Hyperparameter Tuning for Address Validation using Optuna

MARIYA EVTIMOVA
UFR27,
Paris 1 Sorbonne- Pantheon, CRI,
90 Rue de Tolbiac, 75013 Paris,
FRANCE

Abstract: - Public institutions generally share personal information on their websites. That allows the possibility to find personal information when performing internet searches quickly. However, the personal information that is on the internet is not always accurate and can lead to misunderstandings and ambiguity concerning the accessible postal address information. That can be crucial if the information is used to find the location of the corresponding person or to use it as a postal address for correspondence. Many websites contain personal information, but sometimes as people change the web address, information is not up to date or is incorrect. To synchronize the available personal information on the internet could be used an algorithm for validation and verification of the personal addresses. In the paper, a hyperparameter tuning for address validation using the ROBERTa model of the Hugging Face Transformers library. It discusses the implementation of hyperparameter tuning for address validation and its evaluation to achieve high precision and accuracy.

Key-Words: - Hyperparameter Tuning, Hugging Face Transformers Library, Optuna, Machine Learning algorithm, Address validation, ROBERTa model, Postal address

Received: June 2, 2022. Revised: August 29, 2023. Accepted: October 5, 2023. Available online: November 13, 2023.

1 Introduction

The proliferation of digital information on the internet is experiencing exponential growth, with numerous websites, particularly those affiliated with public institutions, disseminating personal data. Nonetheless, it's important to note that the data obtained from diverse internet sources is not always correct, [1]. An increasing number of individuals are turning to online resources to access pertinent address-related information, as the internet's unhindered accessibility facilitates swift communication, [2]. However, this accessibility carries the inherent risk of encountering outdated information, potentially leading to misinterpretations and severed connections with intended recipients, [3]. Thus, the accurate acquisition of address information stands as pivotal for effective communication.

It's crucial to emphasize that the composition of postal address entries displays greater diversity compared to conventional descriptors, [4], [5]. Conversely, not all components of an address are essential for practical use. For instance, the provided address above notably excludes the "Postal code" entry. Similarly, instances exist where suite details are absent. A significant challenge in verifying internet-sourced addresses pertains to the intricacies of vocabulary usage. Language nuances can breed semantic confusion due to the presence of synonyms (words with identical meanings) and polysemy (a single word with divergent meanings).

As an example, the term "Avenue" might be abbreviated variably as "Av.," "av," or "Ave." Further exemplifying polysemy, the street name "rue de Tivoli" is present in both Marseilles and Paris, illustrating this phenomenon, [6].

It's pertinent to acknowledge that the order of address components differs among various institutions. Notably, certain establishments prioritize the inclusion of street numbers before street names, while others position house addresses after the street name.

Complications arising from erroneous address databases, characterized by duplicate entries and inconsistent variants, give rise to a complex landscape where accurate data retrieval becomes both arduous and unreliable. The data classification encompassing names and addresses is underscored by idiosyncratic attributes that contribute to distinct complexities in their management. This data domain is particularly susceptible to volatility due to the dynamic nature of institutional affiliations, address changes, and name modifications. Moreover, data input for names and addresses frequently exhibits

cluttered tendencies, as front-end interfaces allow free-form entry, often incorporating comments and additional data without validation.

Further complicating matters is the subjective nature inherent in the representation of names and addresses. Individuals possess the liberty to express identical entities in varied ways, despite referring to the same entity. Unfortunately, a universally accepted standard or framework that could encapsulate name and address data while simultaneously evaluating its quality remains conspicuously absent. This challenge is compounded by the intricate interplay of cultural contexts within France, which significantly influence the interpretation and management of name and address data.

2 Related Work

2.1 The Sections Present Recent Relative Research of the ROBERTA Language Model Applied in Address Verification

In the study conducted by, [7], a deep learning methodology is introduced to enhance the quality of global address data utilized for imported food safety management. The proposed approach involves the classification of user input addresses into administrative divisions specific to the respective country. By transforming the addresses into standardized formats, the quality of the address data is assessed and enhanced.

Within the context of research, [8], a filtering approach named Distill-AER has been put forth. This technique is designed to facilitate the transformation of knowledge extracted from a well-populated labeled dataset of standard addresses in the realm of Big Data. The objective is to adapt this knowledge for the targeted task of recognizing entities within addresses that hold a specific significance. To facilitate this transfer, a labeled spoken dialogue dataset containing address entities is constructed through the utilization of the data augmentation paradigm.

The study, [9], introduces a two-stage address validation methodology that incorporates standardization and classification steps, both leveraging the ROBERTa pre-trained language model. The proposed approach is evaluated through experiments conducted on real datasets, showcasing its effectiveness and reliability.

2.2 Other Machine Learning Techniques Applied for Address Validation

The study, [10], presents HyPASS, a software detailed in their study, which encompasses a hybrid approach combining Software-Defined Networking (SDN) principles with host discovery and address validation techniques. The primary objective of HyPASS is to mitigate source spoofing attacks by enhancing network security measures.

The paper by, [11], introduces an automated probabilistic method, based on a hidden Markov model (HMM), that utilizes national address guidelines and an extensive national address database. This approach aims to process raw input addresses by performing cleaning, standardization, and verification tasks.

The study, [12], introduces a novel robust architecture called DeepParse in their paper, which is specifically designed for postal address parsing. This architecture implements address parsing and also reflects Named Entity Recognition (NER) problems. DeepParse treats input data at various levels such as characters, trigram characters, and words, to extract features and carry out address validation. The model was trained using a synthetically generated dataset and subsequently tested with real addresses.

In the research carried out by, [13], an application of one-dimensional transformation in CNN (Convolutional Neural Networks) is utilized for address parsing.

The results obtained from the evaluation demonstrate a high accuracy for a labeled dataset comprising nearly 20,000 samples. Notably, the proposed network architecture possesses a scalable nature, eliminating the need for any pre- or post-processing stages.

The authors in, [14], introduce a tool designed for the annotation of electronic health records. The research focuses on training random forests to identify patients who are experiencing homelessness. To validate the efficacy of each model, a 10-fold cross-validation technique is employed.

3 Standards for Postal Address in France

3.1 French Standard Overview

In France, postal addresses follow a specific format. Here is the standard structure for a postal address in France:

- Recipient's Name: Full name of the person or organization to which the mail is addressed.
- Building Number and Street Name: The building number comes before the street name. For example, "12 Rue de la Paix" (12, Peace Street).
- Postal Code and City: The five-digit postal code comes before the name of the city. For example, "75001 Paris" (75001 being the postal code for the 1st arrondissement of Paris).
- Cedex (Optional): If the recipient's address includes a Cedex (Courrier d'Entreprise à Distribution Exceptionnelle) number, it should be included on a separate line below the postal code and city. Cedex is used for mail addressed to specific companies, organizations, or government agencies that have their distribution system. For example, "Cedex 2" or "CEDEX 2".
- Country (Not required for domestic mail)

If the mail is being sent from outside of France, the country name should be included in uppercase letters at the last line of the address. For domestic mail within France, the country name is not necessary.

An example of a postal address in France:
Monsieur Jean Dupont
12 Rue de la Paix
75001 Paris
France

It's important to note that the exact format may vary slightly depending on the region or specific requirements of the local postal service, [15].

It is prudent to consistently cross-validate prevailing norms and directives set forth by La Poste, the national postal service of France. Alternatively, one may opt to engage in dialogue with the intended recipient to ascertain any supplementary details that could potentially be deemed requisite.

3.2 Proposal for Address Model

Fig. 1 introduces an address denoted as Adr = {adr $_1$, ..., adr $_n$}, where adr signifies a collection of addresses adr_i, and 'i' denotes a word, while 'n' signifies the address's length. The primary objective of parsing Adr is to allocate a label 'l' to each word adr_i within Adr from the corresponding set of address tags denoted as T;

T = {P, C, CO, SR, BN, CE, SN, PB}.

These tags find their definitions in the address model depicted in Figure 1, whereby the address tags symbolize composite attributes.

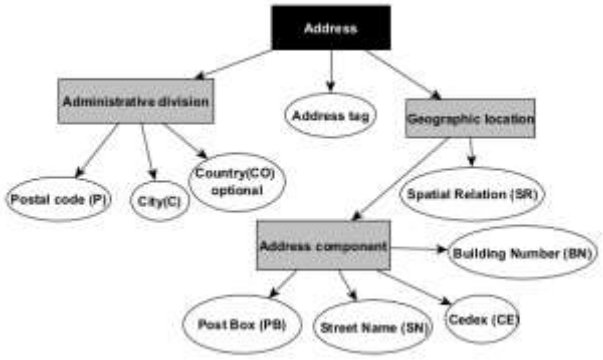

Fig. 1: Address model

There are also several commercial paid applications and services that could be used for the verification of postal addresses like la Poste API, Google Maps Geocoding API, SmartyStreets, Experian Data Quality, and Loqate.

4 Overview of the Tasks Involved in Hyperparameter Tuning

1. Data Preparation:

Collection and preprocessing of the proposed datasets for address validation (French BAN corpus and French higher education).

This step includes:

- cleaning and labeling data with both correct and incorrect addresses
- splitting the data into training, validation, and test sets.

2. Pre-trained Model Selection:

Hugging Face's ROBERTa model is chosen for the address validation task.

3. Tokenization:

Choosing an appropriate tokenization strategy for the data address validation task.

4. Model Architecture:

Include fine-tuning the ROBERTa model for address validation.

5. Hyperparameter Tuning:

Hyperparameters tuning can include modification of the following parameters, [14]:

- Learning Rate: parameter essential for model convergence between 1e-5 and 1e-3.

Techniques such as learning rate schedules can be applied.
- Batch Size: adjustment of the batch size concerning the computer hardware limitations and the size of the data. Normally, the smaller batch sizes correspond to smaller learning rates.
- Number of Epochs: determination of the appropriate number of epochs through experimentation and avoiding overfitting
- Weight Decay: regularization of the parameter to control overfitting.
- Warmup Steps: implementation of warm-up steps for the learning rate scheduler.
- Gradient Accumulation: useful parameter for limited GPU memory.
- Early Stopping: This can be used to avoid overfitting.
- Loss Function: choosing or customization of the loss function that is suitable for address validation. It should reflect the nature of your task, possibly considering token-level or sequence-level metrics.
- Evaluation Metrics: definition of evaluation metrics that are relevant for address validation, such as accuracy, F1 score, or other custom metrics.

6. Other machine learning techniques that could be applied, [16], [17]:
- Regularization: application of dropout or other regularization techniques to prevent overfitting.
- Fine-tuning Strategy: experimentation with different fine-tuning strategies. Techniques like gradual unfreezing of layers or differential learning rates can be used.
- Data Augmentation: augmentation of the dataset with variations of addresses, to improve model robustness.
- Cross-Validation: performance of k-fold cross-validation to assess the model's generalization and identify optimal hyperparameters.
- Grid search or Random search: consider using grid search or random search to automate hyperparameter tuning.
- Monitoring and Logging: implementation of a system to monitor and log the training process and results, allowing tracking of the performance of different hyperparameters.
- Hardware and Parallelism: depending on the resources available, it is possible to explore distributed training to speed up the tuning process.
- Deployment and Inference:
deployment of the model for inference in a production environment.

5 Description of the Algorithm for Hyperparameter Tuning using Optuna

Figure 2 describes the algorithm for hyperparameter tuning that is proposed for address validation.

```
Create a study with maximize direction of optimization and 10 number of trials;
Define a study- specific hyperparameter search space:
        learning_rate:=(1e-6,1e-3);
Load the ROBERTa pre- trained tokenizer and model;
Prepare the training data;
        Define the input addresses;
        Replace the actual labels with 0 for incorrect address and 1 for correct address;
Tokenize the input text;
Creation of DataLoader for the training data
Set up optimizer with the trial suggested learning rate with number=:3 epochs;
Fine- tuning loop:
        For epoch in range of number of epochs
                For batch in DataLoader
Return the evaluation metric value
```

Fig. 2: Pseudo-code of the algorithm for hyperparameter tuning for address validation

6 Experiments and Results

The last phase of the model defined in Figure 2 is described in this section.

Currently, the BAN contains 25 million addresses across France. The parsing and classification of the dataset were conducted using two real-world datasets.

The first dataset is the BAN, which comprises millions of structured addresses extracted from the French database. The second dataset is a collection of 3,683 structured addresses from a French higher education database.

- The French BAN corpus, comprising 25 million addresses extracted from the database.

- The French higher education dataset encompasses a collection of 3,683 addresses.

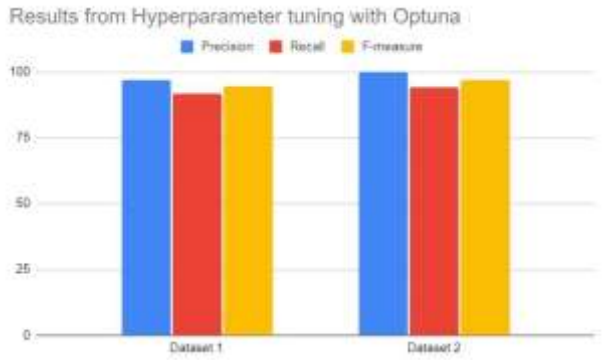

Fig. 3: Evaluation of the datasets

From the data provided from the evaluation represented in Figure 3, it is possible to conclude that the results when using Dataset 2 have high precision and F- measure.

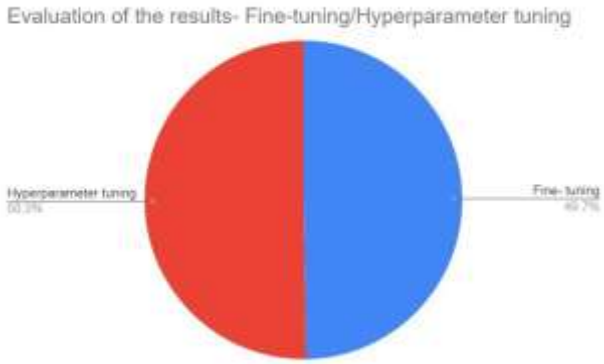

Fig. 4: Comparison of the evaluation results when applying fine-tuning and hyperparameter tuning.

Fig.4 is presented as a comparison of the results with hyperparameter tuning and only with fine-tuning that is presented in the article, [18]. The results show better performance when applying hyperparameter tuning with Optuna.

7 Conclusion

The increasing number of internet sites containing personal information and addresses is constantly growing on the internet, and this data needs to be synchronized and updated, [18]. The websites providing personal information are typically public institutions' websites. This information is generally used by everyone to find addresses (for writing letters or locating office coordinates).

These sites change their content frequently, and the information needs to be updated daily to ensure the online information is accurate, [19], [20]. This has led to the development of an algorithm for address verification using hyperparameter tuning with Optuna, which can be used to align data from various internet sources. The paper describes an algorithm for address verification using the Optuna with ROBERTa model, [21], [22], [23]. The proposed algorithm achieves a 98.5\% accuracy rate and a relatively high F-measure in comparison to the other algorithms applied in address validation, [24], [25], [26].

References:

[1] Cai, Wentao, Shengrui Wang, and Qingshan Jiang. "Address extraction: Extraction of location-based information from the web", *Web Technologies Research and Development-APWeb 2005: 7th Asia-Pacific Web Conference*, Shanghai, China, March 29-April 1, 2005. *Proceedings 7. Springer Berlin Heidelberg*, 2005.

[2] Fedushko, Solomia, and Yuriy Syerov. "Design of registration and validation algorithm of member's personal data", *International Journal of Informatics and Communication Technology 2.2,* 2013, pp. 93-98.

[3] Dakrory, Sara, et al. "Extracting geographic addresses from social media using deep recurrent neural networks", *2021 9th International Japan-Africa Conference on Electronics, Communications*, and Computations (JAC-ECC). IEEE, 2021.

[4] Beverly, Robert, et al. "Understanding the efficacy of deployed internet source address validation filtering", *Proceedings of the 9th ACM SIGCOMM conference on Internet measurement*, 2009.

[5] Nagabhushan, P., Shanmukhappa A. Angadi, and Basavaraj S. Anami. "Symbolic data structure for postal address representation and address validation through symbolic knowledge base", *Pattern Recognition and Machine Intelligence: First International Conference, PReMI 2005*, Kolkata, India, December 20-22, 2005. Proceedings 1. Springer Berlin Heidelberg, 2005.

[6] U.S. POSTAL SERVICE FACILITIES: Improvements in Data Would Strengthen Maintenance and Alignment of Access to Retail Services, *GAO Report,* December 2007:i-61. Accessed August 29, 2023.

[7] Soeng, Saravit, et al. "Deep Learning Based Improvement in Overseas Manufacturer Address Quality Using Administrative District Data", *Applied Sciences* 12.21, 2022, vol. 11129.

[8] Wang, Yitong, et al. "Distill-AER: Fine-Grained Address Entity Recognition from

Spoken Dialogue via Knowledge Distillation", *Natural Language Processing and Chinese Computing: 11th CCF International Conference, NLPCC 2022, Guilin, China*, September 24–25, 2022, Proceedings, Part I. Cham: Springer International Publishing, 2022.

[9] Guermazi, Yassine, Sana Sellami, and Omar Boucelma. "A RoBERTa Based Approach for Address Validation", *New Trends in Database and Information Systems: ADBIS 2022 Short Papers*, Doctoral Consortium and Workshops: DOING, K-GALS, MADEISD, MegaData, SWODCH, Turin, Italy, September 5–8, 2022, Proceedings. Cham: Springer International Publishing, 2022.

[10] Meena, Ramesh Chand, et al. "HyPASS: Design of hybrid-SDN prevention of attacks of source spoofing with host discovery and address validation", *Physical Communication 55*, 2022, vol .101902.

[11] Christen, Peter, and Daniel Belacic. "Automated probabilistic address standardisation and verification.", *Australasian Data Mining Conference*, 2005.

[12] Abid, Nosheen, Adnan ul Hasan, and Faisal Shafait, "DeepParse: A trainable postal address parser.", *2018 Digital Image Computing: Techniques and Applications (DICTA)*, IEEE, 2018.

[13] Delil, Selman, et al. "Parsing Address Texts with Deep Learning Method", *2020 28th Signal Processing and Communications Applications Conference (SIU)*, IEEE, 2020.

[14] Erickson, Jennifer, Kenneth Abbott, and Lucinda Susienka, "Automatic address validation and health record review to identify homeless Social Security disability applicants.", *Journal of Biomedical Informatics, vol.*82, 2018, pp. 41-46.

[15] YANG, Li; SHAMI, Abdallah, "On hyperparameter optimization of machine learning algorithms: Theory and practice.", *Neurocomputing*, 2020, vol.415, pp. 295-316.

[16] Akiba, Takuya, et al., "Optuna: A next-generation hyperparameter optimization framework", *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2019, p. 2623-2631.

[17] Andonie, Răzvan. Hyperparameter optimization in learning systems. *Journal of Membrane Computing*, 2019, 1.4, pp. 279-291

[18] Evtimova, M., "Validation algorithm for aligning postal addresses available on the Internet", MACISE conference, 2023.

[19] Basu, Subhadip, et al. "A novel framework for automatic sorting of postal documents with multi-script address blocks." Pattern Recognition 43.10, 2010, pp.3507-3521.

[20] Andonie, Răzvan. Hyperparameter optimization in learning systems. *Journal of Membrane Computing*, 2019, 1.4, pp. 279-291.

[21] Lewis, Taylor, Joseph McMichael, and Charlotte Looby, "Evaluating Substitution as a Strategy for Handling US Postal Service Drop Points in Self-Administered Address-Based Sampling Frame Surveys." Sociological Methodology 53.1, 2023, pp. 158-175.

[22] De, Shankkha, and Dipti Verma. "Deep Convolutional Transfer Learning approach for Bengali handwritten character recognition from document image." Science and Culture, 2023.

[23] Wolf, Thomas, et al. "Huggingface's transformers: State-of-the-art natural language processing." arXiv preprint arXiv:1910.03771, 2019.

[24] Jain, Shashank Mohan. "Tasks Using the Hugging Face Library." Introduction to Transformers for NLP: With the Hugging Face Library and Models to Solve Problems. Berkeley, CA: Apress, 2022, pp.69-136.

[25] Ushio, Asahi, and Jose Camacho-Collados. "T-NER: an all-round python library for transformer-based named entity recognition." arXiv preprint arXiv:2209.12616, 2022.

[26] Kayed, Mohammed, Sara Dakrory, and Abdelmaged Amin Ali. "Postal address extraction from the web: a comprehensive survey." Artificial Intelligence Review 55.2, 2022, pp.1085-1120.

Contribution of Individual Authors to the Creation of a Scientific Article (Ghost-writing Policy)
I am the only author and contributor of this research.

Sources of Funding for Research Presented in a Scientific Article or Scientific Article Itself
No funding was received for conducting this study.

Conflict of Interest
The author has no conflicts of interest to declare.