DOI: https://doi.org/10.15276/ict.01.2024.37
УДК 004.8

Огляд моделей машинного навчання NER для аналізу мобільних даних у криміналістиці

Ісаченко Ольга Володимирівна
Магістр каф. Інфокомунікаційної інженерії імені В.В. Поповського
E-mail: olha.isachenko@nure.ua
Харківський національний університет радіоелектроніки, пр. Науки, 14. Харків, 61166, Україна

АНОТАЦІЯ

За останні роки машинне навчання (machine learning, ML) широко поширилося у повсякденному житті. На його основі працюють програми зі штучним інтелектом, на основі якого з'вилося багато віртуальних помічників. Важливу роль ML грає і у разних сферах діяльності багатьох підприємств. ML допомагає автоматизувати багато процесів, спрощуючи функціонування компанії.

Named Entity Recognition (NER) моделі дозволяють автоматично вибирати, шукати інформацію за певними критеріями у вилучених, наприклад, логічним методом, мобільних даних.

Підтримка NER моделями Python дає можливість гнучко програмувати конкретні запроси, які стровюються в процесі криміналістичної експертизи. Відкритий код створює унікальну можливість постійно вдосконалювати модель, навчаючи її на наборах даних.

Потужним пакетом для роботи з NER є фреймворк spaCy, який допомагає спростити дані та отримати детальну інформацію з введених даних, обучити модель, зробити налаштування моделі та багато іншого. spaCy сумісний із 64-розрядним CPython 3.7+ і працює на Unix/Linux, macOS/OS X і Windows. Останні версії spaCy доступні через pip і conda.

Ключові слова: машинне навчання; NER; NLP; програмне забезпечення для криміналістичної експертизи; фреймворк SpaCy

Актуальність. Використання моделей машиного навчання NER в мобільній криміналістиці, наряду з класичними методами та програмним забезпеченням, дає можливість удосконалити збір та обробку мобільних даних. Переваги застосування ML моделей полягають у автоматичному скануванні пристрою та класифікації токенів даних.

Мета дослідження: проаналізувати типи та можливості NER моделей, їх застосування у фреймворку SpaCy, як одного з найефективніших при класифікації токенів даних, для подальшого використання у методах криміналістичної експертизи мобільних пристроїв.

Для виявлення певної інформації в тексті використовують моделі машинного навчання Named Entity Recognition – Розпізнавання іменованих об'єктів (далі – NER). NER модель може автоматично вибрати таки категорії даних як імена, номери телефонів, адреси електронної пошти, вуличні адреси, номери кредитних карток або місцезнаходження (Рис. 1).

NER є частиною більшої галузі науки про дані або аналізу даних, відомої як обробка природної мови Natural Language Processing (далі – NLP) [1].

Використання NER в криміналістичній експертизі мобільних даних дає можливість витягувати структуровану інформацію з текстових даних мобільного для подальшого аналізу.

NER зазвичай досягається за допомогою моделі. Вільний текст є вхідними даними в модель. Потім модель використовує схему маркування для виведення слів у тексті, які належать до кожного попередньо визначеного типу сутності [2].

Взагалі існує кілька систем NER:

1) На основі словника. Системи NER на основі словників посилаються на терміни, перелічені в словниках, щоб визначити їх присутність у тексті. Словники можуть являти собою будь-яку колекцію слів, пов'язаних із певною сферою чи доменом.

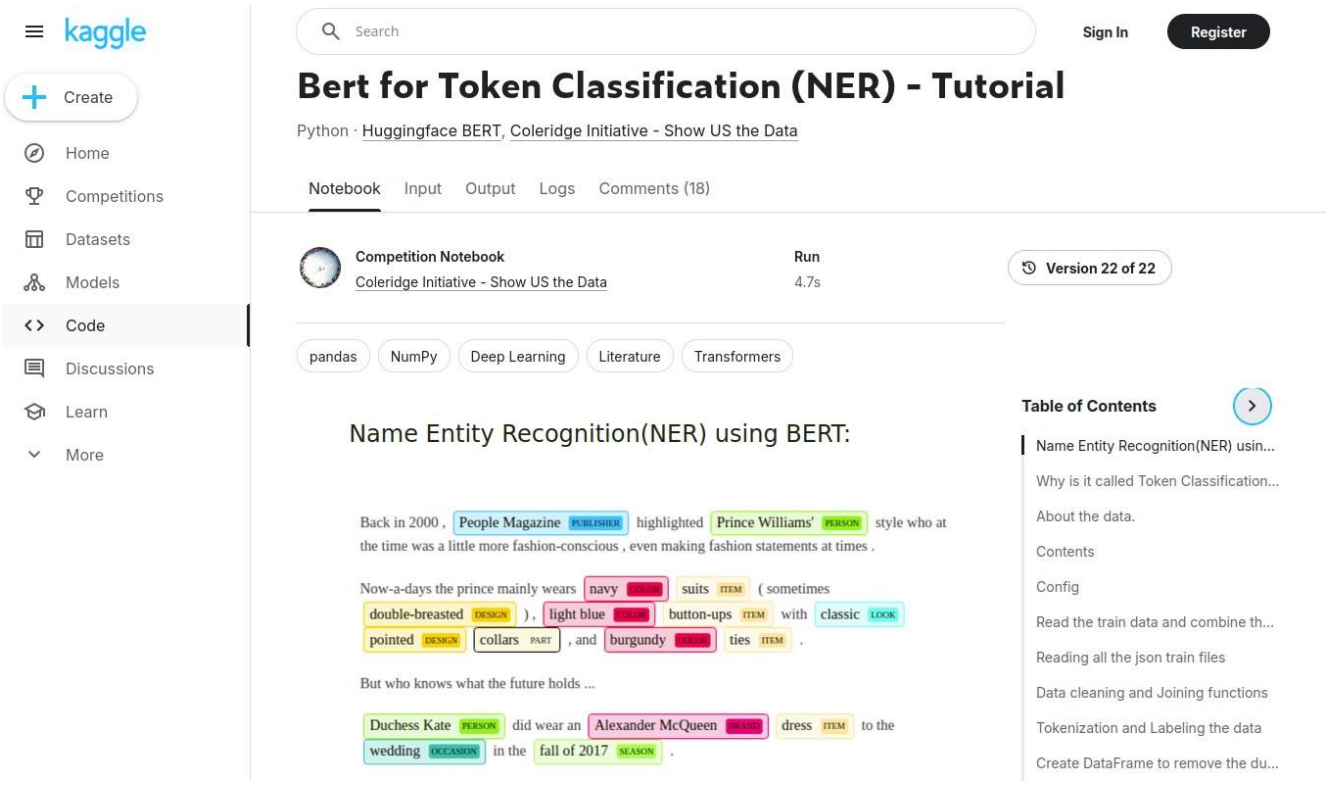

Рис. 1. **Приклад Named Entity Recognition класифікації токенів тексту**

2) На основі правил. Системи NER на основі правил покладаються на набір інструкцій для вилучення іменованих об’єктів із тексту. Повинно створити правила на основі двох типів інструкцій: правила на основі шаблонів, які стосуються форм і структури слів, і правила на основі контексту. Ці правила також можна поєднувати зі словниками.

3) На основі машинного навчання. Системи NER на основі машинного навчання базуються на статистичних моделях, призначених для ідентифікації імен об’єктів. Щоб розробити систему NER на основі ML, модель машинного навчання має бути навчена на анотованих документах. Анотовані документи містять пояснення, які допомагають машині навчитися створювати назви об’єктів на основі інструкцій і минулого досвіду.

4) Гібридні системи. Гібридні системи NER поєднують більше одного з перерахованих вище підходів [3].

Найпоширенішими типами моделей NER є моделі на основі правил, машинного навчання або моделі штучного інтелекту Artificial intelligence (далі – AI) нейронної мережі.

Щоб створити модель AI з машинним навчанням або нейронною мережею, необхідно навчити модель на вільних текстових даних, де вже ідентифіковано типи об’єктів, які цікавлять. У цьому випадку модель NER вивчає шаблони сутностей на навчальних даних, а потім застосовує ці шаблони для ідентифікації сутностей у будь-якому вільному тексті.

Однією з переваг моделі AI NER є те, що вона може навчитися визначати, чи є слово об’єктом інтересу, на основі контексту навколо слова. Наприклад, модель AI NER може використовувати контекст речення, щоб визначити, чи слово Apple стосується організації чи плоду. Модель на основі правил може мати проблеми з чимось подібним.

Ефективність NER визначається як швидкість обробки текстових даних у моделі NER. Зазвичай це вимірюється словами за секунду.

За методологічним визначенням моделі розділяють на:

– рекурентні нейронні мережі (recurrent neural networks, RNN) і довготривала короткочасна пам'ять (long short-term memory, LSTM). RNN – це тип нейронної мережі, розроблений для задач прогнозування послідовності. LSTM, особливий вид RNN, який може навчитися розпізнавати шаблони з часом і зберігати інформацію в «пам’яті» протягом довгих

послідовностей, що робить їх особливо корисними для розуміння контексту та ідентифікації об’єктів;

– умовні випадкові поля (conditional random fields, CRF). CRF часто використовуються в поєднанні з LSTM для завдань NER. Вони можуть моделювати умовну ймовірність цілої послідовності міток [10], а не лише окремих міток, що робить їх корисними для завдань, де мітка слова залежить від міток навколишніх слів;

– трансформери і BERT. Трансформні мережі, зокрема модель BERT (Bidirectional Encoder Representations from Transformers), мали значний вплив на NER. Використовуючи механізм самоконтролю, який зважує важливість різних слів, BERT враховує повний контекст слова, дивлячись на слова, які стоять перед і після нього [2] (Рис. 2).

Логіка роботи моделей NER:

– токенізація тексту, текст розбивається на окремі слова або токени;

– виділення позначок – кожному токену призначаються позначки, які описують його середовище та контекст, такі як попередні та наступні слова, частини речей та інші лінгвістичні характеристики;

– застосування моделі – модель аналізує розпізнавання кожного токена та визначає, є він іменованою сутністю чи ні;

– об’єднання результатів – результати аналізу токенів об’єднуються, щоб сформувати іменовані сутності, і їм призначаються відповідні метки класів, такі як “PER” (для особи) або “ORG” (для організації) інш;

– постобробка – доповнювальна обробка для уточнення результатів або виправлення помилок.

Бібліотеки NER надають готові рішення для виділення іменованих сутностей і можуть бути адаптовані під конкретні завдання та типи даних.

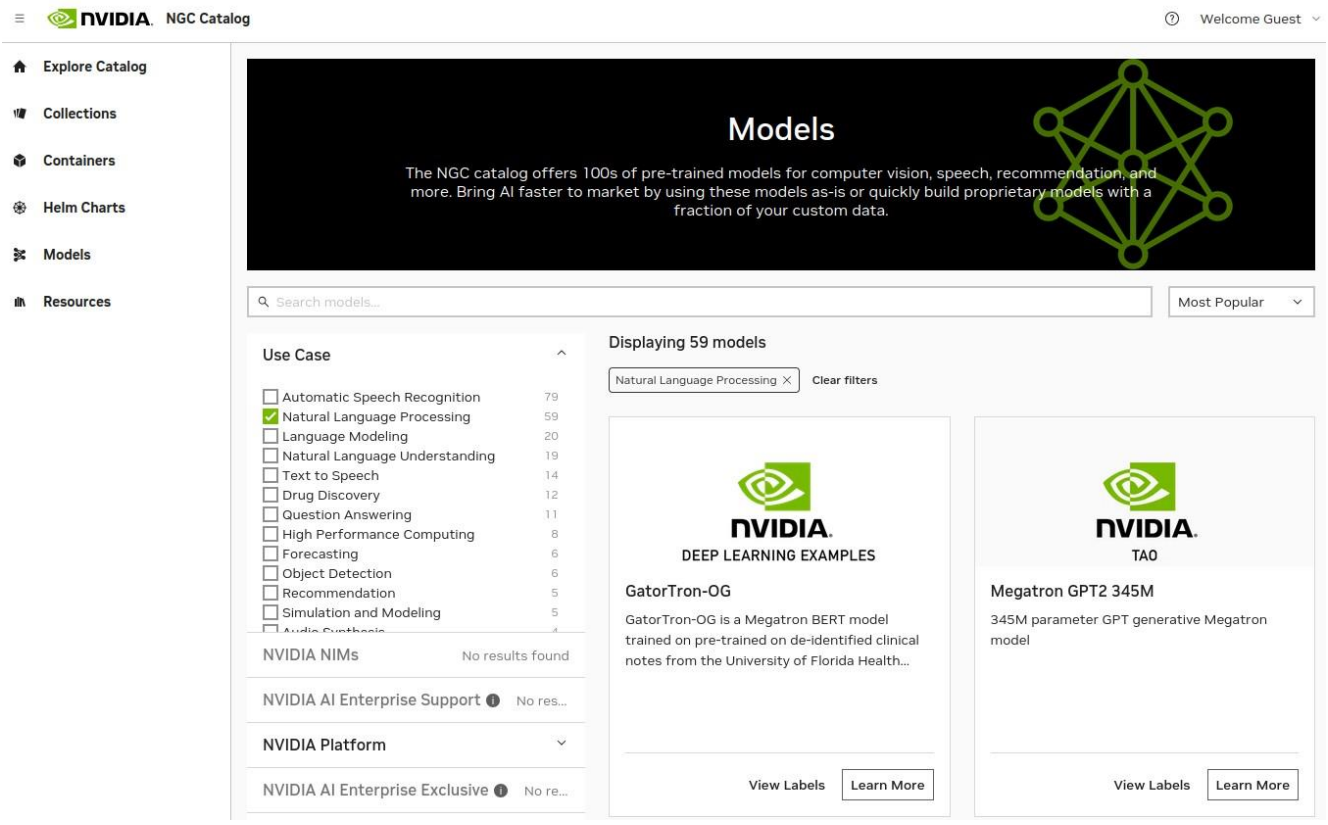

Рис. 2 **Моделі Nvidia Corporation**

Існує декілька способів реалізувати пошук інформації за NER принципом. Один з найпопулярніших – SpaCy. Написана на Python (Python – це потужна мова програмування, яка проста у вивченні. Він має ефективні структури даних високого рівня та простий, але ефективний підхід до об'єктно-орієнтованого програмування [5]) і відома своєю швидкістю та зручністю, SpaCy – це бібліотека програмного забезпечення з відкритим кодом для просунутого NLP. Його створено на основі останніх досліджень і розроблено для використання з реальними продуктами. Вона також має розширену статистичну систему, яка дозволяє користувачам створювати налаштовані екстрактори NER (Рис. 3).

SpaCy має наступні переваги:

– виявлення речень і токенізація: spaCy може розбити вхідний текст на лінгвістично значущі або базові одиниці для подальшого аналізу;

– стоп видалення слів: spaCy може видалити вказані слова, щоб вони не зіпсувати такі завдання, як частотний аналіз слів;

– резюмування тексту: spaCy може зменшити неоднозначність, резюмувати та витягувати з тексту найбільш релевантну інформацію, таку як особа, місцезнаходження чи компанія, за допомогою функції лемматизації та розпізнавання іменованих сутностей;

– переклад мов: spaCy та інші інструменти обробки природної мови можуть використовувати глибоке навчання для перекладу мови та тексту різними мовами, навіть для спеціалізованих галузей і доменів;

– розбір залежностей і подібності: spaCy призначає мітки синтаксичних залежностей, які вивчають, як слова пов'язані одне з одним у заданому тесті;

– зіставлення на основі правил: spaCy може витягувати певні шаблони в тексті, наприклад повні імена, номери телефонів і дні народження;

– spaCy постачається з вбудованим візуалізатором під назвою displaCy, який візуалізує синтаксичний аналіз залежностей або іменовані сутності в браузері чи блокноті Jupyter.

Останній випуск spaCy 3.0 містить багато вдосконалень, які допомагають створювати, налаштовувати та підтримувати моделі NLP [6].

Висновки. Розпізнавання іменованих об'єктів (NER) – це важливе завдання в обробці природної мови (NLP), яке передбачає ідентифікацію та класифікацію іменованих об'єктів у тексті за попередньо визначеними категоріями, такими як імена осіб, організації, місця розташування, вираження часу, кількості, грошові значення, відсотки, тощо.

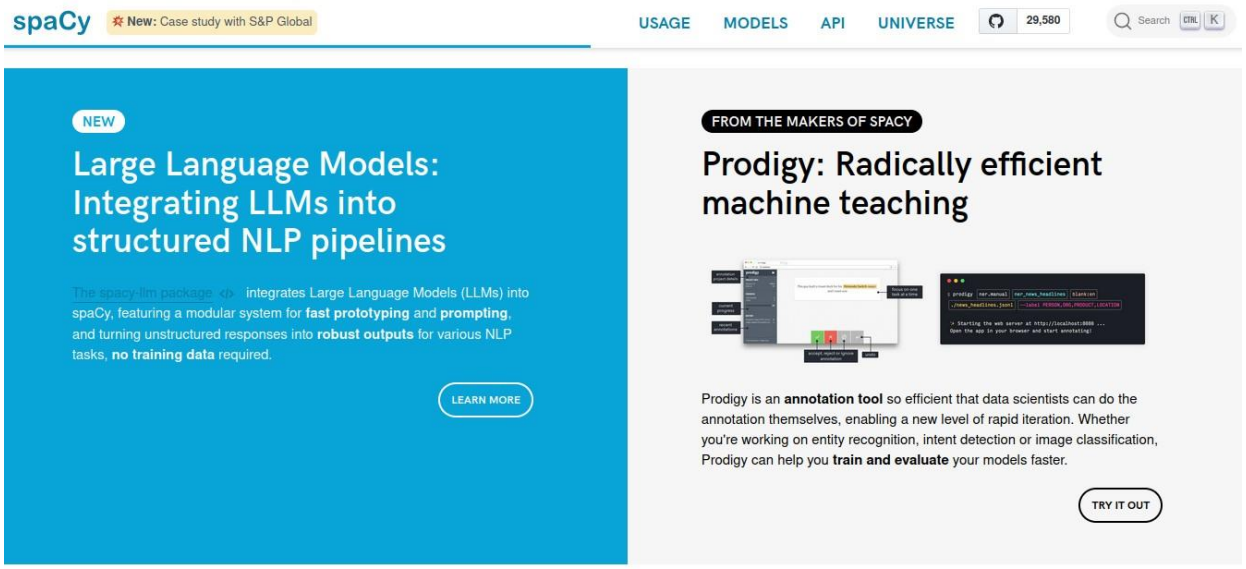

Рис. 3. **Можливості SpaCy фремворку**

Розпізнавання іменованих об’єктів може використовувати в різних сферах, так застосування у мобільній кримінаістиці дасть змогу швидкого пошуку інформації, класифікаціі вмісту вилучених мобільних даних.

SpaCy робить створення системи NER надзвичайно простим, тому як варіант, для невеликих об’ємів даних та нескланих умов, можно використовувати цей фреймворк.

СПИСОК ЛІТЕРАТУРИ

1. Ferrara J. “Understanding named entity recognition (NER) models“. – Available from: https://www.tonic.ai/guides/named-entity-recognition-models. – (Дата звернення: 26.08.2024).

2. «Інформація з вебсайту “IBM”». – Доступно з: https://www.ibm.com/topics/named-entity-recognition. – (Дата звернення: 22.08.2024).

3. Wu R., Guo X., Du J., Li J. “Accelerating neural network inference on FPGA-based platforms – A survey”. *Electronics.* 2021; 10 (9): 1025. DOI: https://doi.org/10.3390/electronics10091025.

4. Schulze J. “ What is named entity recognition (NER) and how does it work?”. – Available from: https://www.coursera.org/articles/named-entity-recognition. – (Дата звернення: 10.09.2023).

5. Firuzan A., Modarressi M., Reshadi M. & Khademzadeh A. “Reconfigurable network-on-chip based convolutional neural network accelerator”. *Journal of Systems Architecture.* 2022; 129: 102567. DOI: https://doi.org/10.1016/j.sysarc.2022.102567.

6. «Інформація з вебсайту Wikipedia, “spaCy”». – Доступно з: https://en.wikipedia.org/wiki/SpaCy. – (Дата звернення: 24.08.2024).

7. «Підручник з Python». – Available from: https://docs.python.org/uk/3/tutorial/index.html. – (Дата звернення: 22.08.2024).

8. “Інформація з вебсайту “SpaCy””. – Доступно з: https://domino.ai/data-science-dictionary/spacy. – (Дата звернення: 21.08.2024).

9. Mattingly W. “Introduction to named entity recognition”. – Доступно з: https://ner.pythonhumanities.com/intro.html 2021. – (Дата звернення: 23.08.2024).

10. Sohi G. “Instruction issue logic for high-performance interruptible, multiple functional units,pipe-lined computers”. *IEEE Trans Comput*. 1990; 39 (3): 349–359. DOI: https://doi.org/10.1109/12.48865.

11. Nurvitadhi E., Venkatesh G., Sim J., Marr D., et al. “Can FPGAs Beat GPUs in accelerating next-generation deep neural networks?” *In: Proceedings of the 2017 ACM/SIGDA International Sympo-sium on Field-Programmable Gate Arrays*. 2017. p. 5–14. – Available from: https://jaewoong.org/pubs/fpga17-next-generation-dnns.pdf. – [Accessed: 22, Dec. 2022].

DOI: https://doi.org/10.15276/ict.01.2024.37
UDC 004.8

Estimates of the accuracy of identification of a nonlinear dynamic system using step test signals

Olha V. Isachenko
Master, faculty Infocommunication Engineering named after V.V. Popovsky
E-mail: olha.isachenko@nure.ua
Kharkiv National University of Radio Electronics, Nauky Ave. 14. Kharkiv, 61166, Ukraine

ABSTRACT

Machine learning (ML) has become widespread in everyday life. On its basis, programs with artificial intelligence work, on the basis of which many virtual assistants have evolved. ML plays an important role in various spheres of activity of many enterprises. ML helps to automate many processes, simplifying the functioning of the company.

Named Entity Recognition (NER) models allow to automatically select and search for information according to certain criteria in mobile data extracted, for example, by the logistic method.

Support for NER by Python models makes it possible to flexibly program specific requests that are generated in the forensic examination process. Open source creates a unique opportunity to continuously improve the model by training it on datasets.

A powerful NER package is the spaCy framework, which helps to simplify data and extract detailed information from input data, train a model, perform model tuning, and more. spaCy is compatible with 64-bit CPython 3.7+ and runs on Unix/Linux, macOS/OS X and Windows. The latest spaCy releases are available over pip and conda.

Keywords: Machine learning; NER; NLP; forensics software; SpaCy framework