Intelligent Data Analysis 28 (2024) 841–865
DOI 10.3233/IDA-230588
IOS Press

SocialNER2.0: A comprehensive dataset for enhancing named entity recognition in short human-produced text

Adel Belbekri[a,*], Fouzia Benchikha[a], Yahya Slimani[b] and Naila Marir[a]
[a]*Lire Laboratory, University of Constantine 2 – Abdelhamid Mehri, Algeria*
[b]*Joint Group for Artificial Reasoning and Information Retrieval (JARIR), Manouba University, Tunisia*

Abstract. Named Entity Recognition (NER) is an essential task in Natural Language Processing (NLP), and deep learning-based models have shown outstanding performance. However, the effectiveness of deep learning models in NER relies heavily on the quality and quantity of labeled training datasets available. A novel and comprehensive training dataset called SocialNER2.0 is proposed to address this challenge. Based on selected datasets dedicated to different tasks related to NER, the SocialNER2.0 construction process involves data selection, extraction, enrichment, conversion, and balancing steps. The pre-trained BERT (Bidirectional Encoder Representations from Transformers) model is fine-tuned using the proposed dataset. Experimental results highlight the superior performance of the fine-tuned BERT in accurately identifying named entities, demonstrating the SocialNER2.0 dataset's capacity to provide valuable training data for performing NER in human-produced texts.

Keywords: Big data, deep learning, user-generated texts, text analysis, named entity recognition

1. Introduction

The internet has become ubiquitous because of the continuous advancement of technology, such as communication networks, connected devices, and the democratization of social network usage, community sites, and blogs. These evolutions have generated a large amount of data, presenting an opportunity to extract new sources of knowledge instead of from traditional sources. These information sources are extremely valuable to businesses and governments. They represent commercial or sociocultural analysis more because they are closer to the consumer/citizen. Furthermore, these data sources provide an important foundation for developing intelligent tools capable of direct human interaction, such as personal assistants like Google Assistant, ChatGPT, and Siri. These latter have assimilated into our daily lives, assisting us in various tasks, ranging from answering questions to providing personalized recommendations.

Using textual data from social media platforms has been a major focus for researchers in various fields. This massive amount of data has facilitated studies ranging from opinion exploration to linguistic variation analysis. However, it is important to note that the nature of messages posted on these platforms frequently differs from that of more traditional data sources used in NLP models. Messages on social media platforms are typically less formal in tone than formal texts. They have distinct linguistic characteristics

*Corresponding author: Adel Belbekri, Lire Laboratory, University of Constantine 2 – Abdelhamid Mehri, Algeria. E-mail: adel.belbekri@univ-constantine2.dz.

such as shortened words, colloquial expressions, emoticons, irregular grammar, and spelling. Therefore, traditional NLP tools face a new challenge in processing and analyzing the informal language found in user-generated data [1]. To address this issue, the NLP community has created linguistic resources and NLP pipelines specifically for social media data [2,72,27,32,74,29,73].

Several NLP applications, such as knowledge extraction, information retrieval, indexing and automatic translation, spelling or grammar checking, text classification, summary production, and entity linking, aim to develop methods and tools to address these challenges. The task of NER plays a transversal role in these various fields [3]. NER emerged as a subtask of information retrieval in the 1990s [4]. Its primary goal is to detect and classify textual objects in various contexts, such as the names of organizations, persons, and places. However, NER remains a difficult learning problem due to the scarcity of training data [5]. Thus, the challenge is to generalize small datasets to obtain large ones. Traditional approaches rely on orthographic features and language-specific knowledge resources, but developing them for new languages and domains can be costly [6]. This includes new types of knowledge sources like user-generated data from social network resources [7].

Deep learning approaches have been widely used to solve NER problems [8,71,67,69,66,70,68,3]. The availability of large labeled datasets is clearly a factor in deep learning's optimal performance [9]. However, available NER datasets for human-produced texts are limited in data quality or quantity [10]. In fact, the number of datasets that meet the following criteria is limited: (i) datasets that cater to short textual content; (ii) datasets that are accurately annotated with named entities and their corresponding types; (iii) datasets that are substantial in quantity to satisfy the requirements of large-scale models. In a previous work [11], a dataset called SocialNER was proposed by incorporating data from well-known existing datasets [12,45,40,43,44], with the objective of satisfying quality and quantity requirements. This dataset, however, was limited to only four entity categories (person, location, organization, and miscellaneous) and had imbalanced data.

This paper presents a new version of the SocialNER dataset. The number of entity types was increased from four to twelve, adding new categories: products and brands, languages, sports, events, creativity, nature, food, and dates. This expansion broadens the retrievable types covered by social platform texts and user-generated content. Data imbalance is also addressed, focusing on the types' distribution and the nature of the content. To facilitate model learning, an oversampling technique is used with Word2Vec models [13] trained on Glove [14] and Google News corpora [15]. In addition, data from traditional datasets was used to handle well-written text. Next, extensive testing is carried out on SocialNER2.0 and other datasets to demonstrate the effectiveness of the proposed dataset when used with pre-trained BERT models [16]. The aim is to outperform existing approaches in terms of F1 score and to evaluate the performance of existing NER models on the dataset. The main contributions of this work can be summarized as follows:

- Different datasets from different NLP tasks (including question answering, information retrieval and entity linking) are combined. The principle of data fusion is applied to combine these datasets efficiently and extract useful information. Furthermore, new datasets that were not used in previous work are introduced. To enhance the annotation process, word information extracted automatically is incorporated, including Part-Of-Speech tags (POS-tags), chunk tags, and new types retrieved from the DBpedia knowledge graph. Additionally, standard datasets are integrated to extract named entities from well-written user-generated texts. Furthermore, to address the issue of imbalanced data and entity types, supplementary data is generated.
- To validate the efficiency of the dataset for training and benchmarking purposes, a series of experiments are conducted. The experiments aim to answer the following research questions:

(RQ1) Is the SocialNER2.0 dataset suitable for deep learning in named entity detection of human-generated texts, and can it be used to enhance the performance of existing NER models?
(RQ2) Do models trained on SocialNER2.0 perform consistently when applied to user-generated and more formal texts such as journals or books? Furthermore, how does the performance of SocialNER2.0-trained models differ across different types of named entities?
(RQ3) Does the incorporation of heterogeneous datasets of various types and natures in SocialNER2.0 improve the performance of trained models?

The remainder of the paper is structured as follows: Section 2 provides an overview of current research. Section 3 describes the process of creating the SocialNER2.0 dataset. Section 4 presents and discusses the experimental findings. Finally, Section 5 summarizes the paper's findings and highlights some open issues that require further investigation.

2. Related work

Several annotated datasets, including CoNLL-2003 [17] for news articles, OntoNotes [18], and more recently Few-NERD [19] for formal documents, have been created specifically to facilitate NER tasks. These datasets, however, are not always suitable for the NER task in short human-produced texts, such as social media posts, which can be challenging. These short texts are often characterized by a lack of standardization in language use, a high level of noise, and a wide range of colloquial expressions and abbreviations. To address this challenge, some studies [20,27,21,32,37,25,29] have proposed datasets for NER in short human-produced texts, with two main approaches: manual data collection and annotation, and data fusion and augmentation.

2.1. Manual data collection and annotation

Many approaches involve human annotation, which experts or crowd workers can do. Authors in [21] conducted a study introducing an annotated Twitter corpus, a dataset specifically designed for NER in tweets. The corpus comprises four standard categories: person, location, organization, and others. To collect the annotations, the researchers used Amazon Mechanical Turk (MTurk) [22] and CrowdFlower [23]. The aim is to address the issues raised by the informality and brevity of Twitter compared to traditional genres, this dataset is used in [24,64] for testing and evaluations. In [25], the authors present a dataset containing tweets with annotated entities named in ten categories. To perform the annotation, experts in NLP manually labeled the dataset. This research investigates the difficulties conventional NLP tools encounter when processing tweets due to their informal and abbreviated style. Specifically, the study addresses these challenges in POS tagging, chunking, and NER, [26,65] used this dataset for benchmarking purposes.

The Broad Twitter Corpus (BTC), an extensive dataset encompassing named entity annotations, is presented in [27]. The data collection process is a collaborative effort between NLP experts and crowd workers, resulting in comprehensive data collection. The BTC includes the original source text and intermediate annotations and covers a wide range of geographical regions, temporal periods, and types of Twitter users. The named entity annotations in BTC use standard categories: persons, organizations, locations, and miscellaneous entities, [28,63] used the BTC to perform evaluations. Authors in [29] present a Hindi-English CodeMixed corpus for NER and extensive experiments on machine learning models. The proposed corpus comprises tweets from the Indian subcontinent over the last eight years about politics, social events, sports, and other topics. The Twitter Python API3 [30] was used for tweet

collection. Extensive pre-processing was used to remove useless and noisy tweets. This dataset was annotated by two human annotators with linguistic backgrounds and proficiency in Hindi and English. The named entity categories covered by the corpus correspond to the standard NER categories, which include persons, organizations, locations, and miscellaneous entities, this dataset is cited in [31,62,62] as a benchmark dataset.

In a recent study, authors in [32] used Tweebank V2 [33] to create the Tweebank-NER corpus. The researchers utilized Amazon Mechanical Turk to annotate named entities within the Tweebank V2 dataset and then evaluated the quality of the annotations. The annotation process was carried out in accordance with the guidelines outlined in the CoNLL 2003 standard for NER [17]. To help annotators understand the guidelines, multiple examples were provided for each rule. The resulting dataset and models have been made public to facilitate future research in Tweet NLP [34]. The Tweebank-NER corpus incorporates named entity annotations that adhere to standard categories, encompassing persons, organizations, locations, and miscellaneous entities.

The manual annotation process has introduced subjectivity into the above datasets, resulting in inconsistencies and imprecision. These constraints can have an effect on the quality of labeled data, affecting the performance of models trained on it and their ability to identify and classify entities in text. Furthermore, relying on a single data source limits the trained model's ability to generalize and adapt to various conditions and contexts. Single-source datasets may have inherent biases, a limited perspective, a focus on specific topics or domains, and noise or inaccuracies.

2.2. Data fusion and augmentation

The datasets in this section result from data fusion, data augmentation, or combining the two. The PLONER dataset is presented in [35]. This dataset consists of texts from W-NUT16 [36], CoNLL-2003 [17], and 0ntoNotes [18]. The researchers selected samples covering only three entity types: persons, organizations, and locations. This dataset was utilized for benchmarking purposes, as described in their paper.

In [20], the authors expanded the UlyssesNER-Br corpus for the NER task by adding comments pertaining to bills in the Brazilian Portuguese language. They also supplemented the annotated corpus with a formal corpus to see if combining formal and informal texts from the same domain could improve NER. The researchers ran experiments with a Bidirectional Long Short-Term Memory-Conditional Random Fields (BiLSTM-CRF) model and a Conditional Random Fields (CRF) model. Then, they used the proposed dataset to fine-tune the BERT model for the NER task. The authors concluded that formal texts help with entity identification in informal texts, with the optimized BERT model performing the best.

In [37], the authors conducted a study on generating weakly labeled data for NER in social media. They employed a data generation approach that utilized Freebase triplets and sentence matching [38]. They obtained the weakly labeled data by annotating named entities in unlabeled texts when they appeared together in sentences. The authors proposed two augmentation techniques to strengthen the weakly labeled data: alias augmentation and typo augmentation. Alias augmentation involved gathering alternative names from Wikidata to create new named entity pairs. In contrast, typo augmentation introduced different types of typos to increase the likelihood of matching named entity candidate pairs in social media texts. By implementing these methods, they successfully increased the volume of weakly labeled data and enhanced the performance of NER, specifically in the realm of social media. Table 1 lists key features for comparing and positioning SocialNER2.0 in relation to existing works.

Table 1
Comparison of existing NER datasets

Dataset	Information sources	# of entity types	# of tokens (K)	Availability	Annotation type	Language	Nature of text
Finin et al. [21]	Tweets	04	007	Unavailable	Manual crowdsourcing annotation	English	Informal
Ritter et al. [25]	Tweets	10	046	Unavailable	Manual expert annotation	English	Informal
The Broad Twitter Corpus [27]	Tweets	04	165	Available	Manual expert annotation + crowdsourcing annotation	English	Informal
Hindi-English CodeMixed corpus [29]	Tweets	04	005	Available	Manual expert annotation	Codemixed hindi-english	Informal
The PLONER dataset [35]	Tweets, Reddit, Forums, Books and Journals	03	050	Available	Not provided	English	Informal + Formal
Tweebank-NER [32]	Tweets	04	024	Available	Manual crowdsourcing annotation	English	Informal
UlyssesNER-Br corpus [20]	Comments about bills	07	138	Available	Manual crowdsourcing annotation	Brazilian Portugese	Informal + Formal
Kim et al. [37]	Twitter + Wikipedia	06	019 Tweet + 120 Wiki	Unavailable	Automatic	English	Informal + Formal
The proposed SocialNER2.0 dataset	Tweets, Reddit, Forums, Dbpedia, Books, Journals, Generated data with Word2Vec trained on Glove and Google News	12	696	Available	Automatic	English	Informal + Formal

Despite these efforts, there is still a lack of specialized training datasets for NER in short, human-produced text. As previously stated, the manual annotation process can lead to imprecise labeling of datasets. The relatively small number of tokens and entities in these manually annotated datasets also represents a second weakness. Furthermore, these datasets have limitations regarding entity types, particularly those found in user-generated texts. The small number of tokens and entities in these datasets also makes generalization difficult for models that require larger training sets. Furthermore, models' ability to handle the diverse range of entities commonly found in user-generated texts is hampered by insufficient representation of various entity types.

This paper proposes an enhanced dataset with a wide range of entity types to address these limitations. An automated process was used to generate the training dataset, which uses data sources from other NLP challenges, open-linked data knowledge graphs, and synthetically generated data based on the integrated datasets. SocialNER2.0 approaches can be classified as fusion and augmentation, with a combination of both in terms of knowledge sources and the nature of the text used. While there are currently no standard training datasets for NER in short human-produced text, the proposed dataset aims to advance the state of the art in this field. The SocialNER2.0 dataset generation process is detailed in the section that follows.

3. SocialNER2.0 dataset construction

A well-curated dataset serves as the foundation for successful deep-learning models. This section describes the method used to create the SocialNER2.0 dataset, which includes several key steps to ensure its quality and effectiveness for the NER task. To begin with, data is selected by mining pre-existing datasets to retrieve relevant data. This step ensures that the dataset contains a wide variety of user-generated texts. Next, relevant data are extracted from various fields. Data enrichment is also carried out, which involves supplementing the extracted data with additional information such as named entity type, chunk position, and POS tags. This enrichment process improves the dataset's granularity and depth, allowing for more accurate NER analysis.

Data conversion prepares the dataset for training and evaluation [39]. This includes aggregating and structuring merged datasets in CoNLL-2003 format [17]. In addition, new synthetic data is generated based on existing data to solve any data imbalance problems that may arise. Furthermore, data from conventional datasets are integrated to efficiently process user-generated correctly written texts, maximizing dataset coverage and relevance. Once the dataset is constructed, it is validated. This involves fine-tuning the pre-trained BERT model on the proposed dataset SocialNER2.0. The fine-tuning process enhances the model's performance and adaptability to the dataset. Finally, the dataset is validated by evaluating the fine-tuned pre-trained BERT model performance on existing datasets, ensuring its compatibility and benchmarking its effectiveness in NER tasks. Overall, the dataset construction process involves data selection, extraction, enrichment, conversion, balancing, and validation steps, all geared toward creating a comprehensive and high-quality dataset for NER analysis on user-generated texts. Figure 1 depicts the process for generating the SocialNER2.0 dataset.

3.1. Data selection

A training dataset is constructed by extracting information of interest for the NER task from several established datasets. The datasets chosen are specifically for evaluating NLP challenge tasks in which NER plays a significant role. As a critical sub-task, NER significantly impacts the outcomes of these tasks. The subsequent datasets are used: LC-QuAD, QALD-5, QALD-6, QALD-7, QALD-8, QALD-9,

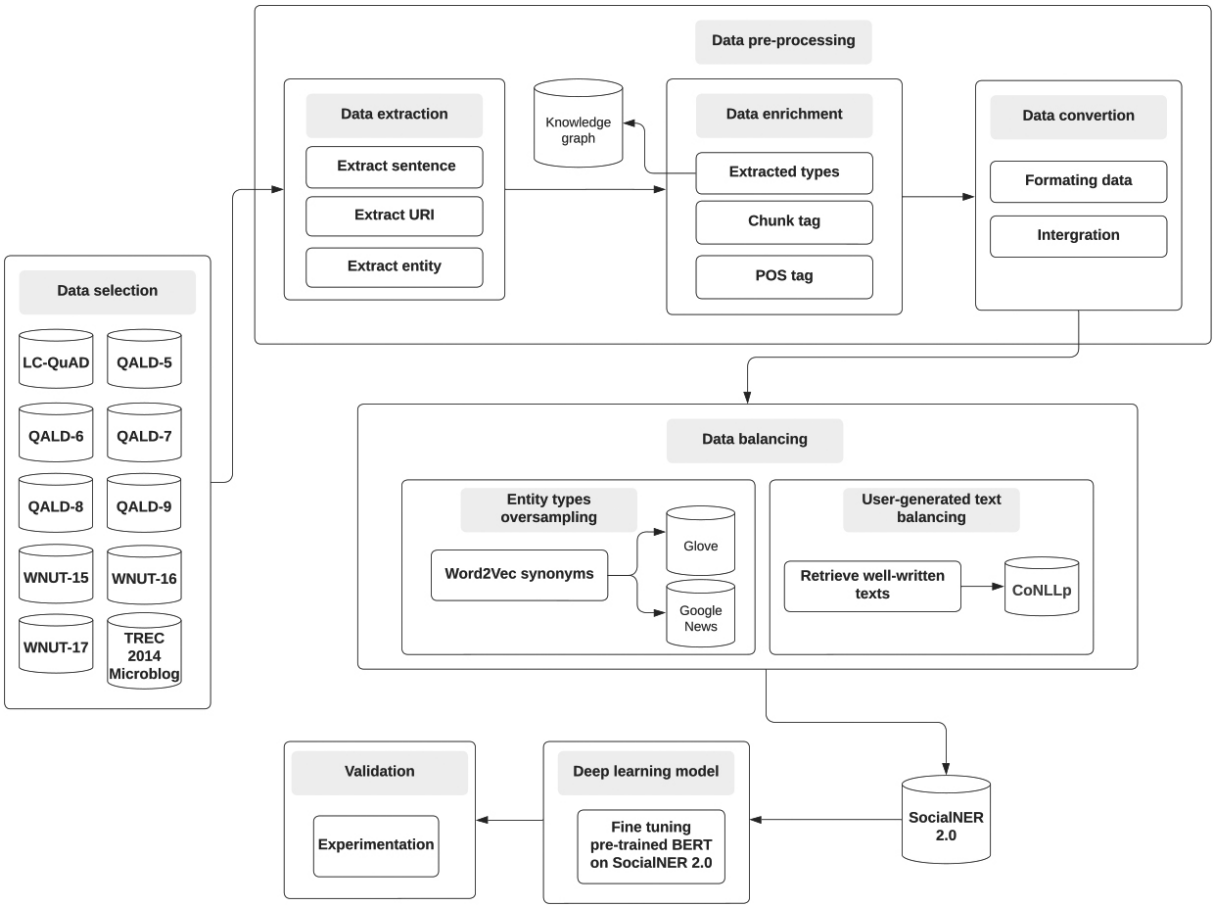

Fig. 1. SocialNER2.0 generation process.

W-NUT15, W-NUT16, W-NUT17, and TREC 2014 microblog. For instance, the QALD datasets contain natural language questions such as *"What is the currency of china?"* as well as corresponding SPARQL queries that can extract information about named entities like country names. While these datasets contain brief and informal textual data labeled for evaluating NLP tasks like question-answering, they provide relevant data that can be used for NER. Table 2 summarizes the datasets fields used for the fusion process.

3.1.1. Data sources

To ensure the correctness of the data in our SocialNER2.0 dataset, we adopted a methodology that leverages the reliability of several widely used and previously validated datasets, including QALD, WNUT, LC-QuAD, and CoNLL. These datasets are recognized as gold standards in their respective domains and have been extensively used in prior NLP research, providing a solid foundation for our work.

- **LC-QUAD** is a benchmark dataset for Question Answering over Linked Data. It consists of 5,000 questions, with 3,000 questions for training and 2,000 for testing. The questions are written in natural language and cover various topics, such as movies, music, and geography [40].
- **QALD-5, QALD-6, QALD-7, QALD-8 and QALD-9.** The "QALD-X", which stands for Question Answering over Linked Data, is a series of challenges designed to advance the field of QALD systems.

Table 2
Extracted fields from the datasets used for the fusion process

Dataset	Format	Fields used	Observation
LC-QUAD	JSON	Question label URI of the entity	Entity type is retrieved using a SPARQL query on the entity URI
QALD-5 QALD-6 QALD-7 QALD-8 QALD-9	JSON	Question SPARQL query	A Python code is used to retrieve the resource URI from the SPARQL query (field on the original dataset). Then, the same SPARQL code in LC-QUAD is used to retrieve the category
TREC 2014 Microblog	JSON	Question Query annotation	The category is retrieved using a SPARQL query on the entity URI
W-NUT15 train W-NUT16 train W-NUT17 train	CoNLL	Token Ner_tag	/

QALD-5, QALD-6, QALD-7, QALD-8, and QALD-9 represent different editions of this challenge. Each focuses on evaluating and promoting the development of question-answering techniques in specific contexts. QALD-5 [41], held in 2015, emphasized multilingual question answering and sought to enhance systems' ability to handle questions in different languages. QALD-6 [42] included hybrid questions, which required combining information from structured data and textual resources. In 2017, QALD-7 [43] focused on answering questions using structured and unstructured data sources while incorporating a user interaction component. QALD-8 [44] introduced a new track for machine reading-based question answering to assess systems comprehension and reasoning abilities. Finally, QALD-9 [45] focused on enhancing systems performance by exploring various data augmentation techniques.

- **TREC 2014 Microblog** was created as part of the TREC Microblog track, which aims to advance research in the information retrieval domain specifically focused on microblog data. This dataset has been a valuable resource for researchers and developers, allowing them to train and evaluate their systems for information retrieval tasks using microblog data [12].
- **W-NUT15, W-NUT16 and W-NUT17** are widely recognized as resources used in the field of NLP and knowledge extraction. They were created for the Workshop on Noisy User-generated Text (WNUT), which focuses on analyzing and understanding text data from social media platforms. The W-NUT15 dataset [46] comprises tweets annotated with NER tags, allowing researchers to train and evaluate NER models specifically designed for noisy informal texts. Similarly, the W-NUT16 dataset [36] extends the Workshop on Noisy User-generated Text 2015 tasks by providing data with additional new data as NER classes, including time expressions, ordinal numbers, and more. Finally, the W-NUT17 dataset [47] expands on previous versions by introducing a new challenge of identifying and categorizing emergent entities in social media texts. These datasets serve as valuable benchmarks for developing robust NLP models that can handle the unique characteristics of user-generated content, thereby contributing to advances in information extraction and social media data understanding.

3.2. Data extraction

3.2.1. Entity extraction from SPARQL queries

To process QALD datasets efficiently, a regular expression-based extraction technique is used to retrieve the URI linked to the named entity from the SPARQL query present in the dataset. This is essential,

```
import re
def extract_entities(query):
    pattern="http://dbpedia.org/resource/[^>]+"
    return re.findall(pattern,query)
def entities(query):
    firstModified=[]
    if query=="OUT OF SCOPE":
        return firstModified
    whereString = query[query.index('{')+1:query.rfind('}')-1]
    if "no_query" in whereString:
        return firstModified
    whereString=whereString.replace("\n","")
    whereString=whereString.replace("\t"," ")
    query=whereString
    pattern="res:[^\s]+"
    first=re.findall(pattern,query)
    for entity in first:
        firstModified.append(entity.replace("res:","http://dbpedia.org/resource/"))
    pattern="http://dbpedia.org/resource/[^>]+"
    second=re.findall(pattern,query)
    return firstModified+second
```

Listing 1. The source code of resource extraction process.

```
PREFIX rdfs : <http://www.w3.org/2000/01/rdf-schema#>

SELECT * WHERE
{
    <http://dbpedia.org/resource/China> rdfs:label ?y FILTER (lang(?y) = "en")
}
```

Listing 2. SPARQL query to retrieve the label of the entity.

as QALD datasets are annotated with SPARQL queries that correspond directly to the questions asked. This technique can ensure accurate and efficient retrieval of relevant information associated with named entities. For example, The source code published by [48] (Listing 1) is used to extract the URI of the named entity from the SPARQL query. This technique allows accurate and efficient retrieval of relevant information associated with the named entity.

3.2.2. Entity label extraction

As part of the methodology for processing and analyzing relevant data, a SPARQL query was implemented to retrieve the entity label, specifically designed to extract the necessary information (Listing 2) [11]. This label detects the entity to be tagged with a type; other entities are tagged with the 'O' label.

3.3. Data enrichment

3.3.1. Entity type extraction

The named entity type is retrieved using the corresponding URI (Listing 3). For instance, the URI of the entity *"China"* is used to query DBpedia and retrieve the type of this entity. The data obtained from the knowledge graph (DBpedia) is then incorporated into the existing dataset to enhance its quality. To improve the accuracy of the deep learning model, the data types are categorized into superclass types, also referred to as coarse-grained types [49] (Fig. 2), using the dictionary illustrated with few examples in Fig. 3. This categorization ensures that the model is trained on a more generalized representation of the data, thereby enhancing its overall performance.

```
PREFIX rdf : <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs : <http://www.w3.org/2000/01/rdf-schema#>
PREFIX dbo :  <http://dbpedia.org/ontology/>

SELECT * WHERE
{
    <http://dbpedia.org/resource/China> rdf:type ?y.
    ?y rdfs:label ?x FILTER (lang(?x) = "en")
 }
```

Listing 3. Example of the entity type extraction using the URI of the resource “China”.

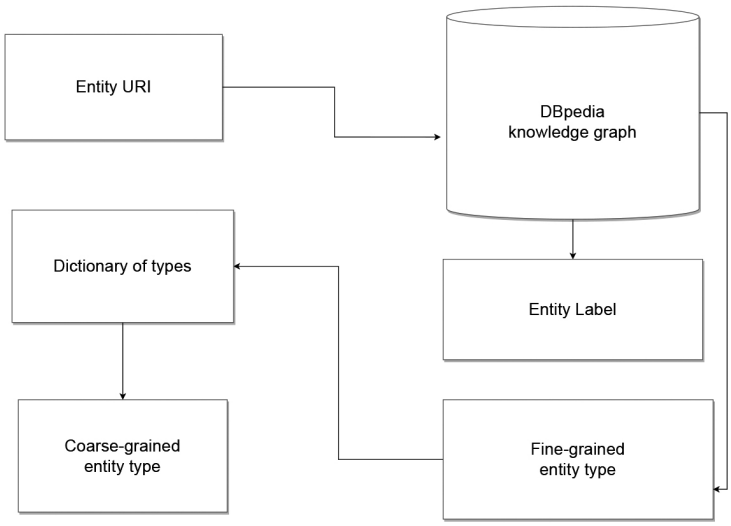

Fig. 2. Entity label and type extraction.

3.3.2. Chunk-tag and pos-tag

To enhance the relevance of the dataset, information about pos-tags [50] and chunk-tags [51] is incorporated. These tags provide additional insights into the grammatical category of words and sentence structure. Including this information in the dataset improves its ability to understand and interpret text more accurately and contextually.

3.4. Data conversion

Data conversion is the process of converting data from various datasets into a suitable format for processing. It entails changing the representation, structure, or format of data so that NER deep learning models can use it effectively. Data conversion ensures data is structured and prepared to enable models to interpret and use information during learning and prediction processes.

3.4.1. Formatting data

Formatting data is of the utmost importance in data processing and analysis, particularly in the field of NLP. CoNLL [17] and JSON (JavaScript Object Notation) [52] are two commonly used formats for representing structured data in NLP. Both formats have distinct advantages and are widely supported by various NLP tools and libraries. The CoNLL format is commonly used for displaying tabular data with linguistic annotations. It has a standardized structure with columns for various linguistic features like

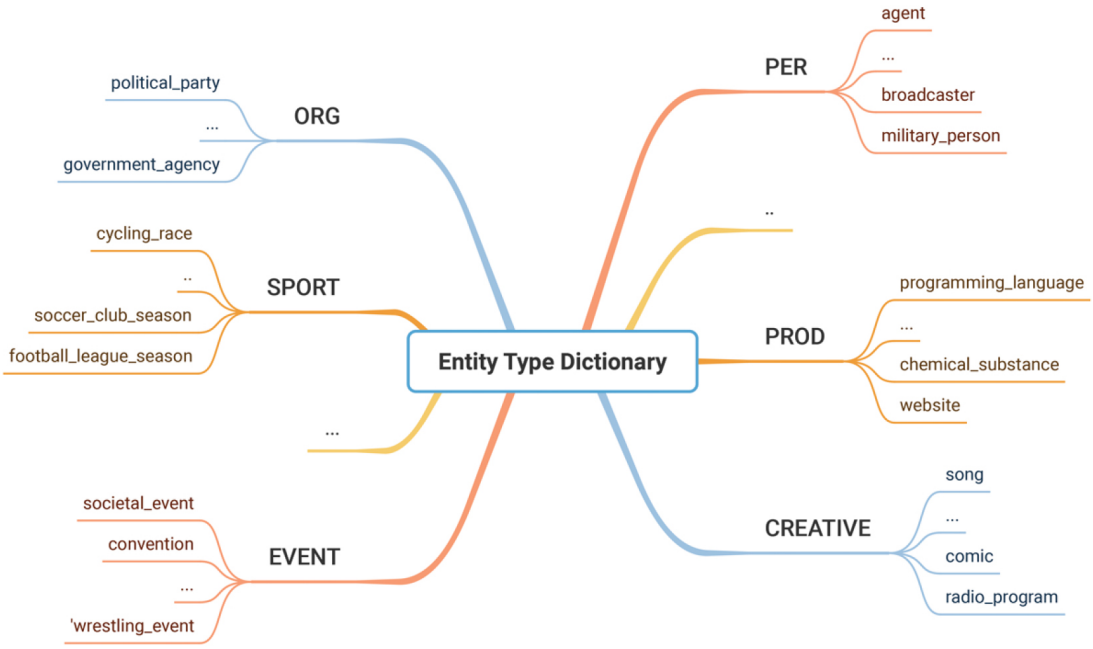

Fig. 3. Correspondence between categories.

word tokens, POS tags, syntactic dependencies, and named entities. The CoNLL format allows for simple data manipulation and is popular in tasks such as NER and syntactic parsing. The data is formatted for neural network processing in accordance with the CoNLL-2003 data format specifications:

- Each line represents a single word.
- Empty lines separate sentences.
- The final element of each line is a tag, indicating whether the word is part of a named entity and its type.
- Four fields appear in each line: the word, its POS tag, its chunk tag, and its named entity tag.
- Words outside named entities are tagged as "O", while those inside named entities are tagged with "I-XXX", where "XXX" denotes the entity type.
- To indicate the beginning of a new entity, the tag "B-XXX" is assigned to the first word of the second entity when two entities of type "XXX" appear consecutively.
- The dataset comprises twenty named entity types: persons (PER), organizations (ORG), locations (LOC), miscellaneous names (MISC), products and brands (PROD), languages (LANGUAGE), sport (SPORT), events, creative works, natural, food, and date. A dictionary (Fig. 3) was used to classify the named entities into the named category to establish a correspondence between the categories of named entities in the knowledge graph and the aforementioned categories.

This tagging schema uses the IOB scheme originally proposed by [51]. The CoNLL-2003 assumes that named entities are non-overlapping and non-recursive; only the top-level entity is annotated if a named entity is embedded in another named entity.

The proposed dataset is available in JSON format, a lightweight data-interchange format renowned for its simplicity and human-readable nature. Numerous programming languages widely support it, and NLP applications commonly employ it for data storage, transmission, and interchange [53]. This format makes it easy to use in different contexts, offering convenience, flexibility, and compatibility. The SocialNER2.0

JSON file has been formatted as a collection of JSON objects, each containing the following key-value associations:

- "Data": This key represents the user-generated data, typically a sentence or a question.
- "Tokens": This key expresses the tokens denoting the data's individual words or units.
- "POS": This key characterizes the part-of-speech tags, which indicate the grammatical category of each token.
- "Chunk": This key symbolizes the chunking tags, which group words together based on their syntactic structure.
- "NER": This key represents the NER tags, which determine if a token corresponds to a named entity, such as a person, location, or organization.

The NER values employed in SocialNER2.0 adhere to the same format described in the previous paragraph regarding the CoNLL format.

3.4.2. Data integration

The process of combining and integrating multiple data sources or datasets to create a unified dataset that can be used for training is referred to as data integration. The aim is to use diverse and complementary information from various sources to improve the performance and generalization capability of the trained model. Models can learn from a more comprehensive and representative set of examples when multiple datasets are combined, capturing a broader range of patterns and relationships in the data. As a result, after formatting our dataset according to the required structure, it is consolidated into a single file for testing and analysis. Furthermore, to enhance the reliability of our dataset, we conducted a series of consistency checks. These checks included random manual inspections to verify that no significant deviations occurred during the automated extraction process.

After analyzing the obtained data, the following observations were made: the distribution of the remaining named entity classes shows a significant class imbalance, 'B-PER', 'I-PER', 'B-LOC', 'I-ORG', and 'B-ORG' labels having the highest counts; however, there are very few examples of some other named entity labels, such as 'B-DATE' (only 10 instances). Furthermore, the overall volume of data is relatively small to meet the requirements of large-scale models. Another observation is that the constructed dataset does not handle correctly written user-generated data regarding spelling and grammar.

3.5. Data balancing

To address the issue of imbalanced data and improve the performance of deep learning models trained on our dataset, we propose to generate synthetic data using Word2Vec [13] to balance out the distribution of named entity labels. Word2Vec is a neural network-based model for language processing that can learn to represent words in a high-dimensional vector space based on their context. These vector representations, or "word embedding," capture the semantic relationships between words and can be used to generate synonyms for a given word [54].

To generate synthetic data using Word2Vec, the named entities in the original dataset are identified and the phrases containing them are extracted. Word2Vec is then employed to generate synonymous phrases for each extracted phrase while ensuring the retention of the original named entity tags. For example, suppose the following phrase with a named entity tag: "John Smith works at Google" is retrieved. In that case, the phrase is extracted, and Word2Vec is employed to generate synonymous phrases, such as "John Smith works for Google", "John Smith is employed at Google", and so on. Retaining the original named entity tag ensures the generated phrases are still labeled with the correct entity type (Algorithm 1). Table 3

Algorithm 1: Data Preprocessing and Synthesis for Named Entity Recognition

Require: Path to the dataset file (`file_path`)
Load pretrained word embeddings models.
Initialize lists for sentences, tags, B-tags, chunks, and POS tags.
Open the dataset file at `file_path`.
Initialize variables for sentence, sentence_tags, chunk_tags, POS tags, B-tag, number_of_NE, and is_NE.
for each line in the dataset **do**
    Strip the line.
    if the line is empty **then**
        if `is_NE` is 1 **then**
            Reset `is_NE` to 0.
            Append B-tag to B-tags if it exists.
            Append sentence, sentence_tags, chunk_tags, and POS tags to their lists.
        end if
    else
        Split line into token, POS, IOB, and NER.
        if NER is "O" and token is not in stopwords and not punctuation **then**
            Initialize `max_similarity` to 0.
            if token exists in model's vocabulary **then**
                Find synonyms with similarity to token.
                Update `max_similarity`, `new_token`, and `is_NE` if necessary.
            end if
            if `max_similarity` $\geqslant 0.7$ **then**
                Set `new_token` to synonym with `max_similarity`.
            else
                Set `new_token` to token.
            end if
        else
            Set `new_token` to token.
        end if
        Append `new_token`, NER, POS, IOB to their lists.
        if NER starts with "B-" **then**
            Set B-tag to NER.
        end if
        if NER is "B-XXX" **then**
            Set `is_NE` to 1 and increment `number_of_NE`.
        end if
    end if
end for
Open a new file named "NE_oversampled" for writing processed data.
for each sentence in sentences **do**
    for each token in the sentence **do**
        Write token, POS, IOB, and NER as tab-separated values to the file.
    end for
    Write a newline to separate sentences.
end for
Ensure: Processed data stored in the "NE_oversampled" file, containing tokens, POS tags, IOB tags, and NER labels.

displays the final balanced distribution of the diverse entity types after employing word embedding-based oversampling to supplement the under-represented classes.

To address the issue of proper spelling and grammar and to enhance the quality of our dataset as well as the system's capability to handle diverse text types, including formal writing, we integrated the final dataset with the traditional CoNLL dataset, known for its high-quality, formal text samples, after generating synthetic data. During the development process, we employed a data augmentation and fusion process and found that our focus on addressing various challenges inadvertently weakened our ability to

Table 3
Distribution of oversampled entity types using Word2Vec

NER types	After integration original % of distribution	Word2Vec oversampled % of distribution
ORG	20.35	12.00
PER	33.54	9.55
LOC	19.69	10.23
CREATIVE	12.21	9.31
PROD	4.79	8.38
MISC	4.24	8.35
EVENT	2.63	7.54
NATURAL	1.02	7.22
LANGUAGE	0.50	7.80
SPORT	0.60	6.90
FOOD	0.39	6.32
DATE	0.06	6.04

Table 4
Characteristics of SocialNer2.0 dataset

Characteristics		Values
# of sentences		5.08260×10^4
# of tokens		6.96568×10^5
# of named entities		5.97640×10^4
# NER tags	"O" label	5.94987×10^5
	"B-XXX" label	5.97640×10^4
	"I-XXX" label	4.18170×10^4

recognize more formal texts. Despite their informal nature, social media and user-generated content can also contain correctly written material. Therefore, we decided to leverage the linguistic rigor inherent in the CoNLL dataset to counteract this weakness. By merging it with our final dataset, we further enhanced the diversity of the content. This integration significantly increased the robustness and effectiveness of the resulting models, ensuring they are well-equipped to handle a wide spectrum of text types, from informal to formal.

3.6. Generated dataset properties

Before the balancing phase, it was observed that the majority of the data belonged to the "O" label, with a count of 111,570. This observation can be attributed to the prevalence of non-named entity words in natural language text. The total number of tokens in the dataset, excluding the "O" label, is 19827, representing 9667 named entities. Available data was increased by generating synthetic data using Word2Vec and adding classical data from the CONLL-2003 dataset. This approach aimed to enhance the performance of deep learning models trained on this dataset, ensuring their capacity to accurately identify and classify named entities in text. Table 4 summarizes the characteristics of the final generated dataset.

4. Experimentation

To evaluate the effectiveness of the SocialNER2.0 dataset in improving the performance of existing NER models, a series of experiments were conducted using a pre-trained BERT model [16]. The choice of the BERT model was motivated by the following factors:

– Performance: BERT has achieved remarkable results across various NLP tasks, showcasing its ability

to understand and process language deeply contextually. This contextual understanding is crucial in NER, as it allows the model to capture the intricate relationships between words and accurately identify named entities within the given text.
- Pre-training on a variety of text sources: it enables BERT to learn highly informative contextual representations of words, allowing it to comprehend the meaning and context of the text. Contextual understanding is especially important in NER, where named entities frequently rely on their surrounding words for identification and classification.
- Capturing semantic relationships: one of BERT's key strengths is its ability to capture rich semantic relationships between words. This feature is especially useful in NER, where the presence of contextually distant words can significantly influence named entity identification and classification.
- Transfer learning and fine-tuning: leveraging a pre-trained BERT model offers the advantage of transfer learning. The pre-training phase allows BERT to acquire a general language understanding that can be fine-tuned using our proposed dataset for specific post-processing tasks such as NER. This fine-tuning process helps adapt the model's representations and parameters to the nuances and characteristics of the NER task and the target dataset.

The aforementioned factors make BERT an excellent choice for NER tasks, where accurate identification and classification of named entities rely heavily on contextual understanding.

4.1. Training and benchmarking data

4.1.1. Training data

Two datasets were used to fine-tune the BERT pre-trained model: the CoNLL benchmark dataset [17] and the proposed dataset SocialNER2.0.

- The CoNLL-2003 dataset is a widely used standard benchmark dataset for evaluating NER models. It consists of news articles from the Reuters Corpus and contains named entity labels for four types of entities: persons, locations, organizations, and miscellaneous [17].
- The SocialNER2.0 dataset is designed for NER in short human-produced texts from social media platforms. It has over 50826 annotated tweets, comments, and search queries in English, totaling 696568 tokens. The dataset includes a wide range of named entities, often absent in traditional NER datasets.

4.1.2. Benchmarking dataset

Benchmarking datasets is essential for evaluating the performance of NER models and comparing their effectiveness [55]. This study uses three benchmarking datasets to assess the performance of the fine-tuned BERT model for NER: Ner_references, W-NUT16 Test, and W-NUT17 Test.

- Ner_references [56] provides a comprehensive set of annotated texts that can be used as a reliable benchmark for NER tasks. This benchmarking dataset, available on Kaggle,[1] contains various examples of named entities from different domains, allowing us to assess the model's ability to recognize entities in various contexts, particularly user-generated data.
- The W-NUT16 test benchmarking dataset [36] is another valuable benchmark, focusing specifically on NER in user-generated texts from social media. Given the informal nature of the text, this benchmarking dataset presents a unique challenge, making it an ideal choice for evaluating the model's performance in handling user-generated content.

[1] https://www.kaggle.com/.

– Similarly, the W-NUT17 test benchmarking dataset [47] concentrates on NER in social media text and presents a more recent collection of annotated data, especially the emerging ones.

To generate the SocialNER2.0 dataset, we used the training splits of W-NUT16 and W-NUT17, excluding validation and test splits. This enhanced SocialNER2.0 breadth while preserving unseen test data for unbiased model evaluation. The fine-tuned BERT models were tested on dedicated W-NUT16 and W-NUT17 test splits, ensuring a clear separation. No overlap occurred between SocialNER2.0 training data and W-NUT test data, preventing data leakage. This approach aligns with best practices, allowing unbiased assessment of the models' generalization to new data.

Using these benchmarking datasets, the precision, recall, and F1-score of the proposed fine-tuned BERT model can be assessed in various domains and text genres. Moreover, this thorough evaluation allows us to assess the model's performance and compare it to existing state-of-the-art approaches, validating its effectiveness in accurately and reliably identifying named entities.

4.2. Experimental setup

The experiments described in this study were carried out on a machine equipped with an Intel Core i7 12700F 12th generation processor running at 2.10 GHz. The machine has 64 GB of RAM and an NVIDIA GeForce RTX 3090 Ti GPU with GA 102 Revision A1 and 24576 MB of memory. The Ubuntu 20.04 operating system served as the foundation, and the neural network models were implemented using the Python 3.9 programming language. The machine was outfitted with the NVIDIA CUDA Toolkit version 11.7 to utilize the GPU's power for accelerated training. The GPU improved the training process by optimizing the computational efficiency of the neural network models. A batch size of 32 examples was used during the training process. The ADAM optimizer [57], a popular deep learning model optimization tool, was used to iteratively update the network weights, optimizing model performance and convergence. The experiments aimed to achieve good performance and promising results for the neural network models used in this study by leveraging this hardware setup and advanced software tools.

4.3. Model training

The BERT model is used for NER from textual data. The training data is read from a text file containing sentences annotated with tags for each word. The BERT tokenizer converts sentences into token sequences, which are then converted into IDs and padded to have equal-length sequences. The tags are also converted to IDs, and attention masks are created to indicate which tokens are actual words and which are padding tokens. After splitting the data into training and testing sets, a pre-trained BERT model with a token classification head for named entity annotation is loaded. The model is then trained on the training set using the ADAM optimizer and log-likelihood loss function [57]. Finally, the model is evaluated on the test set, with precision, recall, and F1-score calculated.

4.4. Evaluation metrics

Precision, recall, and F1-score metrics are widely used and considered appropriate for evaluating the performance of NER models for several reasons. Precision is an important metric in NER because it measures the model's prediction accuracy. It represents the proportion of true positives (named entities that were correctly predicted) among predicted positives. A high precision indicates that the model has a low false positive rate, indicating that it is effectively identifying and classifying named entities. The model's recall measures its ability to find all relevant-named entities in a given text. It denotes the

Table 5
Evaluation metrics on the 20% test set

Metric	Value
True Positive Rate (TPR)	0.980
True Negative Rate (TNR)	0.990
Positive Predictive Value (PPV)	0.990
Negative Predictive Value (NPV)	0.940
False Negative Rate (FNR)	0.001
False Positive Rate (FPR)	0.001
False Discovery Rate (FDR)	0.001
False Omission Rate (FOR)	0.050
F1-score	0.950
Critical Success Index (CSI)	0.980
Accuracy (ACC)	0.980
Balanced Accuracy (BA)	0.980
Matthews Correlation Coefficient (MCC)	0.970
Bookmaker Informedness (BM)	0.980
Markedness (MK)	0.947

proportion of true positives among all positives. A high recall indicates that the model correctly identifies most named entities in the text, reducing the number of false negatives (missed named entities).

The F1-score is a balanced metric that takes precision and recall into account. The harmonic mean of precision and recall yields a single value representing the model's overall performance. The F1-score is especially useful when the number of positive and negative instances is unequal. In NER, where named entities are frequently a minor portion of the text, the F1-score provides a reliable measure of the model's performance by considering precision and recall equally. We gain a comprehensive understanding of the NER models' performance in terms of accuracy, capturing relevant-named entities and achieving a balance between precision and recall by evaluating them using precision, recall, and F1-score. These metrics enable meaningful comparisons between models, facilitate benchmarking against existing approaches, and aid in making informed decisions about the models' effectiveness for NER tasks.

4.5. Experimental results

4.5.1. Training results on SocialNER2.0

A pre-trained BERT model was fine-tuned on the SocialNER2.0 dataset for NER. The model was trained for 48 epochs with a batch size of 32, optimizing with ADAM optimizer and cross-entropy loss. The evaluation of the model's performance on the 20% test set yielded promising results (Table 5). The predicted class distribution showed a robust performance across all named entity types, achieving high precision and recall for each category. The loss function during training decreased steadily (Fig. 4), indicating effective learning throughout the epochs. Overall, the fine-tuned BERT model demonstrated high performance in NER on the SocialNER2.0 dataset, making it promising for identifying named entities in social media text. However, further analysis is required to assess the model's performance on diverse datasets and its generalization capabilities to unseen data. The results of experiments on different datasets are presented in the following subsection, providing a more comprehensive evaluation of the model's effectiveness.

4.5.2. Results and comparison on different datasets

The performance of five refined BERT models was assessed on four different datasets. The models were: fine-tuned on SocialNER2.0 (FBSN), fine-tuned on CoNLLp (FBC), fine-tuned on PLONER (FBP),

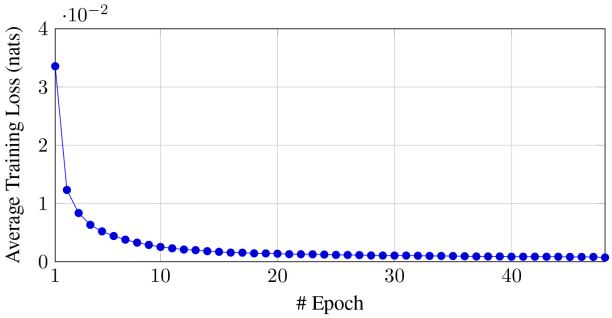

Fig. 4. Average training loss per epoch.

fine-tuned on the Broad Twitter Corpus (FBBTC), and fine-tuned on Tweebankner (FBTB). Balanced accuracy and macro averages (precision, recall, and F1-score) were used to evaluate the performance of the models on each dataset for NER. The results listed in Table 6 (with the best values indicated in bold) show that the FBSN model outperforms the others regarding macro average (precision, recall, and F1-score) on most datasets. It performs well on the Ner_references, W-NUT16, and W-NUT17 datasets, indicating its effectiveness in recognizing named entities in user-generated data. The other models, including FBC, FBP, FBBTC, and FBTB have lower macro averages across all datasets. This suggests limitations compared to FBSN in handling the nuances and characteristics of user-generated content. The evaluation results on the test dataset exhibit a notable variance in performance between the model trained on the SocialNER2.0 dataset and those trained on CoNLLp, Tweebankner, PLONER, and the Broad Twitter Corpus. This divergence suggests that models trained on CoNLLp, Tweebankner, PLONER, and the Broad Twitter Corpus may face challenges in effectively generalizing to the diverse and intricate spectrum of user-generated data. In contrast, the BERT model fine-tuned on SocialNER2.0 showcases superior performance on the test datasets, highlighting its proficiency in named entity recognition in user-generated texts. Moreover, while models trained on CoNLLp, Tweebankner, PLONER, and the Broad Twitter Corpus underperform on the SocialNER2.0 dataset, the model trained on SocialNER2.0 demonstrates improved effectiveness on the CoNLLp dataset. These findings emphasize the significant performance enhancements on test datasets when the BERT model is fine-tuned on the SocialNER2.0 dataset, as opposed to being fine-tuned on CoNLLp, Tweebankner, PLONER, and the Broad Twitter Corpus. These findings highlight the importance of fusion for training data to achieve optimal results in tasks such as NER in user-generated text. This underscores the value of diverse training sets in enhancing model adaptability and accuracy, particularly in social media content's dynamic and varied landscape, where traditional models might not suffice.

We also used the W-NUT17 dataset to validate the results against state-of-the-art works focusing on deep learning approaches for NER in text from social media. The macro average F1-score improved by 10.83%, increasing from 51.43% to 57.00%. The results are reported from [58,37]. The first three papers [59,60,61] employed traditional deep learning architectures such as BiLSTM, CNN, CRF, and topic modeling, combined with distributed word embeddings. In contrast, [37] leveraged BERT representations, upon which our approach also builds. Their model fine-tuned BERT using additional weakly labeled social media data for training. Similarly, our methodology performs BERT fine-tuning but instead utilizes the SocialNER2.0 dataset, which amalgamates and strengthens multiple source datasets through augmentation techniques. Table 7 summarizes the compared works.

In another experiment, the trained models' performance and distribution across various named entity types were evaluated using the W-NUT17 dataset. The radar diagram in Fig. 5 depicts the obtained

Table 6
Results of the fine-tuned BERT models for different benchmarks

Datasets	MAP[6]					MAR[7]					MAF1[8]					BA[9]				
	FBSN[1]	FBC[2]	FBP[3]	FBBTC[4]	FBTB[5]	FBSN	FBC	FBP	FBBTC	FBTB	FBSN	FBC	FBP	FBBTC	FBTB	FBSN	FBC	FBP	FBBTC	FBTB
Ner_references	**0.55**	0.45	0.35	0.26	0.24	**0.46**	0.42	0.33	0.27	0.16	**0.49**	0.42	0.32	0.26	0.16	**0.80**	0.79	0.65	0.71	0.55
W-NUT16	**0.92**	0.16	0.14	0.13	0.07	**0.86**	0.38	0.16	0.22	0.07	**0.85**	0.19	0.15	0.15	0.07	**0.89**	0.77	0.64	0.64	0.51
W-NUT17	**0.57**	0.16	0.13	0.13	0.08	**0.63**	0.24	0.11	0.12	0.08	**0.57**	0.19	0.12	0.12	0.08	**0.83**	0.79	0.65	0.63	0.53
CoNLLp	**0.97**	/	0.45	0.35	0.26	**0.98**	/	0.47	0.28	0.13	**0.97**	/	0.44	0.30	0.14	**0.98**	/	0.75	0.71	0.53
SocialNER2.0	/	**0.35**	0.22	0.23	0.07	/	**0.37**	0.18	0.14	0.07	/	**0.32**	0.19	0.16	0.07	/	**0.86**	0.68	0.62	0.49

[1]FBSN: Fine-tuned BERT model on SocialNER2.0. [2]FBC: Fine-tuned BERT model on CoNLL-2003. [3]FBP: Fine-tuned BERT model PLONER. [4]FBBTC: Fine-tuned BERT model on Board Twitter Corpus. [5]FBTB: Fine-tuned BERT model on TweeBankNER. [6]MAP: Macro Average Precision. [7]MAR: Macro Average Recall. [8]MAF1: Macro Average F1-Score. [9]BA: Balanced Accuracy.

Table 7
Comparison of the state-of-art deep learning NER approaches in human-produced texts, and BERT fine-tuned on SocialNER2.0

Works	% of F1-score on W-NUT17
BiLSTM-CRF model with multichannel word and character embeddings [59]	40.42
LDA topic modeling and distributed word representations [60]	41.81
CNN character embeddings and BiLSTM token embeddings [61]	41.86
BERT-SocialNER [11]	44.00
BERT base model fine-tuned on augmented training data [37]	51.43
Pre-Trained BERT fine-tuned on SocialNER2.0	**57.00**

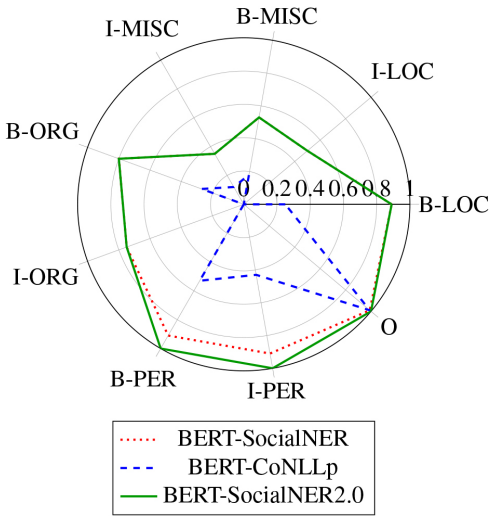

Fig. 5. Performance of different BERT fine-tuned models across entity types.

results. Categories not covered by CoNLLp were excluded to ensure an unbiased comparison. Notably, the BERT-SocialNER and BERT-SocialNER2.0 models' performance is nearly identical, except for the "person" category, where BERT-SocialNER2.0 outperforms. The comparable performance on these shared types reveals the saturation of model effectiveness for these categories. As a result, compared to the other two models, BERT-SocialNER2.0 has the highest coverage, as shown in Table 7. The additional entity categories introduced in SocialNER2.0 that were not originally present help explain this superior overall F1 score achieved by models trained on SocialNER2.0 versus SocialNER. The higher metric indicates improved generalization and adaptability to a broader range of named entities. This finding lends credence to the hypothesis that data fusion improves model performance. It was also emphasized that the MISC category inherently challenges NER models, given its unconstrained heterogeneous nature encompassing unnamed entity types unsuited for predefined categories. Furthermore, it was noted that the substantial variability in possible MISC mentions contributes to higher performance fluctuation.

4.6. Discussion

This study investigated the efficacy of the SocialNER2.0 dataset for deep learning in named entity detection, particularly in the context of human-produced texts. In addition, the dataset's potential to improve the performance of existing named entity detection models, its consistency when applied to different types of text, and the variation in performance between different types of named entities

were assessed. The study also focused on how incorporating data from standard datasets affected the performance of models trained on SocialNER2.0 for NER in user-generated text. Finally, the question of whether SocialNER2.0 can be used as both a benchmark and a training dataset was addressed.

The experiments indicate that the SocialNER2.0 dataset is effective for deep learning in named entity detection for human-produced texts. The ability of models trained on SocialNER2.0 to accurately identify and classify named entities is demonstrated by improved precision, recall, and F1-score across various named entity types. This suggests that the dataset can be used to train models that perform NER in human-generated text. In addition, the results strongly suggest that the SocialNER2.0 dataset can improve the performance of existing named entity detection models. Performance improvements were observed when fine-tuning the BERT model on the SocialNER2.0 dataset, particularly in user-generated texts. The 10.83% increase in performance highlights the dataset's potential to improve model capabilities in dealing with the nuances and complexities of user-generated content.

Regarding the consistency of SocialNER2.0-trained models when applied to different types of texts, the findings show that performance remains relatively stable across user-generated texts and more formal texts such as journals or books. This implies that the models trained on SocialNER2.0 have robustness and generalization capabilities, allowing them to recognize named entities in various text genres. Furthermore, the analysis revealed differences in the performance of SocialNER2.0-trained models across different types of named entities. This indicates the need for additional research and improvement in accurately identifying specific named entity categories. Future research should concentrate on improving the models' performance on various entity types.

Incorporating data from standard datasets improved the performance of SocialNER2.0-trained models for NER in user-generated texts. We observed precision, recall, and F1-score improvements across multiple datasets by combining the SocialNER2.0 dataset with the CoNLLp dataset. The use of standard datasets enabled the detection of named entities in well-written user-generated texts. This clearly demonstrates how using diverse training datasets from various sources can improve the robustness and accuracy of named entity detection models.

Considering the question of whether SocialNER2.0 can be used as both a benchmark dataset and a training dataset, the findings indicate that it can. The dataset is a valuable resource for deep learning model training and a reliable benchmark for evaluating the performance of new models or architectural designs. Because of its extensive annotation and coverage of various named entity types, it is an asset in the field of NER. Moreover, the results indicate that the SocialNER2.0 dataset is effective for deep learning in named entity detection for human-produced texts. The dataset shows promise for improving the performance of existing models, demonstrates consistency across different text genres, and reveals performance variations across different named entity types.

5. Conclusion

This paper addressed the importance of labeled training data for achieving effective performance in NER using deep learning-based approaches. To address this, we propose SocialNER2.0, a new promising dataset specifically designed for deep learning-based NER in user-generated text. Our proposal combines data fusion and enrichment principles by utilizing semantic information from the DBpedia knowledge graph and incorporating POS tags, chunk tags, and synthetic data generated by Word2Vec with Glove and Google News embeddings. Extensive experimentation validated the hypothesis that a diverse and large training dataset leads to more accurate models. Using our proposed and the CoNLLp datasets to fine-tune

the BERT model yields highly promising results, with an average performance improvement of 22.48% observed in user-generated texts.

Furthermore, by a factor of 10.83%, the proposed model outperforms state-of-the-art deep learning-based NER models for social media texts. The CoNLLp fine-tuned model's performance evaluation also establishes its potential as a benchmark for evaluating future models and architectural designs. Future research will focus on enhancing the model's performance using transfer learning and domain adaptation techniques. Consequently, we intend to broaden our research to include NER in different languages, thereby broadening the applicability and generalizability of our proposal.

References

[1] D. Khurana, A. Koli, K. Khatter and S. Singh, Natural language processing: State of the art, current trends and challenges, *Multimedia Tools and Applications* **82**(3) (2023), 3713–3744. doi: 10.1007/s11042-022-13428-4.

[2] R.K. Ando, T. Zhang and P. Bartlett, A framework for learning predictive structures from multiple tasks and unlabeled data, *Journal of Machine Learning Research* **6**(11) (2005).

[3] R. Sharma, S. Morwal and B. Agarwal, Named entity recognition using neural language model and CRF for Hindi language, *Computer Speech & Language* **74** (2022), 101356. doi: 10.1016/j.csl.2022.101356.

[4] M.E. Okurowski, Information Extraction Overview, in:*TIPSTER TEXT PROGRAM: PHASE I: Proceedings of a Workshop held at Fredricksburg, VA, USA, September 19–23, 1993*, Morgan Kaufmann, 1993, pp. 117–121. doi: 10.3115/1119149.1119164.

[5] P. Sun, X. Yang, X. Zhao and Z. Wang, An Overview of Named Entity Recognition, in: *2018 International Conference on Asian Language Processing, IALP 2018, Bandung, Indonesia, November 15–17, 2018*, IEEE, 2018, pp. 273–278. doi: 10.1109/IALP.2018.8629225.

[6] K. Adnan and R. Akbar, Limitations of information extraction methods and techniques for heterogeneous unstructured big data, *International Journal of Engineering Business Management* **11** (2019), 1847979019890771. doi: 10.1177/1847979019890771.

[7] M. Hatmi, C. Jacquin, E. Morin and S. Meignier, Named Entity Recognition in Speech Transcripts following an Extended Taxonomy, in: *Proceedings of the First Workshop on Speech, Language and Audio in Multimedia, Marseille, France, August 22–23, 2013*, CEUR Workshop Proceedings, Vol. 1012, CEUR-WS.org, 2013, pp. 61–65.

[8] M. Bhattacharya, S. Bhat, S. Tripathy, A. Bansal and M. Choudhary, Improving biomedical named entity recognition through transfer learning and asymmetric tri-training, *Procedia Computer Science* **218** (2023), 2723–2733. doi: 10.1016/j.procs.2023.01.244.

[9] C. Sun, A. Shrivastava, S. Singh and A. Gupta, Revisiting Unreasonable Effectiveness of Data in Deep Learning Era, in: *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22–29, 2017*, IEEE Computer Society, 2017, pp. 843–852. doi: 10.1109/ICCV.2017.97.

[10] B. Jehangir, S. Radhakrishnan and R. Agarwal, A survey on Named Entity Recognition-datasets, tools, and methodologies, *Natural Language Processing Journal* **3** (2023), 100017. doi: 10.1016/j.nlp.2023.100017.

[11] A. Belbekri and F. Benchikha, SocialNER: A Training Dataset for Named Entity Recognition in Short Social Media Texts, in: *Artificial Intelligence Doctoral Symposium. AID 2022, Alger, Algiers, September 18–19, 2022 Communications in Computer and Information Science, vol 1852*, Vol. 1852, Springer Nature Singapore, 2022, pp. 278–289. doi: 10.1007/978-981-99-4484-2_21.

[12] J. Lin, Y. Wang, M. Efron and G. Sherman, Overview of the TREC-2014 Microblog Track, in: *Proceedings of The Twenty-Third Text REtrieval Conference, TREC 2014, Gaithersburg, Maryland, USA, November 19–21, 2014*, NIST Special Publication, Vol. 500-308, National Institute of Standards and Technology (NIST), 2014, pp. 1–8.

[13] K.W. Church, Word2Vec, *Natural Language Engineering* **23**(1) (2017), 155–162. doi: 10.1017/S1351324916000334.

[14] J. Pennington, R. Socher and C.D. Manning, Glove: Global Vectors for Word Representation, in: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25–29, 2014, Doha, Qatar, A Meeting of SIGDAT, a Special Interest Group of the ACL*, ACL, 2014, pp. 1532–1543. doi: 10.3115/v1/d14-1162.

[15] I. Google, Google News corpora, 2013, Accessed: [2023].

[16] J. Devlin, M. Chang, K. Lee and K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2–7, 2019, Volume 1 (Long and Short Papers)*, Association for Computational Linguistics, 2019, pp. 4171–4186. doi: 10.18653/v1/n19-1423.

[17] E.F.T.K. Sang and F.D. Meulder, Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition, in: *Proceedings of the Seventh Conference on Natural Language Learning, CoNLL 2003, Held in Cooperation with HLT-NAACL 2003, Edmonton, Canada, May 31–June 1, 2003*, ACL, 2003, pp. 142–147.
[18] S. Pradhan, A. Moschitti, N. Xue, O. Uryupina and Y. Zhang, CoNLL-2012 Shared Task: Modeling Multilingual Unrestricted Coreference in OntoNotes, in: *Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning – Proceedings of the Shared Task: Modeling Multilingual Unrestricted Coreference in OntoNotes, EMNLP-CoNLL 2012, July 13, 2012, Jeju Island, Korea*, ACL, 2012, pp. 1–40.
[19] N. Ding, G. Xu, Y. Chen, X. Wang, X. Han, P. Xie, H. Zheng and Z. Liu, Few-NERD: A Few-shot Named Entity Recognition Dataset, in: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Association for Computational Linguistics, 2021, pp. 3198–3213. doi: 10.18653/v1/2021.acl-long.248.
[20] R. Costa, H.O. Albuquerque, G. Silvestre, N.F.F. da Silva, E. Souza, D. Vitório, A. Nunes, F. Siqueira, J.P.M. Tarrega, J.V.P. Beinotti, M. de Souza Dias, F.S.F. Pereira, M. Silva, M. de Mattos Gardini, V.A.P. da Silva, A.C.P.L.F. de Carvalho and A.L.I. de Oliveira, Expanding UlyssesNER-Br Named Entity Recognition Corpus with Informal User-Generated Text, in: *Progress in Artificial Intelligence – 21st EPIA Conference on Artificial Intelligence, EPIA 2022, Lisbon, Portugal, August 31–September 2, 2022, Proceedings*, Lecture Notes in Computer Science, Vol. 13566, Springer, 2022, pp. 767–779. doi: 10.1007/978-3-031-16474-3_62.
[21] T. Finin, W. Murnane, A. Karandikar, N. Keller, J. Martineau and M. Dredze, Annotating Named Entities in Twitter Data with Crowdsourcing, in: *Proceedings of the 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, Los Angeles, USA, June 6, 2010*, Association for Computational Linguistics, 2010, pp. 80–88.
[22] I. Amazon, Amazon Mechanical Turk, 2005, Accessed: [2023].
[23] I. Figure Eight, CrowdFlower, 2007, Accessed: [2023].
[24] J. Eisenstein, What to do about bad language on the internet, in: *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2013, pp. 359–369.
[25] A. Ritter, S. Clark, Mausam and O. Etzioni, Named Entity Recognition in Tweets: An Experimental Study, in: *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, EMNLP 2011, 27–31 July 2011, John McIntyre Conference Centre, Edinburgh, UK, A Meeting of SIGDAT, a Special Interest Group of the ACL*, ACL, 2011, pp. 1524–1534.
[26] L. Derczynski, D. Maynard, G. Rizzo, M. Van Erp, G. Gorrell, R. Troncy, J. Petrak and K. Bontcheva, Analysis of named entity recognition and linking for tweets, *Information Processing & Management* **51**(2) (2015), 32–49.
[27] L. Derczynski, K. Bontcheva and I. Roberts, Broad Twitter Corpus: A Diverse Named Entity Recognition Resource, in: *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11–16, 2016, Osaka, Japan*, ACL, 2016, pp. 1169–1179.
[28] O. Sainz, I. García-Ferrero, R. Agerri, O.L. de Lacalle, G. Rigau and E. Agirre, GoLLIE: Annotation Guidelines improve Zero-Shot Information-Extraction, CoRR abs/2310.03668, 2023. doi: 10.48550/ARXIV.2310.03668. https://doi.org/10.48550/arXiv.2310.03668.
[29] V. Singh, D. Vijay, S.S. Akhtar and M. Shrivastava, Named Entity Recognition for Hindi-English Code-Mixed Social Media Text, in: *Proceedings of the Seventh Named Entities Workshop, NEWS@ACL 2018, Melbourne, Australia, July 20, 2018*, Association for Computational Linguistics, 2018, pp. 27–35. doi: 10.18653/v1/w18-2405.
[30] C. X, Versions | Docs | Twitter Developer Platform, 2019, Accessed: [2023].
[31] S. Khanuja, S. Dandapat, A. Srinivasan, S. Sitaram and M. Choudhury, GLUECoS: An evaluation benchmark for code-switched NLP, arXiv preprint arXiv:2004.12376, 2020.
[32] H. Jiang, Y. Hua, D. Beeferman and D. Roy, Annotating the Tweebank corpus on named entity recognition and building NLP models for social media analysis, CoRR, 2022.
[33] L. Yijia, Tweebank v2, 2021, Accessed: [2023].
[34] K. Gimpel, TweetNLP, 2011, Accessed: [2023].
[35] J. Fu, P. Liu and Q. Zhang, Rethinking generalization of neural models: A named entity recognition case study, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34, 2020, pp. 7732–7739.
[36] B. Strauss, B. Toma, A. Ritter, M. de Marneffe and W. Xu, Results of the WNUT16 Named Entity Recognition Shared Task, in: *Proceedings of the 2nd Workshop on Noisy User-generated Text, NUT@COLING 2016, Osaka, Japan, December 11, 2016*, The COLING 2016 Organizing Committee, 2016, pp. 138–144.
[37] J. Kim, Y. Kim and S. Kang, Weakly labeled data augmentation for social media named entity recognition, *Expert Systems with Applications* **209** (2022), 118217. doi: 10.1016/j.eswa.2022.118217.
[38] T. Nayak, N. Majumder, P. Goyal and S. Poria, Deep neural approaches to relation triplets extraction: A comprehensive survey, *Cognitive Computation* **13**(5) (2021), 1215–1232. doi: 10.1007/s12559-021-09917-7.
[39] C. Kleissner, Data Mining for the Enterprise, in: *Thirty-First Annual Hawaii International Conference on System Sciences, Kohala Coast, Hawaii, USA, January 6–9, 1998*, Vol. 7, IEEE Computer Society, 1998, pp. 295–304. doi: 10.1109/HICSS.1998.649224.

[40] P. Trivedi, G. Maheshwari, M. Dubey and J. Lehmann, LC-QuAD: A Corpus for Complex Question Answering over Knowledge Graphs, in: *The Semantic Web – ISWC 2017 – 16th International Semantic Web Conference, Vienna, Austria, October 21–25, 2017, Proceedings, Part II*, Lecture Notes in Computer Science, Vol. 10588, Springer, 2017, pp. 210–218. doi: 10.1007/978-3-319-68204-4_22.

[41] C. Unger, C. Forascu, V. López, A.N. Ngomo, E. Cabrio, P. Cimiano and S. Walter, Question Answering over Linked Data (QALD-5), in: *Working Notes of CLEF 2015 – Conference and Labs of the Evaluation forum, Toulouse, France, September 8–11, 2015*, CEUR Workshop Proceedings, Vol. 1391, CEUR-WS.org, 2015, pp. 1–10.

[42] C. Unger, A.N. Ngomo and E. Cabrio, 6th Open Challenge on Question Answering over Linked Data (QALD-6), in: *Semantic Web Challenges – Third SemWebEval Challenge at ESWC 2016, Heraklion, Crete, Greece, May 29–June 2, 2016, Revised Selected Papers*, Communications in Computer and Information Science, Vol. 641, Springer, 2016, pp. 171–177. doi: 10.1007/978-3-319-46565-4_13.

[43] R. Usbeck, A.N. Ngomo, B. Haarmann, A. Krithara, M. Röder and G. Napolitano, 7th Open Challenge on Question Answering over Linked Data (QALD-7), in: *Semantic Web Challenges – 4th SemWebEval Challenge at ESWC 2017, Portoroz, Slovenia, May 28–June 1, 2017, Revised Selected Papers*, Communications in Computer and Information Science, Vol. 769, Springer, 2017, pp. 59–69. doi: 10.1007/978-3-319-69146-6_6.

[44] R. Usbeck, A.-C.N. Ngomo, F. Conrads, M. Röder and G. Napolitano, 8th challenge on question answering over linked data (QALD-8), in: *Joint Proceedings of the 4th Workshop on Semantic Deep Learning (SemDeep-4) and NLIWoD4: Natural Language Interfaces for the Web of Data (NLIWOD-4) and 9th Question Answering Over Linked Data Challenge (QALD-9) co-located with 17th International Semantic Web Conference (ISWC 2018), Monterey, California, United States of America, October 8th–9th, 2018*, CEUR Workshop Proceedings, Vol. 2241, CEUR-WS.org, 2018, pp. 51–57.

[45] R. Usbeck, R.H. Gusmita, A.N. Ngomo and M. Saleem, 9th Challenge on Question Answering over Linked Data (QALD-9) (invited paper), in: *Joint Proceedings of the 4th Workshop on Semantic Deep Learning (SemDeep-4) and NLIWoD4: Natural Language Interfaces for the Web of Data (NLIWOD-4) and 9th Question Answering Over Linked Data Challenge (QALD-9) co-located with 17th International Semantic Web Conference (ISWC 2018), Monterey, California, United States of America, October 8th–9th, 2018*, CEUR Workshop Proceedings, Vol. 2241, CEUR-WS.org, 2018, pp. 58–64.

[46] T. Baldwin, M.C. de Marneffe, B. Han, Y.-B. Kim, A. Ritter and W. Xu, Shared Tasks of the 2015 Workshop on Noisy User-generated Text: Twitter Lexical Normalization and Named Entity Recognition, in: *Proceedings of the Workshop on Noisy User-generated Text*, Association for Computational Linguistics, Beijing, China, 2015, pp. 126–135. doi: 10.18653/v1/W15-4319.

[47] L. Derczynski, E. Nichols, M. van Erp and N. Limsopatham, Results of the WNUT2017 Shared Task on Novel and Emerging Entity Recognition, in: *Proceedings of the 3rd Workshop on Noisy User-generated Text, NUT@EMNLP 2017, Copenhagen, Denmark, September 7, 2017*, Association for Computational Linguistics, 2017, pp. 140–147. doi: 10.18653/v1/w17-4418.

[48] A. Sakor, I.O. Mulang', K. Singh, S. Shekarpour, M. Vidal, J. Lehmann and S. Auer, Old is Gold: Linguistic Driven Approach for Entity and Relation Linking of Short Text, in: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2–7, 2019, Volume 1 (Long and Short Papers)*, Association for Computational Linguistics, 2019, pp. 2336–2346. doi: 10.18653/v1/n19-1243.

[49] L. Liu and M.T. Özsu (eds), Encyclopedia of Database Systems, Springer US, 2009. doi: 10.1007/978-0-387-39940-9.

[50] P.A. Heeman, POS Tags and Decision Trees for Language Modeling, in: *Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, EMNLP 1999, College Park, MD, USA, June 21–22, 1999*, Association for Computational Linguistics, 1999, pp. 1–9.

[51] L.A. Ramshaw and M.P. Marcus, Text chunking using transformation-based learning, Natural language processing using very large corpora, 1999, 157–176. doi: 10.1007/978-94-017-2390-9_10.

[52] T. Bray, RFC 8259: The JavaScript object notation (JSON) data interchange format, RFC Editor, 2017.

[53] A. Znotiņš and E. Cīrule, NLP-PIPE: Latvian NLP tool pipeline, *Human Language Technologies – The Baltic Perspective* **307** (2018), 183–189. doi: 10.3233/978-1-61499-912-6-183.

[54] T. Mikolov, E. Grave, P. Bojanowski, C. Puhrsch and A. Joulin, Advances in pre-training distributed word representations, in: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7–12, 2018*, European Language Resources Association (ELRA), 2018, pp. 52–55.

[55] A. Paullada, I.D. Raji, E.M. Bender, E. Denton and A. Hanna, Data and its (dis) contents: A survey of dataset development and use in machine learning research, *Patterns* **2**(11) (2021), 100336. doi: 10.1016/j.patter.2021.100336.

[56] Jaswani, Naman, NER_dataset, 2019, Accessed: [2023].

[57] Z. Zhang, Improved adam optimizer for deep neural networks, in: *26th IEEE/ACM International Symposium on Quality of Service, IWQoS 2018, Banff, AB, Canada, June 4–6, 2018*, IEEE, 2018, pp. 1–2. doi: 10.1109/IWQoS.2018.8624183.

[58] J. Li, A. Sun, J. Han and C. Li, A survey on deep learning for named entity recognition, *IEEE Transactions on Knowledge and Data Engineering* **34**(1) (2022), 50–70. doi: 10.1109/TKDE.2020.2981314.

[59] B.Y. Lin, F.F. Xu, Z. Luo and K.Q. Zhu, Multi-channel BiLSTM-CRF Model for Emerging Named Entity Recognition in

Social Media, in: *Proceedings of the 3rd Workshop on Noisy User-generated Text, NUT@EMNLP 2017, Copenhagen, Denmark, September 7, 2017*, Association for Computational Linguistics, 2017, pp. 160–165. doi: 10.18653/v1/w17-4421.

[60] P. Jansson and S. Liu, Distributed Representation, LDA Topic Modelling and Deep Learning for Emerging Named Entity Recognition from Social Media, in: *Proceedings of the 3rd Workshop on Noisy User-generated Text, NUT@EMNLP 2017, Copenhagen, Denmark, September 7, 2017*, Association for Computational Linguistics, 2017, pp. 154–159. doi: 10.18653/v1/w17-4420.

[61] G. Aguilar, S. Maharjan, A.P. López-Monroy and T. Solorio, A Multi-task Approach for Named Entity Recognition in Social Media Data, in: *Proceedings of the 3rd Workshop on Noisy User-generated Text*, Association for Computational Linguistics, Copenhagen, Denmark, 2017, pp. 148–153. doi: 10.18653/v1/W17-4419.

[62] R. Priyadharshini, B.R. Chakravarthi, M. Vegupatti and J.P. McCrae, Named entity recognition for code-mixed Indian corpus using meta embedding, in: *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, IEEE, 2020, pp. 68–72.

[63] P. Lison, J. Barnes, A. Hubin and S. Touileb, Named Entity Recognition without Labelled Data: A Weak Supervision Approach, in: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5–10, 2020*, D. Jurafsky, J. Chai, N. Schluter and J.R. Tetreault, eds, Association for Computational Linguistics, 2020, pp. 1518–1533. doi: 10.18653/V1/2020.ACL-MAIN.139. https://doi.org/10.18653/v1/2020.acl-main.139.

[64] M. Miceli, J. Posada and T. Yang, Studying up machine learning data: Why talk about bias when we mean power, *Proceedings of the ACM on Human-Computer Interaction* **6**(GROUP) (2022), 1–14.

[65] X. Schmitt, S. Kubler, J. Robert, M. Papadakis and Y. LeTraon, A replicable comparison study of NER software: StanfordNLP, NLTK, OpenNLP, SpaCy, Gate, in: *2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, IEEE, 2019, pp. 338–343.

[66] W. Li, Y. Du, X. Li, X. Chen, C. Xie, H. Li and X. Li, UD_BBC: Named entity recognition in social network combined BERT-BiLSTM-CRF with active learning, *Engineering Applications of Artificial Intelligence* **116** (2022), 105460. doi: 10.1016/j.engappai.2022.105460.

[67] A. Goyal, V. Gupta and M. Kumar, A deep learning-based bilingual Hindi and Punjabi named entity recognition system using enhanced word embeddings, *Knowledge-Based Systems* **234** (2021), 107601. doi: 10.1016/j.knosys.2021.107601.

[68] S. Rizou, A. Paflioti, A. Theofilatos, A. Vakali, G. Sarigiannidis and K.C. Chatzisavvas, Multilingual name entity recognition and intent classification employing deep learning architectures, *Simulation Modelling Practice and Theory* **120** (2022), 102620. doi: 10.1016/j.simpat.2022.102620.

[69] M. Khalifa and K. Shaalan, Character convolutions for Arabic named entity recognition with long short-term memory networks, *Computer Speech & Language* **58** (2019), 335–346. doi: 10.1016/j.csl.2019.05.003.

[70] S.-H. Na, H. Kim, J. Min and K. Kim, Improving LSTM CRFs using character-based compositions for Korean named entity recognition, *Computer Speech & Language* **54** (2019), 106–121. doi: 10.1016/j.csl.2018.09.005.

[71] J. Chang and X. Han, Multi-level context features extraction for named entity recognition, *Computer Speech & Language* **77** (2023), 101412. doi: 10.1016/j.csl.2022.101412.

[72] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu and P. Kuksa, Natural language processing (almost) from scratch, *Journal of Machine Learning Research* **12** (2011), 2493–2537. doi: 10.5555/1953048.2078186.

[73] J.P. Turian, L. Ratinov and Y. Bengio, Word Representations: A Simple and General Method for Semi-Supervised Learning, in: *ACL 2010, Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, July 11–16, 2010, Uppsala, Sweden*, The Association for Computer Linguistics, 2010, pp. 384–394.

[74] D. Lin and X. Wu, Phrase Clustering for Discriminative Learning, in: *ACL 2009, Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the AFNLP, August 2–7, 2009, Singapore*, The Association for Computer Linguistics, 2009, pp. 1030–1038.