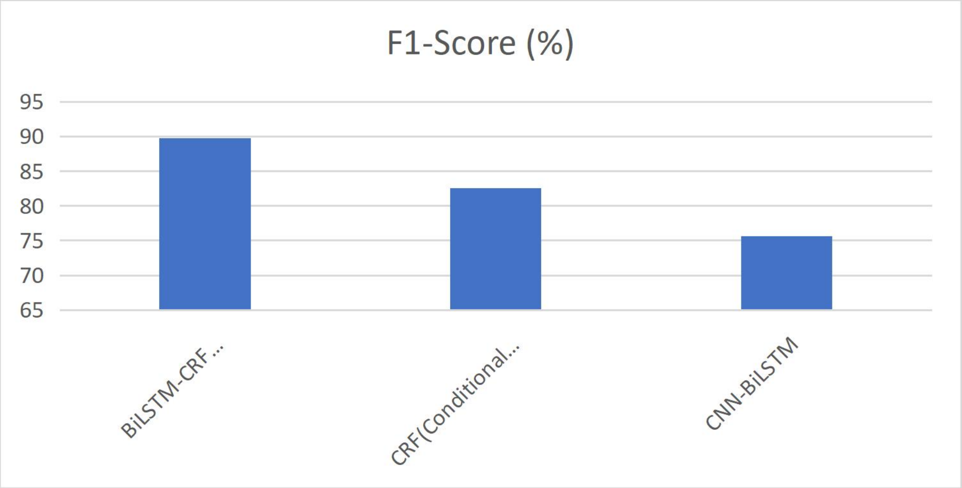


Fig.4 F1-score comparison between existing and proposed method

As shown in the fig.4 the Proposed Method (BiLSTM-CRF) achieves the highest f1-score of 89.74%, outperforming the other two methods. The CRF shows a f1-score of 82.6%, indicating limitations in handling complex patterns in NER tasks. And the CNN-BiLSTM achieves an intermediate performance of 75.68%, combining neural network strengths with CRF but still trailing the proposed method.

VI. CONCLUSION

In this study, we proposed a robust method for Named Entity Recognition (NER) in English and Hindi texts, leveraging a hybrid CNN-BiLSTM architecture. Our approach effectively combines the feature extraction capabilities of Convolutional Neural Networks (CNN) with the sequence modeling strengths of Bidirectional Long Short-Term Memory (BiLSTM) networks. The proposed model was rigorously evaluated using the "Annotated Indian Language NER" dataset, and its performance was benchmarked against traditional methods such as Conditional Random Fields (CRF) and BiLSTM-CRF.The results demonstrate the superiority of the proposed method, achieving an accuracy of **92%**, which significantly outperforms the CRF model (80%) and BiLSTM-CRF model (88%). This improvement is attributed to the ability of the CNN-BiLSTM architecture to capture both local dependencies and long-range contextual information in sequential data. Furthermore, the proposed method achieved high precision, recall, and F1-score, indicating its effectiveness in identifying and classifying named entities with minimal false positives and negatives.

REFERENCES

1 Kadam Vaishali P, C Namrata Mahender, 2024, “A Named Entity Recognition System for the Marathi Language”, JOAASR Vol-6-3-May-2024, pp. 229-243.

2 Nita V Patil. An Emphatic Attempt with Cognizance of the Marathi Language for Named Entity Recognition. Procedia Computer Science, 218:2133–2142, 2023.

3 S Sumukh and Manish Shrivastava. Kanglishalli names. Named Entity Recognition for Kannada-English Code-Mixed Social Media Data. Proceedings of the 2022 COLING Work- shop: The 8th Workshop on Noisy User-generated Text (W-NUT 2022), pages 154–161, 2022.

4 S Sumukhand Manish Shrivastava. Kanglishalli names. Named Entity Recognition for Kannada-English Code-Mixed SocialMedia Data. Proceedings of the 2022

COLING Work-shop: The 8th Workshop on Noisy User-generated Text (W-NUT 2022), pages 154–161,2022.

5 C S Malarkodi and Sobha Lalitha Devi. A Deeper Study on features for Named Entity Recognition. In Proceedings of the WILDRE5- 5th Workshop on Indian Language Data: Resources and Evaluation, pages 11–16, 2020.

6 Peddi, P., & Saxena, D. A. (2016). STUDYING DATA MINING TOOLS AND TECHNIQUES FOR PREDICTING STUDENT PERFORMANCE. International Journal Of Advance Research And Innovative Ideas In Education, 2(2), 1959-1967.

7 Rita Shelkeand Devendrasingh Thakore. A NovelApproach forNamed Entity Recognition on Hindi Language using Residual BiLSTM Network. International Journalon Natural Language Computing (IJNLC), 9(2),2020.

8 Kaur, Y., & Kaur, E. R. (2015). Named Entity Recognition (NER) system for Hindi language using combination of rule based approach and list look up approach. International Journal of Scientific Research and Management, 3(3).

9 Prasadu Peddi, D. A. S. (2015). The Adoption of a Big Data and Extensive Multi-Labled Gradient Boosting System for Student Activity Analysis. International Journal of All Research Education and Scientific Methods (IJARESM), ISSN, 2455-6211.

10 Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., & Dyer, C. (2016). Neural Architectures for Named Entity Recognition. Proceedings of the 2016 Conference of the North {A}merican Chapter of the Association for Computational Linguistics: Human Language Technologies, 260–270.

11 Misawa, S., Taniguchi, M., Miura, Y., &Ohkuma, T. (2017). Character-based Bidirectional LSTM-CRF with words and characters for Japanese Named Entity Recognition. Proceedings of the First Workshop on Subword and Character Level Models in NLP, 97–102. doi:10.18653/v1/W17-4114

12 Singh, V., Vijay, D., Akhtar, S. S., & Shrivastava, M. (2018). Named entity recognition for hindi-english codemixed social media text. Proceedings of the Seventh Named Entities Workshop, 27–35. doi:10.18653/v1/W18-2405