applied sciences

MDPI

Article

Improving Named Entity Recognition for Social Media with Data Augmentation

Wenzhong Liu [1,2] and Xiaohui Cui [1,3,*]

[1] The Engineering Research Center of Cyberspace, Yunnan University, Kunming 650504, China
[2] The Pilot School of Software, Yunnan University, Kunming 650504, China
[3] The School of Cyber Science and Engineering, Wuhan University, Wuhan 430001, China
* Correspondence: xcui@whu.edu.cn

Abstract: Social media is important for providing text information; however, due to its informal and unstructured nature, traditional named entity recognition (NER) methods face the challenge of achieving high accuracy when dealing with social media data. This paper proposes a new method for social media named entity recognition with data augmentation. First, we pre-train the language model by using a bi-directional encoder representation of the transformer (BERT) to obtain a semantic vector of the word based on the contextual information of the word. Then, we obtain similar entities via data augmentation methods and perform substitution or semantic transformation on these entities. After that, the input into the Bi-LSTM model is trained and then fused and fine-tuned to obtain the best label. In addition, our use of the self-attentive layer captures the essential information of the features and reduces the reliance on external information. Experimental results on the WNUT16, WNUT17, and OntoNotes 5.0 datasets confirm the effectiveness of our proposed model.

Keywords: social media; named entity recognition; data augmentation; BERT

Citation: Liu, W.; Cui, X. Improving Named Entity Recognition for Social Media with Data Augmentation. *Appl. Sci.* **2023**, *13*, 5360. https://doi.org/10.3390/app13095360

Academic Editor: José Salvador Sánchez Garreta

Received: 1 March 2023
Revised: 21 April 2023
Accepted: 23 April 2023
Published: 25 April 2023

1. Introduction

In recent years, social media platforms have emerged as a vital channel for accessing information. The vast amount of text data generated on these platforms presents an opportunity for natural language processing (NLP) research. Named entity recognition (NER) is a critical task in NLP, which involves identifying and categorizing entities in text data. NER has been extensively studied and applied in various domains, such as information extraction [1], machine translation [2], and question-answering [3].

However, NER in social media data continues to be a challenging task. The unique characteristics of social media text, such as the use of non-standard language, creative spelling, slang, and abbreviations, render it difficult to apply strict syntactic rules to the text data. Additionally, social media entities are diverse and sparse, which renders the task difficult to generalize [4]. Traditional NER approaches have attempted to address these challenges by using existing gazetteers and embeddings that are trained on large social media texts and external features in order to improve social media NER [5,6]. However, these approaches rely on additional efforts to obtain the extra information and are vulnerable to noise in the resultant information. Training embeddings for the social media domain may introduce many unfamiliar expressions to the vocabulary, leading to inaccurate entity classification.

Data augmentation has been extensively researched and widely applied in NLP tasks, effectively to resolve challenges such as limited data volume, sample imbalance, and specific domain social media NER tasks. By introducing additional samples and diversity, data augmentation can improve the robustness and generality of models while reducing overfitting. Several data augmentation techniques have been commonly used, such as the utilization of synonyms, antonyms, and word form variations, which increase the amount

and diversity of training data [7–9]. Additionally, semantic transformation techniques can make use of semantic role annotation information to perform semantic transformation, thereby further increasing data diversity and richness [10]. In addition to traditional data augmentation techniques, language model generation has also demonstrated its effectiveness as a data augmentation method [11] by generating new samples by using language models, it can help alleviate the impact of insufficient training data while improving the generalization ability of the model.

To address the above issues and improve the robustness and accuracy of NER in social media, this paper proposes a framework that is based on data augmentation for social media NER (DA-NER). First, we use a pre-training language model, the bidirectional encoder representations from transformers (BERT) [12], to extract the semantic vector of each word based on its contextual information. Then, we obtain similar entities via data augmentation methods and subject the entities to replacement or semantic transformation. Specifically, we calculate the cosine similarity between the entity embeddings in the training data and the entities of the predefined knowledge base by pre-training the word embedding model GloVe [13]; then, we select the entities with high cosine similarity and weight them to the corresponding entities to obtain the semantic vector after data augmentation. This balances data augmentation and context to improve model performance and generalization ability. The two vectors are then fed into the Bi-LSTM-CRF [14] model for training and then fused and fine-tuned to obtain the optimal labels. In addition, our use of the self-attention layer captures the underlying information of the features and reduces the dependence on external information. Our approach is experimented on three widely used benchmark datasets in the social media domain. The experimental results show that our approach of the DA-NER model is effective and achieves state-of-the-art results on all datasets.

The contributions of this paper can be concluded as follows:

- We apply data augmentation techniques to social media NER, which effectively address the challenges of data sparsity and category imbalance. Experimental results demonstrate that the incorporation of data augmentation can significantly improve entity recognition performance and enhance the generalization and robustness of the model.
- In this paper, the attention mechanism is integrated into the Bi-LSTM model, which assigns weights to different words use the selection type. This enables the model to leverage contextual semantic associations and effectively address the challenges associated with acquiring local features.
- We conducted extensive experiments on three benchmark datasets for social media NER. The results demonstrate that our method outperforms other approaches and achieves outstanding performance on social media NER benchmark datasets.

2. Related Work

2.1. Named Entity Recognition

Named entity recognition has a rich history in research, with the majority of models regarding it as a sequence labeling task, which identifies named entities with specific meaning from text, such as people's names, locations, organizations, dates, times, and so on, to aid in comprehending the meaning and structure of the text. Traditional sequence labeling models are based on Hidden Markov Models (HMM) [15] or probabilistic graph models such as CRF [16].

The emergence of deep learning has sparked considerable interest among researchers in neural-based models. Recent research on NER has explored various models, including convolutional neural networks (CNN) [17–19], recurrent neural networks (RNN) [20], and transformers [21,22]. These models use contextual information and semantic representations to automatically learn features and rules, eliminating the need for manual design and overcoming its drawbacks. Additionally, the emergence of pre-trained language models, such as BERT [12] and GPT [23], has further improved the effectiveness of NER. These

models are pre-trained on large-scale unlabeled texts and fine-tuned on labeled data, which can significantly improve the accuracy and robustness of NER.

However, the existing NER methods for social media often struggle with informal language, spelling errors, and the use of non-standard abbreviations and expressions. These factors contribute to the scarcity and sparsity of labeled data, which renders it more difficult to train NER models that can achieve high accuracy on social media text. Yang et al. [22] implement bidirectional contextual information by pre-training a language model. It predicts the possible words at a location by randomizing the input sequence and then trains a word vector with context. Nie et al. [24] enhance the NER model by using syntactic information or semantically related text. Shahzad et al. [25] assist the contextual model training by introducing social media images. Hu et al. [26] obtain more local features and location information via multi-window loops and combining global information and multiple local features to predict entity labels. These existing research works have certainly helped to improve the performance of NER on social media data. However, they have their limitations. Pre-training the model on large-scale data can introduce noise to the already noisy social media data, making it difficult to accurately capture the features of entities. Similarly, using semantic or image information via direct introduction can also introduce noise and affect the ability of the model to identify entities accurately. Additionally, some of these methods may suffer from overfitting to specific entities or situations, reducing their generalizability to different domains and datasets. Therefore, our proposed method aims to address these limitations by utilizing data augmentation techniques that are specifically designed for social media data, allowing the model to capture the diverse and nuanced language used in social media while avoiding the introduction of noise.

2.2. Data Augmentation

Obtaining the high quality of text representation is crucial for good model performance for many NLP tasks. However, due to sparse data and category imbalance in social media NER, the generalization ability and robustness of these models are still challenging. To address these challenges, researchers have proposed a variety of approaches to data augmentation, for example, data augmentation using synonyms, antonyms, and word form variations; semantic transformation based on semantic role annotations; and the generation of new samples via language models.

Data augmentation is a method to increase data diversity, include synonym replacement, antonym replacement, random insertion, random deletion, etc. Zhang et al. [7] used synonym dictionaries for data enhancement and classification, but limitations in dictionary size and lexicality prevented the replacement of all words. Wei et al. [8] replaced names and places with aliases or related words to generate similar but not identical data, but their method performed word-by-word substitution independently, potentially damaging semantic fluency with too many substitutions.

Semantic transformation is the transformation of semantic information from the original training data to extend the training data. These methods include inserting new entities into the text data, inserting qualifiers of entity types around the entities, and transforming the entity types. Wang and Yang et al. [10] used word embedding to enhance text classification, replacing each original word with a similar word based on cosine similarity to ensure diverse augmented data. Liu [27] combined lexicon and word embedding methods to build a lexicon of relevant words for each word and continuously improved it during training, solving the issue of limited applicability to specific words. However, these methods may not effectively handle multiple words and could still harm semantic fluency.

Language model generation is the use of pre-trained language models (e.g., GPT, BERT, etc.) to generate new training samples. In NER tasks, this method can generate new data that are similar but not identical to the original training data. Jiao et al. [11] developed a method for data augmentation by using the word splitter of BERT to obtain multiple word pieces from each piece of raw data and then constructed a candidate set of replacement words for each word piece. The candidate set was constructed by using a combination of

the word vector-based method and the mask model-based method. The mask language model considers contextual information during prediction to solve the problem of multiple meanings of words and generates semantically fluent sentences. However, this method often requires heuristics to determine the position of the mask to ensure that the augmented sentence retains the original semantics. Overall, this method is effective in generating diverse and semantically fluent sentences but requires careful consideration of the mask position during augmentation.

The key difference between our approach and others is that we do not rely on direct data expansion or semantic transformation. Instead, we fuse the entities that have been enhanced with additional data after training, which helps to avoid any negative effects on the inter-contextuality of the data. This also ensures that the additional entity information takes into account the association between different contexts, which can lead to more accurate results in NER.

3. Method

The social media NER task is commonly approached as a sequence tagging task, where the input sequence $X = x_1, x_2, \cdots, x_n$, which consists of n tokens, is annotated with corresponding named entity (NE) labels $Y = y_1, y_2, \cdots, y_n$ of the same length. In line with this approach, we propose a neural model with data augmentation for social media NER. First, we pre-trained a BERT language model to extract semantic vectors of each word in the training data. Then, we used the GloVe model to obtain similar entities via data augmentation methods. The GloVe model helped us to obtain similar entities for each named entity in the training data, which were then subjected to replacement or semantic transformation. Finally, we fed the augmented data, along with the original data, into the Bi-LSTM-CRF model for training and used a self-attention layer to capture the underlying information of the features and reduce the dependence on external information. By fine-tuning the model parameters, we obtained the optimal labels for each named entity in the test data. The architecture of our model is depicted in Figure 1.

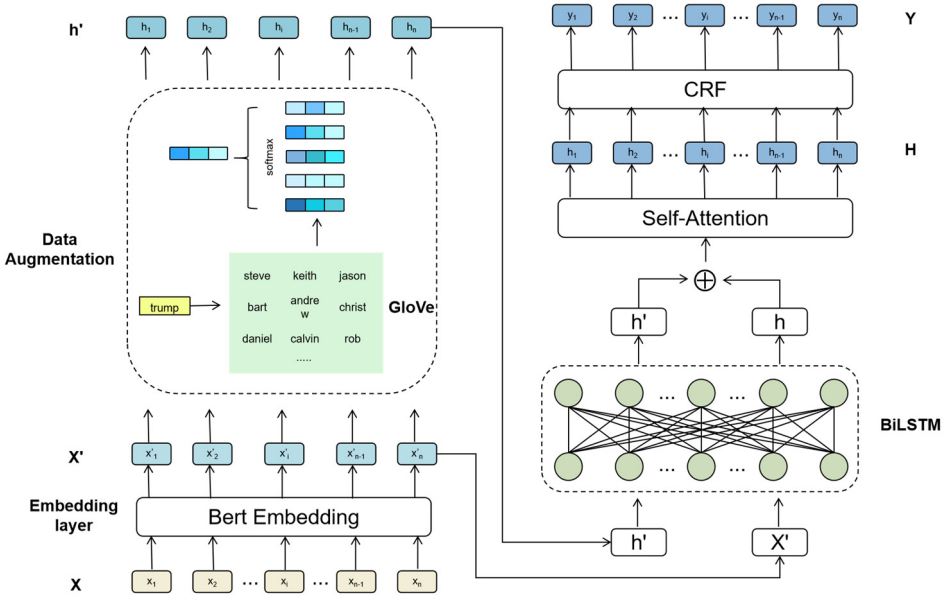

Figure 1. The architecture of DANER.

3.1. BERT

Word embedding is a popular technique in the field of NLP that maps words to a low-dimensional space, effectively addressing the issue of sparse text features and placing similar words closer together in a semantic space. Traditional methods of generating word vectors, such as word2vec and Elmo [28], as well as other pre-trained language models,

tend to be context-independent, rendering it difficult to accurately represent the various meanings of polysemous words. The BERT model uses a multi-layer bidirectional transformer as an encoder, for which each unit consists of a feed-forward neural network and a multi-head attention mechanism that enables each word representation to integrate information from both its left and right context. We use BERT to preprocess word embeddings to obtain the vector $X' = x'_1, x'_2, \cdots, x'_n$.

3.2. Data Augmentation

Due to the issues of data sparsity in social media, it is not easy to obtain high-quality text in the social media domain. Therefore, we propose social media NER data augmentation by using the most similar words in pre-trained embeddings to replace or mask the representation of each token in the input sentences.

We used GloVe [13] to preprocess word embeddings; for each token $x'_i \in X'$, the top n words that are most semantically similar to x'_i were extracted based on cosine similarity and are represented as:

$$Z_i = z_{i,1}, z_{i,2}, \cdots, z_{i,j}, \cdots, z_{i,m} \tag{1}$$

Then, we used a separate embedding matrix e_i to map all extracted words $z_{i,j} \in Z_i$ to their corresponding embeddings. To improve the accuracy of predicting NE labels for x_i in the given context, it is crucial to differentiate the contributions of different words since not all similar words are helpful for this task. To be specific, for every token x_i, attention is used to assign weights to each word $z_{i,j} \in Z_i$. This weight assignment process is carried out as follows:

$$\mu_{i,j} = \frac{exp^{(h_i \cdot e_{i,j})}}{\sum_{j=1}^{m} exp^{(h_i \cdot e_{i,j})}} \tag{2}$$

where h_i is the vector of x_i in the contextual encoder with the same dimension as the embedding dimension of Z_i. Then, we applied the weights to the word z_i, which was used to compute the final vector h'_i of Z_i:

$$h'_i = \sum_{j=1}^{m} \mu_{i,j} \cdot e_{i,j} \tag{3}$$

3.3. Bi-LSTM

Contextual information is crucial when processing sequential data to accurately understand and predict each token in the sequence. To extract this information, Bi-LSTM can be used as an encoder and decoder. By processing both the forward and backward information of the sequence, bidirectional LSTM can better capture the long-range dependencies between elements and extract context vectors. The vector representation of the hidden layer output of LSTM model is defined as follows:

$$f_t = \sigma\Big(W_{fx} \cdot x_t + W_{fh} \cdot h_{t-1} + b_f\Big) \tag{4}$$

$$i_t = \sigma(W_{ix} \cdot x_t + W_{ih} \cdot h_{t-1} + b_i) \tag{5}$$

$$\widetilde{c}_t = \tan \mathrm{h}(w_{cx} \cdot x_t + w_{ch} \cdot h_{t-1} + b_c) \tag{6}$$

$$c_t = f_t * c_{t-1} + i_t * \widetilde{c}_t \tag{7}$$

$$o_t = \sigma(W_{ox} \cdot x_t + W_{oh} \cdot h_{t-1} + b_o) \tag{8}$$

$$h_t = o_t * \tan \mathrm{h}(c_t) \tag{9}$$

where W_{fx}, W_{fh}, and b represent weight matrices and bias vectors connecting the input x_t and the previous output h_{t-1} to the LSTM unit. The sigma symbol σ represents the sigmoid activation function. The input gate, forgetting gate, and output gate are represented by i_t, f_t, and o_t. The point multiplication operation is denoted by $*$. The variable $\widetilde{c}_t$ represents the input modulation gate, and c_t and h_t represent the cell state and output of the LSTM unit at time t, respectively.

In the encoder, the input sequence is encoded into a context vector that is passed to the decoder for the next prediction step. In the decoder, this vector serves as contextual information that is combined with the current state of the decoder and previous output to generate the next output. Specifically, the hidden state of Bi-LSTM can be expressed as follows:

$$h_i = \overrightarrow{h}_i \bigoplus \overleftarrow{h}_i \tag{10}$$

where $\overrightarrow{h}_i$ and $\overleftarrow{h}_i$, respectively, represent the hidden states of the forward and backward LSTM at position i; $\bigoplus$ represents the concatenation operation.

3.4. Tagging

To extract the sequence context information of vectors h_i and h'_i, we can input them into Bi-LSTM separately and obtain their respective context vectors. These vectors are fused by the self-attention mechanism to obtain a new vector H representing:

$$\begin{aligned} H &= Attention\left(W^Q h_i, W^K h'_i\right) \\ &= softmax\left(W^Q h_i, \left(W^K h'_i\right)^T\right) \cdot \left(W^K h'_i\right) \end{aligned} \tag{11}$$

where W^Q and W^K are the query mapping matrix and the key mapping matrix, respectively; $(\cdot)^T$ denotes the matrix transpose operation; $H = [h_1, h_2, \cdots, h_n]$ and $E = [e_1, e_2, \cdots, e_n]$ are the list of hidden vectors and embeddings of input sequence X. Furthermore, pre-trained word embeddings capture a significant amount of external information from a large-scale corpus, and different types of word embeddings may contain different information. To make the most of this information, we propose a simple merging method that concatenates their embedding vectors:

$$e_i = e_i^1 \bigoplus e_i^2 \bigoplus \cdots \bigoplus e_i^T \tag{12}$$

where the final word embedding for x_t is given by e_i, and T is the set of all embedding types. After obtaining the final embeddings, a conditional random field (CRF) decoder is utilized to predict the labels $\widetilde{y}_i \in L$ in the output sequence Y:

$$\widetilde{y}_i = \underset{y_i \in L}{argmax} \frac{exp^{(W_H \cdot H_i + b_H)}}{\sum_{y_{i-1} y_i} exp^{(W_H \cdot H_i + b_H)}} \tag{13}$$

where W_H and b_H are the trainable parameters to model the transition from the previous y_{i-1} to the current label y_i.

4. Experimental and Analysis

4.1. Set

In our experiments, we used three social media benchmark datasets including WNUT16 (WN16) [29], WNUT17 (WN17) [30], and OntoNotes 5.0 (ON5e) [31]. The OntoNotes 5.0 is an English multitask corpus of over 90,000 documents and 160,000 sentences funded by the National Science Foundation. It includes text data of various types such as news reports, movie scripts, telephone conversations, lecture recordings, blog posts, etc. The WNUT 2016 is assembled on the basis of the Twitter NER task and contains both training and development sets. The WNUT 2017 is the second dataset and is derived mainly from Twitter. Also included are Reddit, YouTube, and StackExchange. All of these datasets are used for NER tasks and have been widely used in research in the fields of social media text

analysis and NLP. These datasets feature coverage of common abbreviations, misspellings, and non-standard spellings that render NER tasks more challenging. These datasets can be used to train and evaluate NER algorithms. For all datasets, we used their original segmentation, and Table 1 reports their statistics on entity types, sentences, and overall number of entities in the train/dev/test sets.

Table 1. The statistics of all datasets with respect to the number of sentences (Se.) and named entity (NE.).

Dataset	NE Types		Train	Dev	Test
WN16	10	Se.	2.4 K	1.0 K	3.9 K
		NE.	1.5 K	0.7 K	3.5 K
WN17	6	Se.	3.4 K	1.0 K	1.3 K
		NE.	2.0 K	0.8 K	1.1 K
ON5e	18	Se.	59.9 K	8.5 K	8.3 K
		NE.	81.8 K	11.1 K	11.3 K

For model implementation, we followed the BIOES tagging scheme used by [19]. For the text input, we used both ELMo and bert-case large types of embeddings by default. In the contextual encoding layer, we utilized the Bi-LSTM encoder and set the exit rate to 0.2. To extract entity-similar words, we leveraged pre-trained GloVe word embeddings to obtain the top 10 most similar words (and n = 10) for each word. For the data augmentation module, we randomly initialized the embeddings of the extracted words to represent the information carried by these words.

All experiments were conducted with a batch size of 64 and 100 epochs for each method, with hyperparameters tuned on the development set. We used precision (P), recall (R), and F1 score as evaluation metrics, which is consistent with previous works.

4.2. Baseline

To evaluate the effectiveness of our proposed approach in social media NER, we compared it with other state-of-the-art models. Below is a brief description of these models.

- **Bi-LSTM-CRF** [14] uses a bi-directional LSTM (Bi-LSTM) to model the language and obtain both left-to-right and right-to-left sentence encodings. This enables the model to capture the context of a word from both directions. Then, a Conditional Random Field (CRF) layer is added to capture the dependencies between the tags and make predictions based on the entire sentence.
- **BERT** [12]. BERT is a language model that is specifically designed for pre-training deep bidirectional representations from unlabeled text. It does so by conditioning on both left and right contexts in all layers, making it highly effective in capturing contextual information in natural language processing tasks.
- **XLNET** [22] is an autoregressive model that implements bidirectional contextual information by per-training language models. It predicts the possible words at a certain position by randomly arranging the input sequence and then trains a contextualized word vector with context.
- **AESINER** [24] improves entity recognition by leveraging syntactic information or semantically relevant texts.
- **InferNER** [25] is a method designed for Named Entity Recognition (NER) in short texts. It utilizes word-, character-, and sentence-level information without relying on external sources. Additionally, it can incorporate visual information and includes an attention component that computes attention weight probabilities over textual and text-relevant visual contexts separately.
- **HGN** [26]. HGN obtains more local features and location information through multi-window loops and combines global information and multiple local features to predict entity labels.

4.3. Results and Analyses

In this subsection, we present the results of our experiments evaluating the effectiveness of our proposed model and compare it with existing studies. The results are presented in Table 2.

Table 2. The performance of methods on datasets (%).

Methods	WN16			WN17			ON5e		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
Bi-LSTM-CRF	-	-	-	-	-	-	86.04	86.53	86.28
BERT	-	49.02	54.36	-	46.73	49.52	-	-	89.6
XLNET	55.94	57.46	56.69	58.68	49.18	53.51	89.72	91.05	90.38
AESINER	-	-	55.14	-	-	50.68	-	-	90.32
IfterNER	-	-	-	-	-	50.52	-	-	-
HGN	**59.74**	59.26	59.50	**62.49**	53.10	57.41	90.29	91.56	90.92
Ours	58.37	**65.23**	**61.61**	58.11	**68.89**	**63.04**	**90.42**	**91.65**	**91.03**

The results demonstrate that our proposed model outperforms the baseline in all experiments. On the WN16 and WN17 datasets, our model surpasses HGN by 2.11% and 5.63%, respectively. However, on the ON5e dataset, we achieve only a slight improvement of 0.11%. This could be attributed to the difficulty of entity labeling on this dataset, which has a large number of entity types. Our model significantly improves performance compared to BERT and Bi-LSTM-CRF on all datasets, indicating that combining data augmentation techniques can effectively enhance the performance of our model.

Although our proposed model achieves higher recall and F1 scores than HGN on the WN16 and WN17 datasets, there is still room for improvement in NER, as the accuracy of our model lags behind that of HGN. We suspect that this is because HGN uses a Hero and Gang module, which enhances the correlation between global and local feature information, resulting in better entity location information. Our model, on the other hand, focuses on incorporating richer word information and using Bi-LSTM to capture long-term dependencies, which does not provide an advantage in capturing global contextual information and local feature location information.

Our approach outperforms most baselines and achieves state-of-the-art performance on the social media NER benchmark dataset. This is mainly because our model expands the semantic information of entities through data augmentation methods, resulting in better performance due to the strong textual and richer word information. However, the use of pre-trained models and data augmentation in our model leads to high memory consumption of GPT, which affects the computation speed of the model. We plan to address this issue in future work.

4.4. Ablation Study

We designed ablation experiments to evaluate the data enhancement effectiveness of the DANER model. The ablation experiments were set up as follows.

- DANER. DANER is the proposed method.
- DANER without attention. The attention mechanism is removed, and the entities are labeled by CRF after direct vector fusion.
- DANER without weight word. In data augmentation, pre-trained vectors are not weighted, and data augmentation and semantic transformation are performed directly on the embedded vectors.
- DANER without data augmentation. The model with data augmentation removed is degraded to BERT + Bi-LSTM + CRF neural network model.

We used the WN16 dataset as the experimental dataset, and the other settings were calibrated in the same way as in the previous experiments. We performed each round of experimental training five times and report the average scores of F1 in Table 3.

Table 3. Ablation study results of the WN16 dataset.

Attention	Weight Word	DA	F1
√	√	√	**61.61**
×	√	√	60.73
√	×	√	58.36
×	×	√	56.83
√	×	×	55.62
×	×	×	54.96

As can be seen through the experiments, the performance decreases after removing the attention mechanism, which indicates that the use of attention can effectively explore the information location and labeling of entities. The absence of weighting the vectors enhances the probability of non-entities acquiring labels, which leads to a decrease in the experimental results. When the data enhancement module is removed from the model, the performance receives a greater impact, which illustrates the importance of integrating data enhancement into the model.

5. Discussion

Through the above comparison experiments and ablation experiments, we can observe that our model is able to perform well on social media datasets with sparse data and category imbalance. The experiments show that the data augmentation approach greatly improves the performance of our task and can effectively expand the data. The weighting approach is effective in avoiding the noise introduced in data augmentation, thus improving the robustness of the model. Our model adds attention to fine-tuning the weighted fused vectors to improve the F1 value. With the addition of attention, it can effectively capture the contextual information in the space to better identify the entities.

6. Conclusions and Future Work

In this paper, we propose a data augmentation-based approach to social media NER. To solve the problems of data sparsity and category imbalance in social media datasets, we extracted contextual information by pre-training BERT and expand the semantic information of entities via data augmentation. Considering the problems such as noise caused by directly adding semantic information, we trained both vectors simultaneously, fine-tuned the vector fusion via weighting, and captured the feature information by using attention to obtain the best label. Our experiments on three social media benchmark datasets showed that our model surpasses previous studies and achieves state-of-the-art results, demonstrating the effectiveness of our model for social media NER.

However, our model lags behind in accuracy when it comes to identifying entities, and we plan to address this by optimizing the shortcomings of the model related to the correlation of local and global features in future work. We will also apply our model to other data-sparse NER tasks, such as low-resource NER. We hope that our research results will facilitate the development of NER.

Author Contributions: W.L.: conceptualization, methodology, software, validation, formal analysis, investigation, data curation, writing—original draft, visualization, writing—review and editing. X.C.: resources, writing—review and editing, supervision, funding acquisition. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the Yunnan Province Science Foundation under Grant No. 202001BB050076 and in part by the Fund Project of Yunnan Province Education Department under Grant No. 2022j0008.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Humphreys, L. Populating Legal Ontologies Using Information Extraction Based on Semantic Role Labeling and Text Similarity. Ph.D. Thesis, University of Luxembourg, Luxembourg, 2016.
2. Babych, B.; Hartley, A. Improving machine translation quality with automatic named entity recognition. In Proceedings of the 7th International EAMT Workshop on MT and Other Language Technology Tools, Improving MT through Other Language Technology Tools, Resource and Tools for Building MT at EACL 2003, Budapest, Hungary, 13 April 2003; pp. 1–8.
3. Pang, L.; Lan, Y.; Guo, J.; Xu, J.; Su, L.; Cheng, X. Has-qa: Hierarchical answer spans model for open-domain question answering. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; pp. 6875–6882.
4. Ritter, A.; Clark, S.; Etzioni, O. Named entity recognition in tweets: An experimental study. In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, Edinburgh, UK, 27–31 July 2011; pp. 1524–1534.
5. Peng, N.; Dredze, M. Named entity recognition for chinese social media with jointly trained embeddings. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, 17–21 September 2015; pp. 548–554.
6. Aguilar, G.; Maharjan, S.; López-Monroy, A.P.; Solorio, T. A multi-task approach for named entity recognition in social media data. *arXiv* **2019**, arXiv:1906.04135.
7. Zhang, X.; Zhao, J.; LeCun, Y. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*; National Science Foundation: Alexandria, VA, USA, 2015; Volume 28.
8. Wei, J.; Zou, K. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv* **2019**, arXiv:1901.11196.
9. Coulombe, C. Text data augmentation made simple by leveraging nlp cloud apis. *arXiv* **2018**, arXiv:1812.04718.
10. Wang, W.Y.; Yang, D. That's so annoying!!!: A lexical and frame-semantic embedding based data augmentation approach to automatic categorization of annoying behaviors using# petpeeve tweets. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, 17–21 September 2015; pp. 2557–2563.
11. Jiao, X.; Yin, Y.; Shang, L.; Jiang, X.; Chen, X.; Li, L.; Wang, F.; Liu, Q. Tinybert: Distilling bert for natural language understanding. *arXiv* **2019**, arXiv:1909.10351.
12. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
13. Pennington, J.; Socher, R.; Manning, C.D. Glove: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 1532–1543.
14. Huang, Z.; Xu, W.; Yu, K. Bidirectional LSTM-CRF models for sequence tagging. *arXiv* **2015**, arXiv:1508.01991.
15. Zhou, G.; Su, J. Named entity recognition using an HMM-based chunk tagger. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, PA, USA, 7–12 July 2002; pp. 473–480.
16. Ratinov, L.; Roth, D. Design challenges and misconceptions in named entity recognition. In Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009), Boulder, Colorado, 4–5 June 2009; pp. 147–155.
17. Collobert, R.; Weston, J.; Bottou, L.; Karlen, M.; Kavukcuoglu, K.; Kuksa, P. Natural language processing (almost) from scratch. *J. Mach. Learn. Res.* **2011**, *12*, 2493–2537.
18. Ma, X.; Hovy, E. End-to-end sequence labeling via bi-directional lstm-cnns-crf. *arXiv* **2016**, arXiv:1603.01354.
19. Lample, G.; Ballesteros, M.; Subramanian, S.; Kawakami, K.; Dyer, C. Neural architectures for named entity recognition. *arXiv* **2016**, arXiv:1603.01360.
20. Yang, Z.; Yang, D.; Dyer, C.; He, X.; Smola, A.; Hovy, E. Hierarchical attention networks for document classification. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego, CA, USA, 12–17 June 2016; pp. 1480–1489.
21. Akbik, A.; Blythe, D.; Vollgraf, R. Contextual string embeddings for sequence labeling. In Proceedings of the 27th International Conference on Computational Linguistics, Santa Fe, NM, USA, 20–26 August 2018; pp. 1638–1649.
22. Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R.R.; Le, Q.V. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems*; National Science Foundation: Alexandria, VA, USA, 2019; Volume 32.
23. Ethayarajh, K. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. *arXiv* **2019**, arXiv:1909.00512.
24. Nie, Y.; Tian, Y.; Song, Y.; Ao, X.; Wan, X. Improving named entity recognition with attentive ensemble of syntactic information. *arXiv* **2020**, arXiv:2010.15466.
25. Shahzad, M.; Amin, A.; Esteves, D.; Ngomo, A.-C.N. InferNER: An attentive model leveraging the sentence-level information for Named Entity Recognition in Microblogs. In Proceedings of the The International FLAIRS Conference Proceedings, North Miami Beach, FL, USA, 17–19 May 2021; Volume 34.
26. Hu, J.; Shen, Y.; Liu, Y.; Wan, X.; Chang, T.-H. Hero-Gang Neural Model For Named Entity Recognition. *arXiv* **2022**, arXiv:2205.07177.
27. Liu, S.; Lee, K.; Lee, I. Document-level multi-topic sentiment classification of email data with bilstm and data augmentation. *Knowl.-Based Syst.* **2020**, *197*, 105918. [CrossRef]
28. Peters, M.E.; Neumann, M.; Zettlemoyer, L.; Yih, W.-T. Dissecting contextual word embeddings: Architecture and representation. *arXiv* **2018**, arXiv:1808.08949.

29. Strauss, B.; Toma, B.; Ritter, A.; De Marneffe, M.-C.; Xu, W. Results of the wnut16 named entity recognition shared task. In Proceedings of the 2nd Workshop on Noisy User-Generated Text (WNUT), Osaka, Japan, 11 December 2016; pp. 138–144.
30. Derczynski, L.; Nichols, E.; Van Erp, M.; Limsopatham, N. Results of the WNUT2017 shared task on novel and emerging entity recognition. In Proceedings of the 3rd Workshop on Noisy User-Generated Text, Copenhagen, Denmark, 7 September 2017; pp. 140–147.
31. Pradhan, S.; Moschitti, A.; Xue, N.; Ng, H.T.; Björkelund, A.; Uryupina, O.; Zhang, Y.; Zhong, Z. Towards robust linguistic analysis using ontonotes. In Proceedings of the Seventeenth Conference on Computational Natural Language Learning, Sofia, Bulgaria, 8–9 August 2013; pp. 143–152.