Named Entity Recognition of Kumauni Language using Machine Learning (ML)

Vinay Pant[1*], Dr. Rupak Sharma[2], Dr. Shakti Kundu[3]

[1]Research Scholar Department of Computer Science and Engineering, SRMIST (Deemed to be University) Delhi NCR Campus, Modinagar, India, Email: vp5570@srmist.edu.in
[2]Associate Professor, Department of Computer Application, SRMIST (Deemed to be University) Delhi NCR Campus, Modinagar, India, Email: rupaks@srmist.edu.in
[3]Associate Professor, School of Engineering and Technology, Computer Science Engineering, BML Munjal University(BMU),Gurugram, Haryana, India, Email: shaktikundu@gmail.com
*Corresponding Author

Received: 13.07.2024 Revised: 16.08.2024 Accepted: 10.09.2024

ABSTRACT
Communication between humans is impossible without the use of language. Natural Language Processing (NLP) is a technique that is used so that computers can comprehend various natural languages. Named Entity Recognition (NER) is a subtask of information extraction that aims to discover and categorize the components in given text into pre-defined categories. NER is an abbreviation for the phrase "named entity recognition". Machine translation, question-answering systems, and automatic summarization are examples of the types of NLP tasks that can benefit from NER. The purpose of this research is to investigate whether or if it is possible to create a chatbot that converses in the Kumaon language, as well as any potential difficulties that could arise in the process. In addition to this, the authors provide an in-depth examination of Kumaoni as well as a mapping of the language into other languages to make its use in industrial processing more accessible. In this study author utilize a previously researched based on the named entity recognition of languages from the databases such as Scopus, web of science, IEEE, Google Scholar, Cite SeerX, Cross Ref etc. in their study. The research activity possesses the capacity to fundamentally transform the field of linguistic evaluation in the Kumaon region by employing sophisticated machine learning (ML) techniques to explore the complex domain of NER (named entity recognition) for the Kumauni language.

Keywords: Natural Language Processing (NLP), Machine Learning, Named Entity Recognition (NER), Kumauni Language

1. INTRODUCTION

The Kumaoni language is one of the 325 official languages of India. It is spoken mostly in the Uttarakhand districts of Almora, Nainital, Pithoragarh, Bageshwar, and Udham Singh Nagar. There are also few native speakers in the neighboring countries of Himachal Pradesh and Nepal. Kumaoni could also be seen as one of the sub-groups of the Pahari language family. In 1998, it was expected that some 2,360,000 people understood the Kumaoni language. Several individuals know this Kumani by one of its other names including Kamaoni, Kumaoni, Kumau, Kumwani, Kumgoni, Kumman, and Kunayaoni are only a few examples. It is a member of the large Indo-Aryan (IA) language family [1]. Several different languages are spoken in the Kumaon area. Kumaoni dialects cannot be reliably categorized according to any one system. As shown in Figure 1, the IA language family is represented graphically using tree charts.

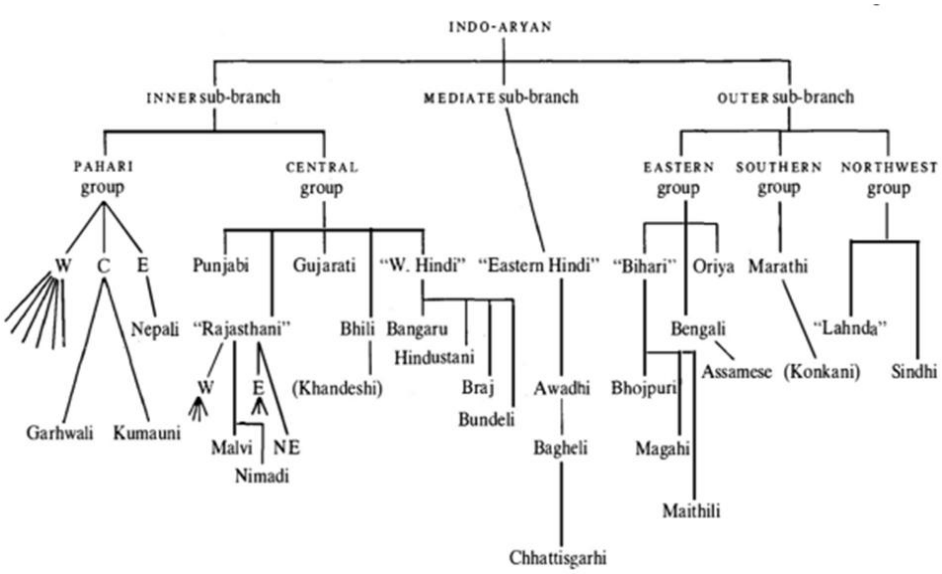

Figure 1: The Indo-Aryan (IA) language family [2]

S Kali (or Central) Kumaoni is communicated mostly in Almora and the northern part of Nainital. Pithoragarh is home to the Kumaoni speakers in the region's far northeast. The region of Kumaon in the south-east Nainital. The area west of Almora and Nainital is where the Western Kumaoni language is spoken.

More specifically languages are given in below [3]:

- Johari of the Malla and Talla Johar (Milam, Munsiyari)
- Danpuriya of Danpur (Bageshwar, Kapkot)
- Bhabhri of Haldwani and Ramnagar
- Askoti of Askot
- Kumaiyya of Champawat
- Pachhai of Pali-Pachhhau (Ranikhet, Dwarahat)
- Khasparjiya of Almora
- Gangoli of Ganai-Gangoli (Kanda, Berinag, Gangolihat)
- Sirali of Sirakot (Didihat)
- Soriyali of Sor Valley (Pithoragarh)
- Rhau-Chaubyansi, (Nainital)
- Phaldakotiya of Phaldkot

It has been stated that there are some Kumaoni speakers living in western Nepal [4].

Natural Language Processing (NLP) is a subfield of AI that combines linguistics and computing to aid in human-computer interaction (HCI). The primary goals of NLP are Natural Language Understanding (NLU) and Natural Language Generation (NLG). An example of an NLP application is Named Entity Recognition (NER). It is also a subtask of many NLP applications, such as information retrieval, machine translation, question answering, extraction, and text summarization [5]. Take this statement as an example: On August 7th, 2014, Ramu enrolled at Amrita in Coimbatore to pursue his master's degree. Ramu represents a person entity, Amrita represents an organization entity, Coimbatore represents a location entity, and August 7th, 2014, represents a date entity in this context. A significant amount of effort has been put into NER for Indo-European languages like English. Capitalization is used to emphasize names in several languages, including English. It is not possible to capitalize words in most Indian languages, including Telugu, Hindi, Tamil, Kanada, Malayalam, Urdu, and Bengali. Upper and lower cases are not differentiated in Indian scripts. Accordingly, NER development for Indian languages is quite challenging [6].

Machine Learning (ML) approaches, including both supervised and unsupervised methods, are used in the rule-based approach used by NER systems. Decision Trees (DE), Support Vector Machines (SVM), Maximum Entropy Markov Models (MEMM), Hidden Markov Models (HMM), Conditional Random Fields (CRF), and other ML algorithms are extensively utilized [7]. Traditional techniques for Indian languages include rule-based approaches that rely on a gazetteer, dictionary, and named entity patterns. Several domain-specific applications make use of NER, from genetics and tourism to others. NER in the tourist industry is primarily concerned with localization and the identification of physical features, such as buildings, roads, rivers, lakes, and oceans. This category also includes parks, museums, and other

religious and cultural buildings, as well as monuments. Deoxyribonucleic acid (DNA), proteins, and gene names are only some of the things that the genetic NER system must be able to recognize [8].

1.1. Kumaoni Language

The Kumaon people have their own unique language, which is called Kumaon. The people that live in the Kumaon area of Uttarakhand state are known as Kumaouni, and the language they speak is also called Kumaouni. Kumaouni has several antonyms, including Kunayaoni, Kumman, Kumgoni, Kumawani, Kamaoni, Kumau, etc. Pithoragarh, Nainital, Bageshwar, Almora, Champawat, and Udam Singh Nagar are the six districts that comprise Uttarakhand's Kumaon division. The Himalayan Mountains to the north and Nepal to the east are the geographical boundaries of this area. The locals in a few regions of Nepal and Himachal Pradesh speak a language called Kumaouni. Kumaoni is one of India's officially recognized languages. It is also considered a branch of the Pahari language family.

Moreover, it employs the Devnagari script in its writings. The Garhwali language, which is closely related to this one, also contains several regional dialects. This language is also extensively spoken in several of Uttarakhand's districts. Therefore, it has been established that speakers of Hindi and Nepali who are fluent in the Kumaoni idiom are also fluent in those languages. The government of India has recognized these idioms as its own. It has been observed that the use of the Kumaoni language is declining, and United Nations Educational, Scientific and cultural organization (UNESCO) has classified it as a "language of concern," calling for more safeguards and revitalizing activities. Additionally, the Kumaoni language is written using the same Devanagari script as the spoken language [9].

The Kumaoni tongue is very important in the media industry. A wide variety of media, from movies and plays to folk music and FM radio, are based on this language. Kumaoni language is used in several films, including Megha Aa, Teri Saun, Aapun Biran, and Madhuli. These films have been huge successes and award-winners. Additionally, the Kumaoni language is used in a variety of plays performed at local theaters [10]. The Ramleela play was the most anticipated and important play of the season. Subsequently, Dinesh Pandey and Mohan Upreti made significant contributions to elevating the quality and promoting this kind of narration from the Kumaon region. The Kumaoni language went on to be used for the transmission of rap and folk music. Many plays, tales, and even musical performances were presented in the Kumaoni language as part of the entertainment, religious, cultural, and traditional segments. The Kumaoni language and its rich cultural history have also greatly benefited from the efforts of many internationally acclaimed singers and musicians. In addition, the dhol, turri, damoun, dholki, thali, daur, masakbhaja, and bankora are played during the narration and singing of Kumaoni songs.

In addition, several radio stations are broadcasting Kumaoni-language music, tales, and a wide variety of other programs. Multiple radio stations regularly air programming from all these entertainment categories. Akashwani Luck has launched a brand-new radio show called Utterayana. This show is targeted at Chinese border areas. In the next year, 2010, a plethora of Kumaoni-language air programs covering a wide range of topics, from entertainment to the environment to culture to agriculture, were produced. As a result of widespread participation, radio broadcasts and stations quickly gained clout and appeal among Kumoans [11].

1.2. Named Entity Recognition

Authors need to reevaluate whether NER is a solved issue or not to establish assessment processes that are appropriate for the demands of the time and that can accurately gauge the present state of the art in NER [15]. As opposed to the hand-crafted rule-based algorithms used by early systems, supervised algorithms like HMM, DE, Maximum Entropy Models (MEM), Conditional Random Field (CRF), and Support Vector Machine (SVM) are used extensively in current systems. In supervised algorithms, a classifier is constructed by a general inference technique (more often known as learning) [16]. The next step involves manually processing many texts by locating and categorizing named items so that an algorithm could derive their defining features [17]. In the last phase of classification, fresh texts are fed into the classifier in the hopes that it will identify items within them. Characteristic qualities specified for algorithmic consumption are used by inference algorithms; a Boolean variable with the values true if a word is capitalized and false otherwise is an example of a feature. Every word is often represented by one or more Boolean, nominal, and numeric values in the feature vector format, which is an abstraction over text [18].

NER, also known as NER, is a step in the Information Extraction (IE) process that involves locating and categorizing certain kinds of information components that are referred to as Named Entities (NE). Opinion Mining, Ontology Population, and Semantic Annotation are just a few of the many other important sub- fields in Information Management that depend on this structure. NE was coined during the sixth Message Understanding Conference (MUC) when the significance of semantically identifying

individuals, groups, geographic locations, and numerical expressions like time and quantity was first recognized. Most modern NER tools continue to regard these sorts of NE to have originated in MUC, but with some significant modifications. However, the significance of NE remains elusive. This is a topic that has surfaced in earlier writings but has not yet been thoroughly examined: This is because "the idea of NE originated in an atmosphere of NLP applications and is far from being linguistically unambiguous and stable [12]. Strangely, this contradicts the widespread belief that NER is a solved job with success rates much over 95% [13]. The authors suggest that NER is not a solved issue and demonstrate how the absence of consensus on the meaning of NE has serious consequences for the development and assessment of NER technologies. There is no easy solution to this issue in the current state of NER assessment forums since they focus on such a wide variety of various activities [14].

Authors need to reevaluate whether NER is a solved issue or not to establish assessment processes that are appropriate for the demands of the time and that can accurately gauge the present state of the art in NER [15]. As opposed to the hand-crafted rule-based algorithms used by early systems, supervised algorithms like Hidden Markov model (HMM), Decision Tree (DT), Maximum Entropy Models (MEM), Conditional Random Field (CRF), and Support vector machine (SVM) are used extensively in current systems. In supervised algorithms, a classifier is constructed by a general inference technique (more often known as learning) [16]. The next step involves manually processing many texts by locating and categorizing named items so that an algorithm could derive their defining features [17]. In the last phase of classification, fresh texts are fed into the classifier in the hopes that it will identify items within them. Characteristic qualities specified for algorithmic consumption are used by inference algorithms; a Boolean variable with the values true if a word is capitalized and false otherwise is an example of a feature. Every word is often represented by one or more Boolean, nominal, and numeric values in the feature vector format, which is an abstraction over text [18].

NER is the procedure of assigning an entity status to anything that has a name or label of its own, such as a place, a company, or a person. NE could be anything from a specific place to a specific date or even a specific amount of money, and the NER model could be modified to accommodate user-defined NE. The following text has been highlighted to indicate when names are used: Symbiosis Society Pune [LOCATION] was founded in 2008 [TIME] by SB Mujumdar [PERSON], who also built the Symbiosis Institute of Technology [ORGANIZATION]. There are four distinct nouns in this statement: one for ORGANIZATION, one for TIME, one for PERSON, and one for LOCATION. As can be seen in Figure 2, NER is a sequence-tagging job in which the contextual meaning of words is fetched via the use of word embeddings.

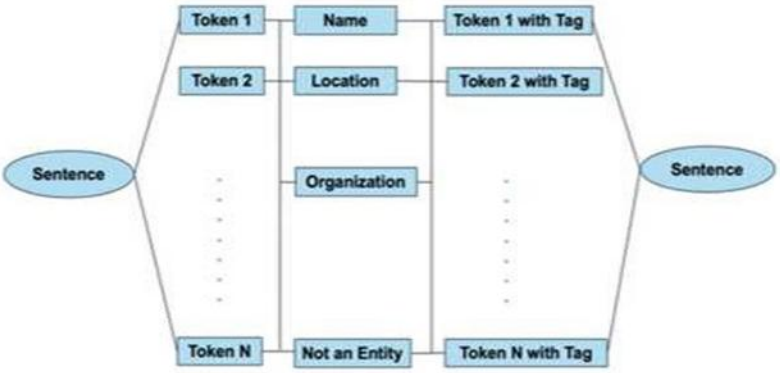

Figure 2: Architecture of NER [19]

A wide variety of things could be recognized, identified, and classified by more advanced NER models. NER technology has a wide variety of applications, including news content classification, content recommendation engine power, customer service, an effective search algorithm, etc. There is a distinct need to recognize words in these contexts. Words like "assault," "crime," "politics," and so on are examples of those that must be recognized when categorizing material for news providers. Words like "clothing," "shoes," "color," "size," "etc." are crucial for the fashion website's database search engine. A new NER model could be developed for each distinct need, or the current model can be modified to locate a new set of target words. Numerous tools and libraries for NER have been developed in both Python and Java. In this study, the author discussed and analyzed the performance of several NLP tools for NER models, including Apache OpenNLP, Python's SpaCy, and TensorFlow. Time, accuracy, and performance quality are among the many criteria used to evaluate different NER tools. NER could make use of a selection of learning strategies, including but not limited to semi-supervised learning, supervised learning,

and unsupervised learning [20]. Support vector machines (SVM), hidden Markov models (HMM), decision trees (DT), conditional random fields (CRF), etc. are only some of the supervised learning tools available. 'Bootstrapping' is used in semi-supervised learning, whereas 'clustering' is used in unsupervised learning. NER could employ a variety of NLP libraries, and its models could be produced using these tools. These libraries are designed with a subset of languages in mind, creating a domain for recognizing name entities. There are two different programming languages used to create the Stanford NE: Java for SpaCy and TensorFlow in Python. All these libraries have their own built-in NER models for certain things like people, businesses, locations, etc. Different conditions apply to problem statements and applications in the actual world. It is necessary to develop or modify NER models because various applications need entities of varying sorts to be detected. Not only accuracy is a consideration when comparing NER models, but also prediction time, model size, and training simplicity are important factors [21].

1.3. Application of NER

There are several applications where NER is required, such as:

- **Machine Translation**

There are few industries in India that are expanding as quickly as India's information technology (IT) market. There is a high need for document translation services in a country like India, where the number of spoken languages is high. Most state governments operate in various regional languages, although the Union Government's official papers and reports are bilingual (Hindi and English). To ensure effective communication, it is necessary to translate these reports and papers into the appropriate regional languages. Proper nouns are not to be translated but rather transliterated when using machine translation. As a result, Named Entity Identification (NEI) presents a significant difficulty for the Machine Translation (MT) community [22].

- **Intelligent Document Access**

Most of the material on the Internet is in English, which is utilized by just around 10% of the Indian population. Search engines are often used to get access to this data, but when it comes to queries written in Indian languages, the search engines' capabilities fall short. This indicates that most Indians are unable to access the information that is available on the Internet, which contributes to the development of a digital divide. Many Indians have a passing familiarity with the English language, but they have too limited a vocabulary to be effective search engine users. It would be crucial for query processing to be able to identify the specified entity in the document. Higher-quality results from these search engines are possible using NER

- **Cross-Language Information Retrieval (CLIR)**

CLIR refers to a method for retrieving documents when the language of the documents is different from the language of the queries made by a human user. CLIR could be differentiated from monolingual information retrieval by the user's capacity to submit a query in one language and get a document in a different language.

- **Summarization**

The ever-expanding World Wide Web is essentially a vast data warehouse containing a wide range of information on real-world things including people, places, and organizations. The query logs of all the web and Altavista search engines revealed that between 11 and 17 percent of all searches included a person's name together with other phrases, while another 4 percent consisted of only the person's name. Conventional search engines just provide a list of URLs that include the query's name with no attempt at summarizing. It's inefficient and time-consuming for users to go through websites and manually summarize the content. Automatic summarizing of the information available on the Web about entities like people, places, and organizations is required to enhance search results for entity searches [23].

- **Question Answering System**

A NER serves two purposes in the context of question answering: first, it could be used to exclude strings (like phrases) that are unlikely to contain the answer, and second, it can be used to identify candidates for the most precise responses. In the process of answering the inquiry, the intended response type is identified and then mapped to a database of NE classifications. A text fragment is analyzed using NER to identify the various NE types contained inside. A text is disregarded or harshly punished if it lacks any NE whose type matches the type of the intended response.

1.4. Approaches of NER

There are several methods for detecting NER. These methods are:

- **Rule-based or Linguistic approach**

NER detection is performed manually by linguists using written rules in rule-based or linguistic methods. These are regulations that are unique to the language. The most notable rule-based NER systems are a)

lexicalized grammar; b) listings in a gazetteer. c) Word-trigger list [24]. NER refers to the technique of recognizing certain words or phrases inside a text and assigning them to predetermined categories. This method is comprised of two primary subsidiary steps. first, determining which words are proper nouns based on the order in which they appear in the text; second, putting proper nouns into the categories of a rule-based system that has been established for them. It's also known as a linguistic method or hand-crafted rules. In this method, the researchers write the rules for the system and any language they are interested in manually. A parse tree or other abstract representation of the source text is generated by rule-based systems during the parsing process [25].

- **Machine Learning Approach**

The goal of the NER method used in a NL-based NER system is to transform the identification task into a classification task, which is then solved using a statistical model for classifying the data. ML techniques are often known as corpus-based approaches. Systems use statistical models and ML algorithms to examine text for relationships and trends. Using ML techniques, the systems detect and categorize nouns into classes such as people, places, times, and so on. NER could be accomplished using a semi-supervised, supervised, or unsupervised ML model [26]. Supervised learning creates a model using only labeled data. The goal of semi-supervised learning is to facilitate learning using both labeled data and relevant information from unlabeled data. Unsupervised learning is a kind of machine learning that is intended to be able to learn with no labeled data or with very little labeled data [27].

- **Hybrid Approach**

The hybrid approach to stammering involves the use of more than one treatment method. In stemming, for instance, there are methods like lookup tables and suffix/prefix stripping. If the term is in the look-up table (which preserves the inflected word and its root forms), the root form is returned; otherwise, the inflected word is subjected to stripping rules [28].

2. RESEARCH METHODOLOGY

The preferred reporting criteria for meta-analyses and systematic reviews (PRISMA) statements serve as the foundation for the methodology used in the current systematic review (SR). In the area of evidence-based research, PRISMA offers a standardized framework for performing and disclosing systematic reviews and a meta-analysis. PRISMA was developed to standardize the reporting of studies that evaluate the efficacy of treatments, but it can be altered to record systematic reviews that concentrate on other issues, like research methodology, the creation of theoretical frameworks, or hypothesis development.

2.1. Search Strategy

A Systematic Literature Review (SLR) focuses on the prospects of the NER of the Kumauni Language using ML and an appropriate case-base structure for the collected cases to enable their effective retrieval as part of this endeavor. The study of "Named Entity Recognition of Kumauni Language using Machine Learning (ML)" was compiled using the key academic literature databases such as Web of Science, Scopus, Google Search, Cross Ref, Elsevier etc. In order to learn more about earlier studies that addressed their study questions and themes, researchers used SLRs. Table 1 shows the various search strategies that were used in this study.

Table 1. Search Keyword Strategy

Sr. No	Keywords
1.	Named Entity Recognition (NER) of Kumauni Language
2.	Information Extraction from Kumauni Language
3.	Kumauni Language Text Analysis
4.	Cross Language NER
5.	Challenges in Kumauni Language NER
6.	Usage of Machine Learning (ML) model in NER of Kumauni language
7.	Machine Learning (ML) usage in text mining of Kumauni Language

2.2. Brand Selection for Search

On January 1, 2022, the author checked the Journal of International, Journal of Machine Learning, Cybernetics (ISSN: 1868808X), and Elsevier to see how frequently the most significant brands had been used in prior research. It employs a significant keyword search without any restrictions for each request. Then the results were separated into two categories: validating and reliability investigations or information-collecting investigations. It conducts two groups of inquiries to determine which brand is the

most suitable. Researchers developed a product keyword search for corporations that are either (1) one of the top five bestselling products in 2019 and 2020 or (2) have launched 10 or even more distinct devices during the first batch. Then it filtered the papers' titles, descriptions, and methodology parts from the resultant list. The process of identification was carried out to (1) eliminate publications that are not relevant to the study and (2) determine other products that were employed in these trials. Afterward, the author made a list of such products and made another batch of inquiries, one per newly discovered product. In the results section, researchers summarize the observations, including the search term utilized for every product.

2.3. Scrutinizing of paper for study

The Primary Studies (PS) selection procedure is divided into 4 stages: detection, admissibility, inclusion, and multiple screening. The initial step involves identifying each potentially important study, there were 8416 results in the first search. A systematic examination of the several databases including Scopus, Web of Science, Elsevier, Cite SeerX, Cross Ref, and some conferences of IEEE. The source of this study also includes some website and books. Subsequently, screening of the databases removes the 216 duplicate records of the data, and the automation tools filter 7750 records and provide 450 screened records. The second step does a preliminary evaluation by screening titles, keywords, and abstracts. At this point, 8384 records have been excluded because it does not fit the inclusion requirements, particularly in terms of the scope of research and optimization topic. These two records marked as comprise and those marked as unclear were forwarded for additional review. According to Figure 3, an SR database assessment is shown. It is also necessary to go through by hand the bibliography of all pertinent publications and review papers. The rest of the documents were thoroughly examined. The additional data and abstracts of the papers were analyzed to determine which research should be included and excluded in the present systematic review, and the following criteria were used as discussed in Table 2.

Table 2. Inclusion and Exclusion criteria of systematic review

Inclusion Criteria	Exclusion Criteria
Paper Should be in Peer-reviewed	Paper don't focus on body stress related study
Paper should be in English language.	Grey literature
No time frame limit for publication.	Duplicate Research and Publication
Paper should be published in research or full article publication.	Ph.D. theses, working paper, and project deliverable.

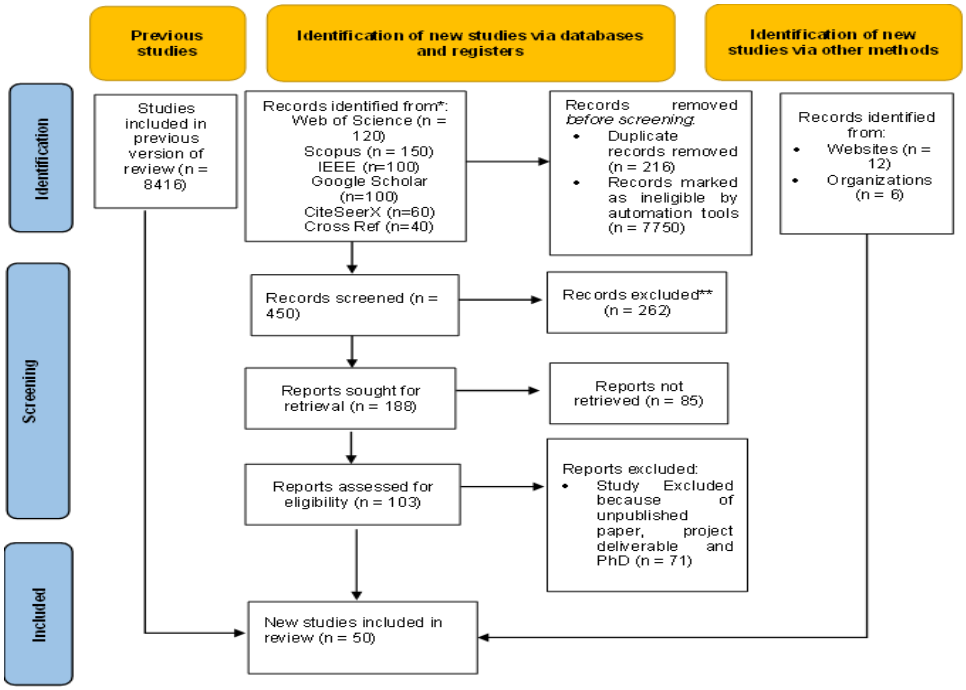

Figure 3. Prisma Diagram of Systematic Review of published article in databases.

2.4. Objective of the current study

Kumaoni is often categorized as a dialect of the Pahari language. The number of native speakers of Kumaoni was estimated to be 2,360,000 in 1998. This Kumani is known by many different names. NER is a subtask of information removal that aims to find and classify components in text into specified categories, including names of people, places, companies, dates, amounts of money, and percentages. The goal of this is to make use of both labeled data and relevant information from unlabel data in learning. The goal of this study is to examine the opportunities and obstacles associated with developing a chatbot for the Kumaon language. We also give a comprehensive analysis of Kumaon and a mapping of the language to others to facilitate its industrial application in processing. The aim of this study is to analyze how well Kumaoni chatbots can serve as a language-learning tool, as well as any obstacles that can arise during deployment. The fundamental benefit of CRFs over HMMs is their conditional character, which allows the independence requirements necessary for HMMs to provide feasible inferences to be relaxed.

CRFs are used to determine the likelihood that certain output node values would be reached given the state of certain or all input nodes. The provisional probability of a state sequence $s = < s1, s2 \ldots, sT >$ given an examination sequence $o = < o1, o2 \ldots, oT >$ is calculated as:

$$P^{(s|o)} = \frac{1}{Z_o} \exp\left(\sum_{t=1}^{T} \sum_{k=1}^{K} \lambda k \, X \, fk(st-1, st, o, t)\right) \quad (1)$$

Where, $fk\ st - 1, st, o, t$ is a feature function whose weight λk, is to be learned vi training. The values of the feature function may range between -∞…… + ∞ but typically they are binary.

A further goal of these SRs is the development of an open-source knowledge platform to support future research on the topic by collecting and analyzing key findings from previous research, summarizing, comparing them, and identifying the issues and limitations that have arisen because of their work. Research on the NER of the Kumauni language using ML was undertaken by assessing the current level of information in the field. The three main investigation questions were formulated during the design phase of the study, and investigators are discussed and evaluated throughout the article. The following are the investigation questions.

RQ 1: How to analyze the chatbots for the Kumauni dialect?

RQ 2: Which technique is used for NER?

RQ 3: What are the advantages of CRF?

To begin with, a comprehensive review of the existing literature was conducted to meet these goals. Sage, Emerald, Google Scholar, Multidisciplinary Digital Publishing Institute (MDPI), Science Direct, Institute of Electrical and Electronics Engineers (IEEE) Xplore, and Springer link were used to explore the citation indexing databases and internet of publications to locate related papers published in the recent ten years. In addition, an internet search was conducted to find the top wrist-wearable gadget makers. Documents such as white papers, manuals from manufacturers, and peer-reviewed research studies were consulted, and the results were analyzed using the data.

3. Advancements in NER: Unraveling the Kumauni Language with MachineLearning

The following study expands on NER of the Kumauni language using ML. Several researchers explained their findings as seen below.

Gusain et al., (2023) [29] proposed a machine learning model aimed at choosing four languages related to the Northern Indo-Aryan family, commonly referred to as Pahari languages. These languages include Kumaoni, Garhwali, Nepali, and Dogri. A corpus of data gathered by the author, which includes data in Dogri and Nepali, is statistically looked at the word level. We also used this corpus to teach traditional machine learning models to recognize the Pahari language. Character n-grams based Linear Support Vector Machines did the best, achieving 99.28% of the answers right.

Rawat et al., (2022) [30] analyze the potential opportunities and obstacles associated with the implementation of a chatbot in the Kumaon language. Additionally, this study offers a comprehensive examination of the Kumaon language, including its linguistic characteristics and features. Furthermore, it aims to establish linguistic connections between Kumaon and other languages, hence facilitating its practical application in industrial contexts. The present chatbot possesses the capability to provide assistance and cater to diverse requirements and services in the Kumaon language. This research uses a study analysis approach to address the issue of language extinction specifically within the context of the Kumaon language. The uniqueness in this research is a chatbot in the Kumaon language with end-to-end encryption so that the service user has good security.

Kadam et al., (2022) [31] focused on a rule-based stemmers' study. It uses the truncating technique. Even though it is one of the most spoken languages in the world and written in the Devanagari script, Marathi data play is not as common as it is with data from other languages. One of India's most popular languages, Marathi (written script: Devanagari), is employed here, and its common nouns and simple tenses have

been investigated for use in this approach. Diversity of rule-based stemmers have been researched in this study for different Devanagari scripts including Marathi, Hindi, and Kumaoni language. It has been found that a stemmer can be enhanced by specifying new rules in accordance with the grammar specifications of the target language.

Kumar et al., (2021) [32] suggested a deep learning neural network model or extracting semantic information from several corpora. It is divided into two sections. The first phase introduces a data preprocessing method that eliminates ambiguity in the input corpora. In the second phase, a novel deep learning-based long-term short-term memory with rectification in the Adam optimizer and multilayer perceptron is proposed to identify events, relations, and agricultural-based named entity recognition. The algorithm under consideration underwent training and evaluation using four distinct input corpora: agriculture, weather, soil, and pests & fertilizers.

Change et al., (2014) [33] suggested utilizing a recurrent neural network to detect language in the Twitter code switching corpus. The author used word embeddings to train an RNN (Recurrent Neural Network) with only raw information and then used RNN to autonomously acquire meaningful representations. This system is able to surpass the top SVM-based systems that were presented in the EMNLP'14 Code-Switching Workshop by 1% in accuracy or by 17% in error rate reduction when using the same mixed-language Twitter corpus.

Kholiya et al., (2020) [34] discussed that both the Garhwal and the Kumaon regions of Uttarakhand contain villages that have their own holy grooves, which are old temple forests, and it is in these forests that the local populations worship the gods and spirits that reside there. Because of human activity, these protected woods, which are home to a variety of native plant and animal species as well as natural water springs, are on the edge of extinction. The state of Uttarakhand has a long-standing custom of preserving the temple forests that are located close to the villages. These forests are used to perform religious rituals in a sacred grove. Documenting the biodiversity of Uttarakhand's sacred groves is one of the goals of this project. Another objective is to investigate potential ways to preserve these groves and bring back some of the pristine beauty they formerly possessed.

Sobhana et al., (2010) [35] created a Named Entity Recognition (NER) method specifically tailored for Geological literature by employing conditional random fields (CRFs). The system utilizes diverse contextual information and a range of attributes to effectively predict distinct named entity (NE) classes. The training dataset comprises over 200,000 words and has undergone manual annotation using a named entity (NE) tag set consisting of seventeen tags. The system demonstrates the capability to accurately identify and classify 17 distinct types of Named Entities (NEs), achieving a commendable F measure of 75.8%.

Sharma et al, (2018) [36] examined influence of gender and regional cultures on entrepreneurial goals and perceived impediments to entrepreneurship in two distinct regions within a state. Previous scholars have consistently emphasized the importance of conducting research on entrepreneurial intentions that incorporates both gender and cultural dimensions. This is because there is a possibility of an interactive effect between sex and culture, which has not been extensively investigated. Exploring this interaction has the potential to shed light on the conflicting findings that have arisen when examining either sex or culture in isolation.

Ertopcu et al., (2017) [37] designed a novel model for the algorithm of Named Entity Recognition (NER). NLP algorithms are capable of locating numerous entities within a sentence, including money, people, locations, dates, and times. An enormous challenge in these operations is resolving ambiguities regarding whether a given word represents a date, time, or amount of money, or whether it refers to a location, organization, individual, or place. After training a model using a predefined dataset, the author evaluates its performance in comparison to other models. Ultimately, significant results are obtained from a dataset comprising 1400 sentences.

Makhija et al., (2016) [38] examines the diverse stemmers that have been created to represent various languages. The process of stemming involves reducing a word to its root or stem term. Similar in name to a Conflation. Stemming is a critical operation incorporated within applications of Natural Language Processing (NLP). This research endeavors to create a stemmer for the Sindhi language, which is written in Devanagari script. It removes prefixes and suffixes from the stem or root of an inflected word. By minimizing the issues associated with over and under-stemming, this stemmer retrieves information more quickly and efficiently.

Das et al., (2014) [39] provided an analysis of the efficacy of Condition Random Field (CRF)-based Named Entity Recognition (NER) systems in the Indian language, which was presented as a collaborative task at ICON 2013. A collection of language-independent characteristics was taken into account for all languages in this study. One feature that has been incorporated exclusively to the English language is capitalization. Bengali, Hindi, and English gazetteer usage is subsequently examined. English attains the maximum F

measure of 88% using the proposed CRF-based system, while Tamil and Telugu both achieve the lowest F measure of 69%.

Pillai et al., (2013) [40] provided a comprehensive review of NER for an Indian language. In contrast to earlier research that relied heavily on handwritten rules, contemporary NER systems are constructed using Machine Learning models including Maximum Entropy (MaxEnt), Hidden Markov Model (HMM), Maximum Entropy Markov Model (MEMM), Conditional Random Fields (CRFs), and Support Vector Machines (SVM). Among the twenty-two official languages of India, NER development has been limited to a select few, including Bengali, Oriya, Tamil, Telugu, Hindi, Punjabi, Urdu, and Kannada. The strategies implemented and the efficacy of these NER in terms of recall, precision, and F-measure are discussed in this study.

Morwal et al., (2012) [41] presented in detail a Hidden Markov Model (HMM)-based machine learning approach to identify the named entities. The objective of NER is to categorize terms into predetermined groups, such as time, date, location, and organization. The fundamental concept underlying the construction of a NER system using the HMM model is that it is language independent, meaning that it can be implemented in any language domain. The fact that the states of our NER system are not fixed indicates that it is dynamic and can be utilized to one's liking. Furthermore, the corpus utilized by our NER system lacks domain specificity.

Sasidhar et al., (2011) [42] elaborated on the utilization of rule-based approaches and language-dependent features in gazetteer lists to identify Named Entities Recognizing (NER) for the Telugu language. The document provided an overview of the recognition of named entities in two phases. The initial stage consists of identifying nouns through the utilization of Telugu dictionaries, noun morphological stemmers, and noun suffixes. Using transliterated gazetteer lists associated with various Named Entity identifiers, Named Entity suffix features, context features, and morphological features, the second phase identifies the Named Entities. There is a wide range of authors who used the technique and presented their discoveries, as given in table 3.

Table 3. Summarize the table of reviewed literature

Author	**Technique Used**	**Outcome**
Gusain et al., (2023) [29]	Linear Support Vector Model	The model trained for Pahari language identification, was able to determined that character n-grams-based Linear Support Vector Machines are 99.28% accurate.
Rawat et al., (2022) [30]	AI based chatbot	The findings of this study reveal the potentialities and obstacles associated with the deployment of a chatbot in the Kumaon language.
Kadam et al., (2022) [31]	Rule based stemmer	The stemmer's efficiency can be enhanced by specifying new rules in accordance with the grammatical specifications of the language of choice.
Kumar et al., (2021) [32]	Deep CNN network	By extracting semantic information from multi-corporal data, the deep CNN attains an accuracy of 88.10 percent.
Change et al., (2014) [33]	RNN network	The proposed RNN network achieves 1% higher accuracy and 17% lower error rate than the top SVM-based systems reported at the EMNLP'14 Code-Switching Workshop.
Kholiya et al., (2020) [34]	Systematic Analysis	In addition to documenting the biodiversity of Uttarakhand's sacred groves, this study examines conservation and restoration strategies for these areas in a broader sense.
Sobhana et al., (2010) [35]	NER and CRF	The suggested NER (Named Entity Recognition) system has an F-measure of 75.8 percent for recognizing 17 types of NEs.
Sharma et al, (2018) [36]	Systematic Analysis	The findings make it clear that the government must design distinct plans for the Garhwal and Kumaon regions. Differential perceptions of barriers between men and women in the Garhwal region call for action on the part of the state administration.
Ertopcu et al., (2017) [37]	NER algorithm	The findings indicate that the performance of these continuous models in NER classification tasks is equivalent to that of supervised discrete features that are manually constructed.
Makhija et al.,	Stemming	It devised a stemmer for the Sindhi language, which is written

(2016) [38]		in Devanagari slate. It removes prefixes and suffixes from the stem or root of an inflected word. By minimizing the issues associated with over and under-stemming, this stemmer retrieves information more quickly and efficiently.
Das et al., (2014) [39]	Conditional Random Field (CRF)	The CRF model under consideration attains its maximum F measure of 88% for the English language and its minimum F measure of 69% for both Tamil and Telugu, both of which are Indian languages.
Pillai et al., (2013) [40]	HMM, Max Entropy, SVM, RF	The evaluation includes an assessment of the performance and F-measure of different machine learning models, including HMM, maximal entropy, MEMM, CRF, SVM, and RF.
Morwal et al., (2012) [41]	HMM	NER divides words into predefined categories using machine learning and the Hidden Markov Model (HMM). These categories include location names, person names, organization names, dates, and times.
Sasidhar et al., (2011) [42]	Named entity recognition (NER)	The result indicates that a named entity is identified through the utilization of transliterated gazetteer lists that contain information on numerous named entity tags, suffix features, context features, and morphological features.

4. Limitations of NER

The problem of NER is not yet solved, but it is solvable. Especially to the degree that any other domain-specific activity could be considered complete. The issue is that we cannot determine due to the lack of adequate assessment processes and tools in NER. NER was thought to be an issue that had been addressed when the techniques attained a minimal level of performance with a small number of NE kinds, document genres, and typically in the journalistic sector [43]. The effectiveness of existing strategies with additional NE types and other sorts of files is uncertain. There are no resources that are universally recognized that could analyze the new kinds of NE that tools recognize today, and the new evaluation forums, even though they overcome some of the limitations that were present in the past, are not appropriate for measuring the evolution of NER because they evaluate systems with different goals, which are not valid for most NER applications [44]. NER, even though it is regarded to be an essential NLP function, is difficult to do because of the myriad of complexity that is present in any natural language. A limited number of the difficulties are discussed in the following:

- **Ambiguity and Abbreviations**

Language is one of the most significant obstacles when attempting to recognize named items. Identifying terms that could function in various sentences and that have varying meanings an additional difficult task is determining how to separate terms that are quite similar in context. It is possible to write several words or phrases in a variety of ways. Abbreviations can make writing and reading faster and more efficient. Words may be written in both short and extended versions. Another formidable obstacle is words that can't be understood without an explanation.

- **Spelling Variations**

The vowels (a, e, I, o, and u) of the English language play a very significant part in the language. Words do not have a significant impact on the phonetics of a language but have a significant impact on the way it is written and the spelling of the language.

- **Foreign Words**

Words that are not used very often in modern times or words that are not heard by a large number of people are another one of the most significant challenges in this field. Words like person names, location names etc. [45].

5. NER for Indian language

The field of NLP study has made a significant step forward due to the advancement of powerful machine learning algorithms and the production of enormous corpora that have been annotated. Insufficient lexical resources, such as annotated corpora, have prevented significant progress in NER for Indian languages. The lack of capitalization, standardized spelling and spelling variance in English NER make it impractical to utilize directly for Indian languages. Additionally, the uncertainties that are present in Indian languages and which deal with linguistic problems such as,

- Agglutinative nature
- Same meaning as a common name and proper name
- Low parts of speech tagging accuracy for nouns

- Patterns and suffixes

These NLP organizations' shared responsibilities aim to have academics and developers collaborate on a topic to provide the most effective solutions possible. The development of the field's state of the art is encouraged via the use of these competition-like events. IJCNLP 2008 hosted the first-ever workshop on natural language processing for five different Indian languages: Hindi, Bengali, Oriya, Telugu, and Urdu. The FIRE 2013 NER for Indian Languages joint task was recently conducted. All the registered groups sent scripts in a variety of languages, including English, Hindi, Tamil, Malayalam, and Bengali [46].

6. Challenges in the Kumauni language

- **No capitalization**

Capitalization is very important in the English language since it is used to determine which words are proper nouns. However, the idea of capitalization does not exist in Indian languages [47].

- **Morphologically rich**

Root identification is particularly challenging because of the high morphological complexity of Indian languages.

- **Ambiguity**

The ambiguity between common and proper nouns.

- **Lack of standardization and Spell Variations**

The fact that different speakers of the same language could use various spellings to refer to the same thing is a major issue in Indian languages.

- **Less Resources**

Pre-processing steps like part-of-speech tagging and chunking are essential for recognizing NE, but they are underdeveloped because of the lack of previous work in NER with Indian languages. Alternatively, the currently available tools either don't do the job, or they do a bad job [48].

- **Lack of labeled data**

There is a lack of corpora and training data for Indian languages.

- **Agglutinative Nature**

Agglutinative indicates that certain other qualities may be added to the word to make it more complicated. Example: Let us assume the root word to be Rup, and the suffix to be Ali, which means GOD. When we combine these two words, we get the new word Rupali, which is the name of a person.

- **Proper Name Ambiguity**

There is ambiguity in proper names in both the English and the Indian language. Names like "White" could be used either as a color name or a personal name in English. Many Indian-given names also have additional, often more precise, definitions in the dictionary, making Indian naming practices more expensive than those of other languages. A significant degree of uncertainty exists even among proper names.

Example: People vs. Companies: Vimal etc.

People vs. Locations: Gandhinagar (person vs. city) People vs. Organizations: Nirma (person vs. university).

Acronyms vs. Organizations: MRI (Magnetic Resonance Imaging vs. Mental Research Institute). People vs. Months: Shrawan (month of the Indian Calendar).

- **Lack of easy availability of annotated data**

Annotated data and a corpus are becoming more limited as few people work on NER in Indian languages [49].

7. Machine learning-based chatbots of the Kumauni language

The term "chatbot" refers to a piece of application programming that is used to moderate conversations that take place over the internet via the utilization of text-to-discourse or text to facilitate communication with real-life human agents. Also, Michael Mauldin gives it the name "chatterbox". Since chatbots are meant to mimic human conversational behavior, their underlying systems need constant fine-tuning and testing, and many existing bots still can't have a decent conversation or don't meet the basic industry requirements of a Turing test. Table 4 depicts a comparison between English, Devanagari (Indian language), and Kumaoni language [50].

Table 4. English and Devanagari language in Kumaoni

S. No	English	Devanagari	Kumaoni
1	I	Mein	Mi/mei/mai
2	He	Who/Usne	Wou/ull/wu

3	She	Who/Usne	Wu
4	You	Tum/Aap (respect)	Tu/ter
5	It	Yeh	Ya
6	This	Yaeh	Yo
7	That	Vaeh	Wo
8	A	EK	A
9	Yes	Haan	Hoye
10	Come	Aao/Aaiye	Aoh
11	Came	Aaya (he)/ Aayee (she)/ Aaye (plural)	Uul
12	Open	Kholo/Kjoliye (respect)/ Kholna	Khulul/Kholna
13	Opened	Khola (he)/ Kholee (she)/ Khole (plural)	Khol halo
14	Sit	Baitho/ Baithiye (respect)/ Baithna (to sit)	Baithnou
15	Walk	Chalo/ Chaliye (respect)/ Chalna (to walk)	Chalul/ hituul
16	Eat	Khao/ Khaiye (respect)/ Khana (to eat)	Khan
17	Go	Jaao/ Jaaiye (respect)/ Jaana (to go)	Jaan/ janai
18	Went	Gaye	Jaan/janai
19	Run	Daudna	Bhagun/daudun
20	He ate an apple	Usne sev Khaya	Ullsaibkheihaaali

8. Conclusion and Future Scope

The field of Named Entity Recognition (NER) has seen extensive development in the English language and other languages, but only modest progress in the Kumaoni languages. The ML approach is the one that works best to locate named entities when it comes to Kumaoni languages. This paper highlights the need of additional research, particularly in the field of NEM of Kumauni language by machine learning technique.

Using a systematic review method, this study pulls from a wide range of well-known databases. The use of different keywords and strict selection criteria helped make sure that the collection of research studies was representative and up to date. The combination of the 14 recent studies has shed light on the NER of Kumauni language using a ML model. In this research 50 different research were considered which utilizing the PRISMA technique, which not only demonstrates how much more can be learned from the subsequent studies that are drawn from the following databases:

Web of Science = 6, Scopus = 5, IEEE = 12, Google Scholar = 2, Cite SeerX = 6, Cross Ref = 4, Elsevier = 4, Books Google = 2, and Google Search = 5, published thesis = 2, and preprint copy = 2.

The total of References taken is 50.

In the conclusion, the advancement in the domain of language recognition has established a symbiotic relationship between technological innovation and the safeguarding of cultural heritage in the Kumaon region. Ongoing research and active participation from the community contribute to the continuous development of the NER models, which guarantees that the technology remains consistent with the cultural and linguistic environment of the Kumauni language.

The purpose of this research is to investigate the potential benefits and drawbacks of creating a chatbot that can communicate in the Kumaon language. In addition to this, the author provides an in-depth examination of Kumaon as well as a mapping of the language into other languages to make its use in industrial processing more accessible. The research activity possesses the capacity to fundamentally transform the field of linguistic evaluation in the Kumaon region by employing sophisticated machine learning (ML) techniques to explore the complex domain of NER (named entity recognition) for the Kumauni language.

REFERENCES

[1] https://www.indianetzone.com/7/kumaoni_language.htm
[2] Kulkarni-Joshi, S. (2019). Linguistic history and language diversity in India: Views and counterviews. Journal of Biosciences, 44(3), 1–10.

[3] Uttaranchal dialects and languages – Uttarakhand worldwide – Kumaoni and Garhwali – Kumaon and Garhwal dialects -. (March 5, 2012). Archived from the original on 5 March 2012. Retrieved February 8, 2021.
[4] Eichentopf, S. R., Boon, S. A., & Benedict, K. D. (2014). 'A Sociolinguistic Study of Dotyali.' an unpublished [Masters Degree Thesis]. Tribhuvan University.
[5] Srinivasagan, K. G., Suganthi, S., & Jeyashenbagavalli, N. (2014). An automated system for Tamil named entity recognition using hybrid approach. In International Conference on Intelligent Computing Applications (pp. 435–439). IEEE Publications. https://doi.org/10.1109/ICICA.2014.95
[6] Khanam, M., & Humera, Md. A. (2016). Khudhus, and MS Prasad Babu. "Named Entity Recognition using Machine learning techniques for Telugu language." In 7th IEEE international conference on software engineering and service science (ICSESS) (pp. 940–944). IEEE Publications.
[7] Vijayakrishna, R., & Sobha, L. (2008). Domain focused named entity recognizer for Tamil using conditional random fields. In Proceedings of the IJCNLP-08 workshop on named entity recognition for South and South East Asian Languages.
[8] Abinaya, N., John, N., Ganesh, B. H. B., Kumar, A. M., & Soman, K. P. (2015). Amrita_cen@fire-2014: Named entity recognition for Indian languages using rich features. In Proceedings of the 6th Annual Meeting of the Forum for the Information Retrieval Evaluation (pp. 103–111). https://doi.org/10.1145/2824864.2824882
[9] Sharma, D. D. (1989). A linguistic geography of Kumaun Himalayas (A descriptive areal distribution of Kumauni language). Mittal Publications.
[10] First Kumaoni film of Uttarakhand [YouTube]. Archived from the original on 13 December 2021. Retrieved April 16, 2021.
[11] Uttarakhand Kumaoni language. Accessed 2012. https://www.indiamapped.com/languages-in-india/uttarakhand-kumaoni-language/
[12] Speck, R., & Ngomo, A.-C. N. (2014). Ensemble learning for named entity recognition. In The semantic web–ISWC. Proceedings, Part I: 13th International Semantic Web Conference, Riva del Garda, Italy, October 19–23, 2014, 13 (pp. 519–534). Springer International Publishing.
[13] Gao, C., Zhang, X., Han, Mengting, & Liu, H. (2021). A review on cyber security named entity recognition. Frontiers of Information Technology and Electronic Engineering, 22(9), 1153–1168. https://doi.org/10.1631/FITEE.2000286
[14] Ekbal, A., & Saha, S. (2016). Simultaneous feature and parameter selection using multiobjective optimization: Application to named entity recognition. International Journal of Machine Learning and Cybernetics, 7(4), 597–611. https://doi.org/10.1007/s13042-014-0268-7
[15] Marrero, M., Urbano, J., Sánchez-Cuadrado, S., Morato, J., & Gómez-Berbís, J. M. (2013). Named entity recognition: Fallacies, challenges and opportunities. Computer Standards and Interfaces, 35(5), 482–489. https://doi.org/10.1016/j.csi.2012.09.004
[16] Gridach, M. (2017). Character-level neural network for biomedical named entity recognition. Journal of Biomedical Informatics, 70, 85–91. https://doi.org/10.1016/j.jbi.2017.05.002
[17] Trewartha, A., Walker, N., Huo, Haoyan, Lee, S., Cruse, K., Dagdelen, J., Dunn, A., Persson, K. A., Ceder, G., & Jain, A. (2022). Quantifying the advantage of domain-specific pre-training on named entity recognition tasks in materials science. Patterns, 3(4), 100488. https://doi.org/10.1016/j.patter.2022.100488
[18] Miranda, N., Raminhos, R., Seabra, P., Sequeira, Joao, Gonçalves, T., & Quaresma, P. (2011). Named entity recognition using machine learning techniques. In EPIA-11, 15th Portuguese Conference on Artificial Intelligence (pp. 818–831).
[19] Shelar, H., Kaur, G., Heda, N., & Agrawal, P. (2020). Named entity recognition approaches and their comparison for custom ner model. Science and Technology Libraries, 39(3), 324–337. https://doi.org/10.1080/0194262X.2020.1759479
[20] Kanya, N., & Ravi, T. (2012). Modelings and techniques in named entity recognition: An information extraction task (pp. 104–108). https://doi.org/10.1049/cp.2012.2199
[21] Patil, N., Patil, A. S., & Pawar, B. V. (2016). Survey of named entity recognition systems with respect to Indian and foreign languages. International Journal of Computer Applications, 134(16), 21–26. https://doi.org/10.5120/ijca2016908197
[22] Allen, J. (1995). Natural language understanding. Benjamin-Cummings Publishing, Co., Inc.
[23] Artiles, J., Amigó, E., & Gonzalo, J. (2009). The role of named entities in web people search. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (pp. 534–542). https://doi.org/10.3115/1699571.1699582
[24] Dey, A., & Purkayastha, B. S. (2012). Named entity recognition A computational approach [PhD Diss.].

[25] Shah, H., Bhandari, P., Mistry, K., Thakor, S., Patel, M., & Ahir, K. (2016). Study of named entity recognition for Indian languages. International Journal of Information Sciences and Techniques, 6(1/2), 11–25. https://doi.org/10.5121/ijist.2016.6202
[26] Wang, X., Yang, C., & Guan, Renchu. (2018). A comparative study for biomedical named entity recognition. International Journal of Machine Learning and Cybernetics, 9(3), 373–382. https://doi.org/10.1007/s13042-015-0426-6
[27] Patawar, M. L., & Potey, M. (2015). Approaches to named entity recognition: A survey. International Journal of Innovative Research in Computer and Communication Engineering, 3(12), 12201–12208.
[28] Dogra, M., Tyagi, A., & Mishra, U. (2013). An effective stemmer in Devanagari script. In Proceeding of the International Conference on Recent Trends In Computing and Communication Engineering-RTCCE (pp. 22–25).
[29] Gusain, R., Dash, S. R., Parida, S., & Jha, G. N. (2023). Automatic language identification: A case study of Pahari languages. Language Resources and Evaluation, 57(3), 1361–1387. https://doi.org/10.1007/s10579-023-09651-6
[30] Rawat, B., Bist, A. S., Mehra, N., Fazri, M. F., & Terah, Y. A. (2022). Study of kumaon language for natural language processing in end-to-end conversation scenario. IAIC Transactions on Sustainable Digital Innovation, 3(2), 143–149. https://doi.org/10.34306/itsdi.v3i2.534
[31] Kadam, D. P. (2022). Develop a Marathi lemmatizer for common nouns and simple tenses of verbs. In International Conference on Advances in Data-driven Computing and Intelligent Systems (pp. 323–333). Springer Nature Singapore.
[32] Kumar, S., Hanumat Sastry, G., Marriboyina, V., Alshazly, H., Idris, S. A., Verma, M., & Kaur, M. (2021). Semantic information extraction from multi-corpora using deep learning. Computers, Materials and Continua, 1–17.
[33] Chang, J. C., & Lin, C.-C. (2014). 'Recurrent-neural-network for language detection on twitter code-switching corpus.' arXiv preprint arXiv:1412.4314.
[34] Kholiya, D., Rawat, L., & Joshi, P. Ethno-botanical recordings from doonagiri sacred Groove in Dwarahat, Kumaun Himalayas (Uttarakhand), India.
[35] Sobhana, N. V., Mitra, P., & Ghosh, S. K. (2010). Conditional random field based named entity recognition in geological text. International Journal of Computer Applications, 1(3), 143–147. https://doi.org/10.5120/72-166
[36] Sharma, L. (2018). Entrepreneurial intentions and perceived barriers to entrepreneurship among youth in Uttarakhand state of India: A cross-cultural investigation across genders. International Journal of Gender and Entrepreneurship, 10(3), 243–269. https://doi.org/10.1108/IJGE-02-2018-0009
[37] Ertopçu, B., Kanburoğlu, A. B., Topsakal, O., Açıkgöz, O., Gürkan, A. T., Özenç, B., Çam, I., Avar, B., Ercan, G., & Yıldız, O. T. (2017). A new approach for named entity recognition. In International Conference on Computer Science and Engineering (UBMK) (pp. 474–479). IEEE Publications. https://doi.org/10.1109/UBMK.2017.8093439
[38] Makhija, S. D. (2016). A study of different stemmer for Sindhi language based on devanagari script. In 3rd International Conference on Computing for Sustainable Global Development (INDIACom) (pp. 2326–2329). IEEE Publications.
[39] Das, A., & Garain, U. (2014). 'Crf-based named entity recognition@icon 2013.' arXiv preprint arXiv:1409.8008.
[40] Pillai, A. S., & Sobha, L. (2013). Named entity recognition for Indian languages: A survey. International Journal, 3(11).
[41] Morwal, S., Jahan, N., & Chopra, D. (2012). Named entity recognition using hidden Markov model (HMM). International Journal on Natural Language Computing, 1(4), 15–23. https://doi.org/10.5121/ijnlc.2012.1402
[42] Sasidhar, B., Yohan, P. M., Vinaya Babu, A. V., & Govardhan, A. (2011). Named entity recognition in Telugu language using language dependent features and rule based approach. International Journal of Computer Applications, 22(8), 30–34. https://doi.org/10.5120/2602-3628
[43] Ekbal, A., Saha, S., & Sikdar, U. K. (2016). On active annotation for named entity recognition. International Journal of Machine Learning and Cybernetics, 7(4), 623–640. https://doi.org/10.1007/s13042-014-0275-8
[44] Marrero, M., Urbano, J., Sánchez-Cuadrado, S., Morato, J., & Gómez-Berbís, J. M. (2013). Named entity recognition: Fallacies, challenges and opportunities. Computer Standards and Interfaces, 35(5), 482–489. https://doi.org/10.1016/j.csi.2012.09.004

[45] Kamath, S., & Wagh, R. (2017). Named entity recognition approaches and challenges. International Journal of Advanced Research in Computer and Communication Engineering (IJARCCE), 6(2), 259–262.
[46] Hiremath, P., & Shambhavi, B. R. Approaches to named entity recognition in Indian languages: A study. International Journal of Engineering and Advanced Technology (IJEAT), ISSN 2249, 8958.
[47] Kaur, D., & Gupta, V. (2010). A survey of named entity recognition in English and other Indian languages. International Journal of Computer Science Issues (IJCSI), 7(6), 239.
[48] Sharma, P., Sharma, U., & Kalita, J. (2011). Named entity recognition: A survey for the Indian languages. Parsing in Indian Languages, 35–39.
[49] Shah, H., Bhandari, P., Mistry, K., Thakor, S., Patel, M., & Ahir, K. (2016). Study of named entity recognition for Indian languages. International Journal of Information Sciences and Techniques, 6(1/2), 11–25. https://doi.org/10.5121/ijist.2016.6202.
[50] Rawat, B., Bist, A. S., Mehra, N., Fazri, M. F., & Terah, Y. A. (2022). Study of kumaon language for natural language processing in end-to-end conversation scenario. IAIC Transactions on Sustainable Digital Innovation, 3(2), 143–149. https://doi.org/10.34306/itsdi.v3i2.534.