University of Sindh Journal of Information and Communication Technology (USJICT)
Volume 6, Issue 3, September 2022
ISSN-E: 2523-1235, ISSN-P: 2521-5582
Website: http://sujo.usindh.edu.pk/index.php/USJICT/

Named Entity Recognition for Urdu Language: The UNER System, A Hybrid Approach

[1]Saba Rani, [3]Hira Fatima Naqvi, [2]Fida Hussain Khoso, [2]Attia Agha, [1]Dil Nawaz Hakro, [1]Maryam Hameed
[1]Faculty of Engineering and Technology, University of Sindh, Jamshoro
[2]Dawood University of Engineering and Technology, Karachi
[3]Institute of Mathematics and Computer Sciences, University of Sindh, Jamshoro

Abstract: NER is a natural language processing technique that primarily classifies parts of parsed text into well-known named entities. In the domain of natural language processing, the recognition of name entities is used to classify nouns that appear in bulk text data and place these nouns into predefined groups, such as names of people, places, times, dates, organizations, etc. There is a lot of fragmented material and data on the Cyberspace, therefore scholars are working on several languages (i.e: Sindhi, English, etc.), by working on various approaches and techniques depending on their locations, to improve accessibility of filtered information for online users. The NER enhance the quality of NLP in applications including automated summarization, semantic web search, information extraction and retrieval machine translation and question answering, chatbots and others. This study designs an efficient framework to extract noun entities in Urdu using a hybrid approach. The UNER system not only extracts entities by searching through a list of names, but also extracts named entities by recognizing phrases in a given text. The UNER system is designed to recognize Urdu noun entities in pre-defined categories such as places, personal names, titled personal names, organizations, object names, trade names, abbreviations, dates and times, measurements, and text names in Urdu.

Keywords:— UNER, NLP NER, Urdu, Recognition, Entity, Named.

I. NATURAL LANGUAGE PROCESSING

In Human Computer Interaction (HCI), NLP has an important role in processing human languages where machines are processing human languages and scripts which pave the way for many of the researchers in the field of various languages [1][2]. At this time, rapid growth has been seen in development of NLP systems and applications to facilitate / access to relevant information in different languages, through appropriate environments.

Working on different languages is difficult due to the morphological structures and insufficient knowledge of words. Many of the languages have been researched and much of the literature is available but languages like Urdu and its neighboring languages pose more challenges for their processing. As in polymorphic language like Urdu, first word meaning explicitly refers to its interpretation and building sentences by joining letters [3]. The understanding of any languages is depending upon its phonic and phonologic understanding, grammar, practice and some elementary information.

A. Named enity Recognition

Named entity recognition is an important task in NLP, mainly to identify known named entities in the analyzed segmented text [5]. Today, NER systems are developed in multiple languages, which ease the users to access useful information about places, people, organizations, and other oun entities. This is done through different techniques like optical character recognition (OCR), which captures data from pictures and alters it into editable format to categorize specific noun entities. The automatic indexing of documented data contained in images called localization and detection [6]. NER is to identify and classify all nouns from any file, document or paragraph, for example: personal names can be masculine or feminine names, place names or city names, organizational names, duration, time, date, and various entities. NER research has identified various applications of the NER including chatbots, transliteration, question answers, information extraction and retrieval, machine learning and others. A representative architecture of the NER based system is present in Figure 1.

In Figure 1, an input is simple text, then the entity detector begins by patterns matching using a list search method based on elaboration rules established by grammar. Another method used is a machine learning approach that detects and uses features during list refinement, followed by the system generated list of NEs. Significant work is available on NER in Arabic [6], English [10], Sindh [29][31], Indonesian [33] and many other languages. There are multiple approaches used in NER systems. The Rule-based and hybrid approaches in NER use the spoken syntax and semantics of any language to recognize entities from text [8][9] [10].

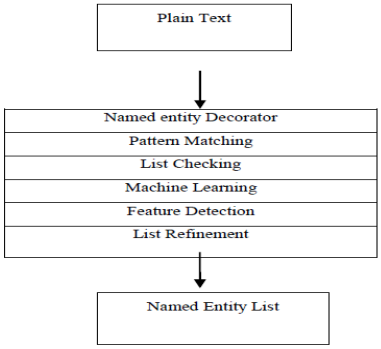

Figure 1: NER system Architecture [7]

B. *Urdu Language*

Language is important for communication; every region has their own mode of communication called language. The national language of Pakistan is Urdu and six Indian states declared Urdu as an official language [11]. There are over 300 million Urdu speakers in the world [12]. Urdu is the fourth largest language in the world. It contains 58 Urdu characters (38 as standard) in its alphabet, mostly a combination of the base character set and the derived character set. Urdu derived characters are formed by linking them to the base character set. As the conditions of their study [13], it is illustrated in Figure. 2. In 4.7% of the total world population, it is the fifth most vocalized language [14].

Urdu is a combination of Arabic and Persian scripts, resulting a mixed script of these two languages is used, called "Nastalik script".

There are twelve scripts used to write Urdu syntax [15]. Urdu consists of 38 elementary characters as shown in [16], known as "horofethahji" [17], as illustrated in Figure 2. Urdu script uses many font styles in which Nastalik is the most common calligraphic font for handwriting [18]. Urdu characters are divided into linking characters and unlinked characters. There are 12 unrelated (isolated) characters in Urdu. They are called unlinked characters because they are not linked to the previous character and cannot be linked to the next character as shown in Figure 4 [19].

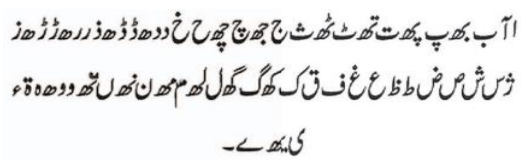

Figure 2: 58 character-set of Urdu [13][20]

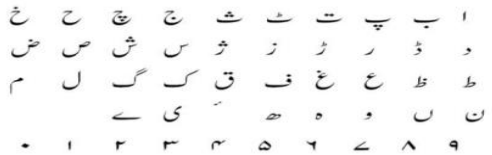

Figure 3: 38 character-set in Urdu [21]

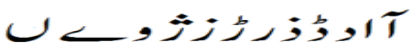

Figure 4: Non-linked Urdu Character [21]

II. RELATED WORK

Named entity recognition is an area of natural language that spans different languages, each with different semantic and grammatical structures. In NER, each method has a specific type of solution. Researchers had presented much of the work [34] on NER using different models, methods and techniques. Literature can be found on the basis of challenges imposed by the language. Various NER systems are available including Punjabi (Shahmukhi) [4], Sindhi [29][31], Arabic [30], Hindi [31], and most of the research is still continued [35] on having the focus of research on proposing novel approaches, but there is still a need for research on Asian languages like Urdu and a lot of work is required due to the challenging nature of the Urdu script.

III. RESEARCH METHODOLOGY

A. *Introduction to UNER System*

The proposed study is titled with relating to name as Urdu Named Entity Recognition (UNER). The national language of Pakistan is Urdu and it's also one of the largest spoken languages in the world [22]. Due to its challenging and complex nature of writing and speaking, Urdu is considered more difficult for its language processing or text processing. The complex structure, number of dots, sounds, shapes and context sensitivity of the Urdu is entirely different from other languages such as Chinese, English, Russian and Korean languages [23].

A system recognizing Urdu NE's helps user to insert editable text in editor, then the system uses hybrid approach to find particular nouns and categorizes these nouns into their named categories. UNER extract twelve NE's i.e: Person, Title Person, Place, Term, Abbreviation, Title Object, size, Organization, Brand, number, Date and Time. The UNER corpus contains predefined list NE's. The database (corpus) is based on online NE's collected from real time data and offline data including books, newspapers and other sources [25]. NER process in Urdu is complicated due to lack of resources such as Urdu dataset labeled NE [24]. Urdu Named Entity Recognition identifies individual NEs from given text, and Named Entity Classification uses the process of placing extracted nouns into specific groups as illustrated in Table 1.

Table 1: A Sentence Containing NE's

English Sentence	Urdu Sentence
Shifa and Muhammad Shayan both are going to Dubai.	شفاء اور محمد شایان دونوں دبئی جارہے ہیں

The example sentence in Table 1 contains three NEs, which are extracted by the system, as illustrated by Table 2.

Table 2: An Example of a NE's categorization

Proper Named Entity	Label (TAG)
محمد شایان شفاء	Person
دبئی	Location

IV. Proposed Research Framework

The system develops its framework on hybrid approach using rule based, list search-based and ML approaches. UNER able to identify 12 noun entities, including person, place, title person, organization, term, date and time value, title object name, numerical value, measurement value, denomination and brand name, etc.

Rule-based approaches include hand-crafted systems that rely on a list of rules in form of algorithms that allow the UNER to identify entities bearing a noun entity. The algorithm of rule-based and ML approaches framework structure is shown in Figure 5.

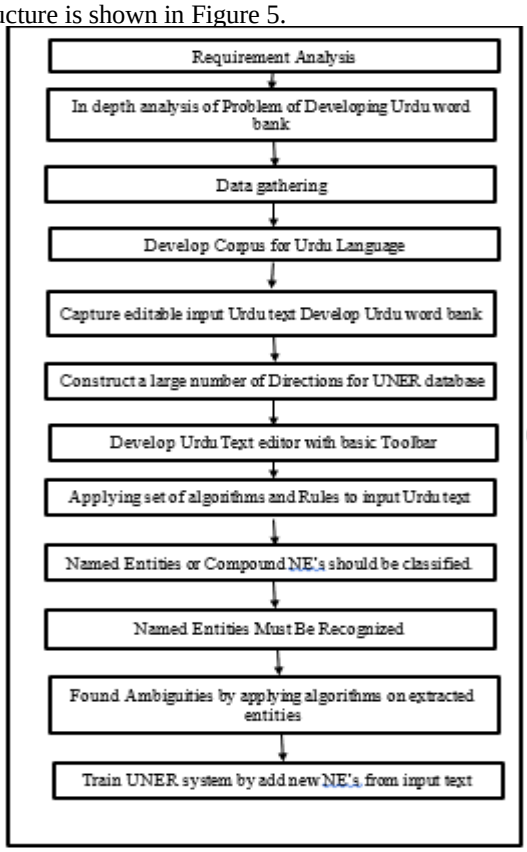

Figure 5: The proposed Hybrid approach and main framework for UNER system

A. Data Collection

This phase focused on compiling named entities for Urdu corpus such as names of person, NE locations, city and country name, NE name, abbreviation, title Person, Object Title, NE Brand, NE Organization, Date and Time, Measure, Number and Term Entities, as shown in Table 3.

B. Development of Urdu Corpus

The corpus design for twelve NE person, NE place, NE person title, NE organization, term, date and time value, object name title, numeric value, measurement value, abbreviation, name and brand name was made and labeled via multiple software.

C. Capture of editable Urdu text

UNER takes data from user in bulk form like no of paragraphs for analyzes it then applies rules to extract NEs.

D. Corpus Rules

UNER used rule-based approach by designing number of algorithms. These algorithms apply on input data and corpus as well.

E. Developing Editor

A custom-built editor has been created for Urdu NER with basic functions, such as: copy, select all, redo, undo, delete all and save to file so that the system can be tested.

F. Classification of Named Entities

This method classifies the text to be extracted NE with the corpus, and extracts the name entity, decomposes the inserted script into sentences, then additional crumbles into words by spaces, then splits data to check in from corpus records, such as names of people, occupations, dates, places, etc.

V. Results and Discussion

A. Designing and modeling UNER corpus

There are many databases on the web for Western languages, which can be used in many NER applications, but few work had done for Urdu, and due to the nonexistence of large publicly available datasets for the Urdu corpus, UNER corpus was created which consisted on twelve groups. UNER corpus contains 20,115 NEs, with 4,841 person NEs, 8,278 location NEs, 199 organization NEs, and others including marks, abbreviations, dates and times, measures, titled persons, terms and titled object entities. The sample corpus data shown in Table 3.

The data for UNER has been collected from various sources including newspapers, news websites and other online sources. The new entities have been created in UNER database. These network elements are stored in the database as different categories. Some of the data collected in the Urdu Corpus are shown in Table 4.

B. UNER Algorithms

The algorithm proposed in this study takes editable text and segments into sentences and the words by means of spaces. The segmented words are matched with the existing entries available and stored in the corpus. The extraction named as NE's recognition, is a knowledge extraction method to classify entities from the given data samples. Various entities such as numbers, places, person names and the spots are the elements of recognition. A hybrid approach

is proposed for coping various problems containing three algorithms. The main algorithm is shown in Fig: 5

Table 3: corpus Tables and its Stored Data

Categories Name	Data stored Example
PersonName	شفاء، محمد شایان
Location Name	حیدرآباد، پاکستان
Title Person Name	علامہ، مسٹر
Designation Name	کمشنر, اسسٹنٹ پروفیسر
Organization Name	اسٹیٹ بینک آف پاکستان ، سندھ بینک
Title Object Name	بار ائٹ لا
Brand Name	ایچ پی، ایپل
Term	شفاء، ممتاز
Number	10245, 100
Measure	انچ, آٹھ سال
Date & Time	مارچ, جون2021ء
Abbreviation	آئی آئی سی ٹی، اي

Table 4: Collected data samples

Single Person NE's	Location NE's
آرزو	حیدرآباد
آئشہ	سہون
شفاء	کوٹری
آذان	ٹنڈو محمد
آسفہ	دادو
آریان	اسلامکوٹ
آرنلڈ	خانپور
آریز	نوشہروفیر
آزان	نصیرآباد
آتقا	عمرکوٹ
آتش	لاڑکانہ

In addition, the problems pertaining to free word order, and entities within entities are solved by the first algorithm and such types of problems are called uncertainty algorithms. These algorithms illustrate the mixed solution model of the UNER system. The proposed algorithms successfully identify and recognize correct named entities of Urdu when a text of Urdu language is given as input.

A paragraph of the Urdu text is given as input to the proposed system. The system segments the given text into words and the marks are identified for the start and the end of the sentence in the tex. The words extracted are then compared with the grouped or categorized data or entities in the database or the corpus of Urdu entities. If the words are matched then all of the matched entities are stored in the table. The table contains the entries such as counting of the entities, matched entities and the frequency of the entities. The system prints the matched entities.

C. UNER Custom-built application

Urdu named entity recognition custom built application consists of multiple interfaces, the first one is starting interface. The Application's Main View at the Start The key starting graphical user interface of the application "Urdu Named Entity Recognition (UNER)" is shown in Figure 6.

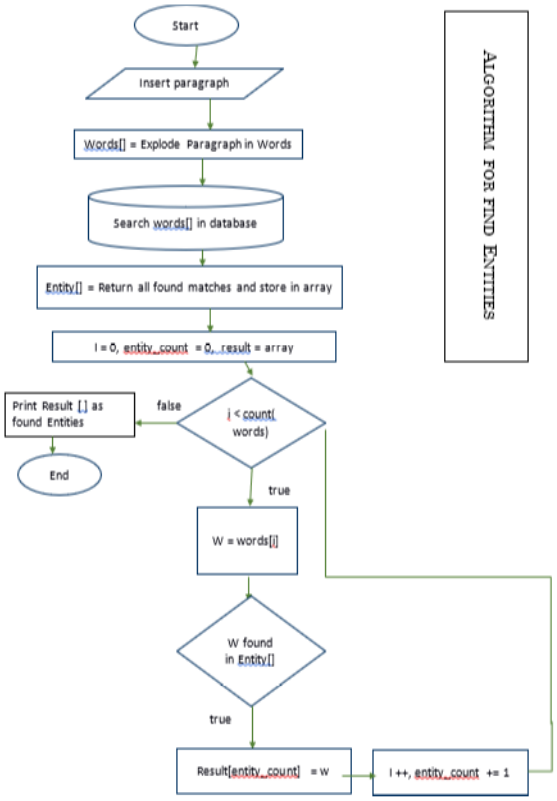

Figure 5: Main UNER Algorithm for NE's Extraction

Figure 6: Main Interface of Application (UNER)

This panel consists of a list of predefined Urdu sentences box and basic toolbar containing six buttons for UNER text editor are shown in figure 7, these buttons are for 'copy, undo, redo, select all, clear and save' actions. The copy function use for copy text from editor.

Figure 7: UNER interface with buttons

The rest of buttons like undo is used for removing last actions in editor and redo button used for recover last undo actions. The select all button selects all text from UNER editor, and clear all button used to clear whole UNER editor. Save button saves all UNER editor text in notepad.

Figure 8: Main Editor (UNER)

In Figure 8, the UNER text editor is shown, which use to take Urdu text as input or select sentences from example sentences. User can select sentences by double clicking on any of one at a time. These sentences are for testing purpose. The user also input with an Urdu keyboard by clicking on a keyboard button icon as shown in Figure 9. After entering Urdu sentence or a paragraph user can click on extract button for extracting NE's from given text. The editor has basic functions toolbar performing copy, select all, save, cut, undo and redo actions on inserted Urdu text.

Figure 9: UNER Urdu Keyboard

D. Results with UNER application

The extracted named entities interface is shown in Figure 10, where UNER extracts NEs from provided Urdu text. Entities are extracted from the input and classified using 12 predefined classes such as person name, place, organization, term, name, titled person, titled object, brand, measure, number, date and time and abbreviation. The UNER system extracts entities in tabular form. The extracted entities are shown while working in Figure 10.

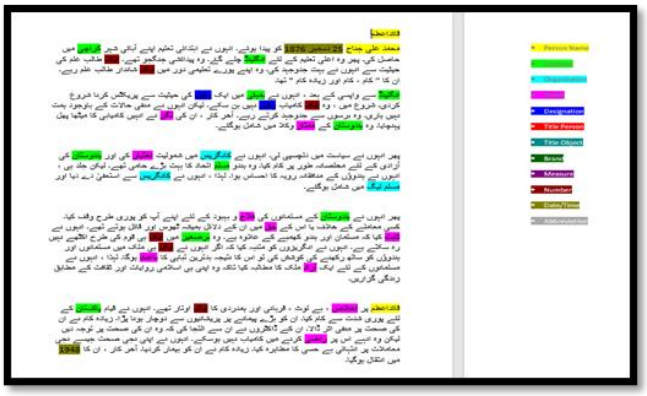

Figure 10: UNER Working

E. Testing and Evaluation

F measure, Precision and the Recall rate (F,P and R) are typically used as quality measures for any NER system. The quality parameters apply to estimate the outcomes of this NER technique for UNER. The accuracy by precision is divided by the number of correct NE's defines the total number of NEs [27]. The recall is calculating the number of correct NEs, found by NER system over the sum of NEs in a text that has been useful in testing purpose.

The F-measure indicates precision and recall in the expression. The accuracy depends on the network elements extracted by the system. Table 5 describes the outcomes of the proposed NE recognition system.

In precision, the total NEs properly identified by the system divided by the sum of NEs, as shown in the given equation

The proposed study tested the system using a class called "(Quaid-e-Azam, قائداعظم)" and the results are shown in Figure 10. The UNER system is tested with 25 documents. A vast amount of inserted data is formed depending on the network elements. The results are shown in Table 5 and Figure 11 as well. The results are calculated for twelve NE's such as Person NE, Title Person NE, Place NE, Organization NE, Date and Time Value, Term, Title Subject, Measure, Numeric Value, Abbreviation, Name, and Business Name.

Table 5: Results of NER System for Urdu Language

Entity	Total NE's in Document	Total Ne's Given by System	Correct NE's	Precision (P)	Recall (R)	F1-Score
Person	130	120	119	0.99	0.91	94.73
Location	110	109	105	0.96	0.95	95.28
Title Person	30	29	27	0.93	0.9	91.25
Organization	150	145	144	0.99	0.96	97.43
Terms	20	18	17	0.94	0.85	88.82
Title Object	10	9	9	1	0.9	94.73
Brand	15	15	15	1	1	100
Abbreviation	4	4	4	1	1	100
Date & Time	30	28	28	1	0.93	96.37
Number	5	5	5	1	1	100
Measurement	22	22	20	0.	0.90	90
Designation	2	2	2	1	1	100
Total accurateness of purposed NER system for Urdu language is						93.75%

Table 5 and Figure 11 illustrate Performance measure in term of Precision Recall and F-Measure, UNER achieved 90-93% accuracy, depending on the NEs stored in the database.

VI. Conclusion and Future Work

A. Conclusion

Artificial intelligence focuses on automating intelligent behavior, making machines smarter and functioning like humans. The relation and interaction between a human and the machine are the core concept of the NLP [2]. The main determination of NE recognition is to find out nouns in the text and classify into classes from a given text or the document or the part of the text [27]. Today, NER systems are developed in different languages and any group of people can easily access data about places, people, organizations and other entities [6]. NE recognition is an important task in NLP, primarily to identify known named entities in the analyzed segment (text) [1]. Language plays an important role in communication. Urdu is the official language of Pakistan [28]. There are over 300 million Urdu speakers in the world [12]. Urdu has become the official language of six Indian states [21]. More than a dozen scripts are used in Urdu, of which Nastaleeq is the most famous calligraphy used to write Urdu [18].

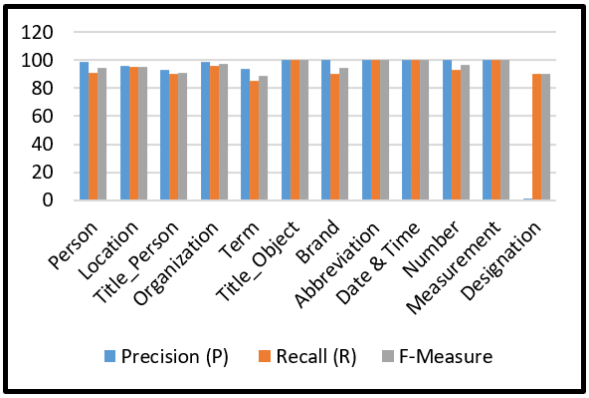

Figure 11: Results of UNER System

The proposed system is capable to identify major entities in Urdu texts. This study was an attempt to develop a system based on Urdu grammar rules and correct syntax. The system will help users to identify and group the identified entities and will help to understand the given text. In this way, with the help of this system a given text (unstructured text) can be understood by the human being easily. The current study recognizes successfully twelve entities of Urdu language from a given text and matches with the given corpus with an aggregate accuracy of 93.75%. The system has the capability to recognize the ambiguous words to some extent. The system is also capable of recognizing complex entities

B. Future Work

The system performance can be improved by finetuning the algorithms used in the study or the further the corpus can be tested by implementing other algorithms. The size of the corpus can be increased by adding more entities and the text entries or Urdu language. A mechanism of deleting conflicting entries can be established so that the system performance can be increased. By implementing efficient algorithms for handling compound and ambiguous entities, the system performance can be enhanced.

.

REFERENCES

[1] J. R. Finkel and C. D. Manning, "Joint parsing and named entity recognition," in Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, 2009.

[2] Y. Kaur and E. R. Kaur, "Named Entity Recognition (NER) system for Hindi language using combination of rule based approach and list look up approach," Int. J. Sci. Res. Manag.(IJSRM), vol. 3, pp. 2300-2306, 2015.

[3] S. Vadera and F. Meziane, "From English to Formal Specifications," Comput. J., vol. 37, pp. 753-763, 09 1994

[4] M. T. Ahmad, M. K. Malik, K. Shahzad, F. Aslam, A. Iqbal, Z. Nawaz and F. Bukhari, "Named entity recognition and classification for Punjabi Shahmukhi," ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), vol. 19, pp. 1-13, 2020.

[5] J. R. Finkel and C. D. Manning, "Joint parsing and named entity recognition," in Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, 2009.

[6] K. Jung, K. I. Kim and A. K. Jain, "Text information extraction in images and video: a survey," Pattern recognition, vol. 37, pp. 977-997, 2004.

[7] V. Gupta and G. S. Lehal, "Named entity recognition for punjabi language text summarization," International journal of computer applications, vol. 33, pp. 28-32, 2011.

[8] A. Goyal, "Named entity recognition for south asian languages," in Proceedings of the IJCNLP-08 Workshop on Named Entity Recognition for South and South East Asian Languages, 2008.

[9] K. Gali, H. Surana, A. Vaidya, P. M. Shishtla and D. M. Sharma, "Aggregating machine learning and rule based heuristics for named entity recognition," in Proceedings of the IJCNLP-08 Workshop on Named Entity Recognition for South and South East Asian Languages, 2008.

[10] P. Kumar and V. R. Kiran, "A hybrid named entity recognition system for south Asian languages," in proceedings of the IJCNLP-08 Workshop on NER for South and South East Asian Languages, 2008.

[11] Wikipediacontributors.2018.Urdu.(2018).https://en.wikipedia.org/w/index.php?title=Urdu&oldid=844110134[Online;accessed10-June-2018]

[12] K. Riaz, "Baseline for Urdu IR evaluation," in Proceedings of the 2nd ACM workshop on Improving non english web searching, 2008

[13] K. Khan, R. U. Khan, A. Alkhalifah and N. Ahmad, "Urdu text classification using decision trees," in 2015 12th International Conference on High-capacity Optical Networks and Enabling/Emerging Technologies (HONET), 2015.

[14] A. Wahab and S. ul Haque and Najam, "Optical Character Recognition System for Urdu.," Journal of Independent Studies and Research, vol. 8, 2010.

[15] K. Khan, R. U. Khan, A. Alkhalifah and N. Ahmad, "Urdu text classification using decision trees," in 2015 12th International Conference on High-capacity Optical Networks and Enabling/Emerging Technologies (HONET), 2015.

[16] A. Rana and G. S. Lehal, "Offline Urdu OCR using ligature based segmentation for Nastaliq Script," Indian Journal of Science and Technology, vol. 8, pp. 1-9, 2015.

[17] Z. Jan, M. Shabir, M. A. Khan, A. Ali and M. Muzammal, "Online Urdu handwriting recognition system using geometric invariant features," The Nucleus, vol. 53, pp. 89-98, 2016.

[18] M. I. Razzak, S. A. Hussain and M. Sher, "Numeral recognition for Urdu script in unconstrained environment," in 2009 International Conference on Emerging Technologies, 2009.

[19] A. Rana and G. S. Lehal, "Offline Urdu OCR using ligature based segmentation for Nastaliq Script," Indian Journal of Science and Technology, vol. 8, pp. 1-9, 2015

[20] K. Khan, R. U. Khan, A. Alkhalifah and N. Ahmad, "Urdu text classification using decision trees," in 2015 12th International Conference on High-capacity Optical Networks and Enabling/Emerging Technologies (HONET), 2015.

[21] A. Rana and G. S. Lehal, "Offline Urdu OCR using ligature based segmentation for Nastaliq Script," Indian Journal of Science and Technology, vol. 8, pp. 1-9, 2015.

[22] W. Khan, A. Daud, J. A. Nasir and T. Amjad, "Named entity dataset for urdu named entity recognition task," Organization, vol. 48, p. 282, 2016.

[23] T. L. Packer, J. F. Lutes, A. P. Stewart, D. W. Embley, E. K. Ringger, K. D. Seppi and L. S. Jensen, "Extracting person names from diverse and noisy OCR text," in Proceedings of the fourth workshop on Analytics for noisy unstructured text data, 2010.

[24] W. Khan, A. Daud, J. A. Nasir and T. Amjad, "Named entity dataset for urdu named entity recognition task," Organization, vol. 48, p. 282, 2016.

[25] A. Wahab and S. ul Haque and Najam, "Optical Character Recognition System for Urdu.," Journal of Independent Studies and Research, vol. 8, 2010.

[26] B. MySQL, MySQL, 2001.

[27] S. Srivastava, M. Sanglikar and D. C. Kothari, "Named entity recognition system for Hindi language: a hybrid approach," International Journal of Computational Linguistics (IJCL), vol. 2, pp. 10-23, 2011.

[28] A. Daud, W. Khan and D. Che, "Urdu language processing: a survey," Artificial Intelligence Review, vol. 47, pp. 279-311, 2017.

[29] D. Nawaz, A. W. A. N. SA, Z. A. BHUTTO, M. Memon and M. Hameed, "Handling ambiguities in Sindhi named entity recognition (SNER)," Sindh University Research Journal-SURJ (Science Series), vol. 49, pp. 513-516, 2017.

[30] Y. Benajiba and P. Rosso, "Arabic named entity recognition using conditional random fields," in Proc. of Workshop on HLT \& NLP within the Arabic World, LREC, 2008.

[31] Y. Kaur and E. R. Kaur, "Named Entity Recognition (NER) system for Hindi language using combination of rule based approach and list look up approach," Int. J. Sci. Res. Manag.(IJSRM), vol. 3, pp. 2300-2306, 2015.

[32] D. N. Hakro, Hakro, M. A., & Lashari, I. A. (2017). Sindhi Named Entity Recognition (SNER). The Government-Annual Research Journal of Political Science., 5(5). 143-154.

[33] Budi, I., & Suryono, R. R. (2023). Application of named entity recognition method for Indonesian datasets: a review. Bulletin of Electrical Engineering and Informatics, 12(2), 969-978.

[34] Hamdi, A., Pontes, E. L., Sidere, N., Coustaty, M., & Doucet, A. (2023). In-depth analysis of the impact of OCR errors on named entity recognition and linking. Natural Language Engineering, 29(2), 425-448.

[35] Chen, Y., Wu, C., Qi, T., Yuan, Z., Zhang, Y., Yang, S., ... & Huang, Y. (2023). Semi-supervised named entity recognition in multi-level contexts. Neurocomputing, 520, 194-204.