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

Vinay Kumar Pant (✉ vp5570@srmist.edu.in)
   SRMIST (Deemed to be University) Delhi NCR Campus

Rupak Sharma
   SRMIST (Deemed to be University

Shakti Kundu
   Manipal University Jaipur (MUJ)

Research Article

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

Posted Date: January 4th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2424051/v1

License: ☋ ☑️ This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: No competing interests reported.
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.

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. 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]. everal 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.
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. South-eastern 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-Pachhhaau (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 with 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
towards 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, NLR development for Indian languages is quite challenging [6].

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. DNA, protein, and gene names are only some of the things that the genetics NER system must be able to recognize [8].

1.1 Kumaoni Language

The Kumaon people have their 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 compensate 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 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 on the decline, and 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 Devanagri 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 of 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 benefited greatly 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), CRF, and 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 subfields 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 HMM, DE, Maximum Entropy Models (MEM), CRF, and 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.
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, effective search algorithm, etc. There is a distinct need for recognizing 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 a 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, we discussed and analyzed the performance of several NLP tools for NER models, with Apache OpenNLP, Python’s SpaCy, and TensorFlow. Time, accuracy, and performing quality are among the many criteria used to evaluate different NER tools. NER could make use of a selection of learning strategies, with but not limited to semi-supervised learning, supervised learning, and unsupervised learning [20]. In the various forms, these methods all have significant advantages. SVM, MEM, HMM, DT, 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 is accuracy a consideration when comparing NER models, but so are prediction time, model size, and training simplicity [21].

1.3 Applications 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/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, NEI presents a significant difficulty for the 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 various 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 AllTheWeb 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 summarising. It’s inefficient and time-consuming for users to go through websites and manually summarise the content. Automatic summarising 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 ML 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 preserved the inflected word and its root forms), the root form is returned; otherwise, the inflected word is subjected to stripping rules [28].

## 2 Review Methodology

The technique described in this Systematic Review (SR) has been based on the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) declarations. PRISMA is a minimal collection of systematic reviews and meta-analyses that are related to evidence. It is usually intended for reporting evaluations that assess the impacts of involvements, though it can also be utilized to record systematic reviews with objectives other than assessing interferences example: prevalence, diagnosis, or prognosis.

### 2.1 Search Strategy

A Systematic Literature Review (SLR) focuses on the Prospects of the NER of the Kumauni Language using ML and appropriate case-base structure for the collected cases to enable their effective retrieval. as part of this endeavor. The initial information sources consist of four primary academic literature collections: Science Direct, IEEE, Springer, and Scopus databases. When researching a certain research issue or topic area, researchers conduct SLRs to acquire information on relevant research in that field.

The last search was on the 25th of January 2022 and repeated database searches were conducted using keywords to locate significant academic material. The essential phrases were searched in the topic and title, as well as the topic, title, and abstract with no time restriction (Scopus and others). Article, review, proceedings paper, bibliography, and article were the only document types allowed, as well as conference paper, article, or review. Table 1 lists the terms that were utilized in Scopus and Near. More than one alternate keyword spelling was also examined. There was a total of eight records in this batch. Table 1 below shows the strategy of search keywords.

<table>
<thead>
<tr>
<th>Keywords</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. What is the Kumaoni language?</td>
</tr>
<tr>
<td>2. What is the role of the Kumouni language in media?</td>
</tr>
<tr>
<td>3. What is Named-Entity-Recognition (NER)?</td>
</tr>
<tr>
<td>4. Why we used NER?</td>
</tr>
<tr>
<td>5. What are the approaches of NER?</td>
</tr>
<tr>
<td>6. What is the application of NER?</td>
</tr>
<tr>
<td>7. Why we used NLP in NER?</td>
</tr>
</tbody>
</table>

### The Brand Selected for Research

On January 1, 2022, checked the Journal of International Journal of Machine Learning and Cybernetics (ISSN: 1868808X), Elsevier to see how frequently the most significant brands had been used in prior research. Used an important word searching with no restrictions for every request. The results were separated into two categories: validating and reliability investigations and information-collecting investigations. Conducted two groups of inquiries to determine what brand was the
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 filtered the papers’ titles, descriptions, and methodology parts from the resultant list. Such identification was performed to (1) eliminate publications that are not relevant to the study and (2) determine other products that were employed in these trials. Then made a list of such products and did another batch of inquiries, one per newly discovered product. Eventually, ten names were added. In the results section, researchers summarize the observations, including the search term utilized for every product.

2.2 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 2375 results in the first search. A systematic examination of several databases and sources, including full-text articles, science direct, springer, Scopus, and IEEE Xplore yielded conference proceedings. To remove multiple copies, all results were screened and analyzed, yielding a total of 217 studies, 28 percent of which were focused on emotion stress detection, 17 percent on scientific viability and consistency, as well as 33 percent were investigatory research on science-based recent applications and innovations. The second step does a preliminary evaluation by screening titles, keywords, and abstracts. At this point, 2335 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.
Figure 3: Flowchart for doing SR of published articles in databases

Table 2: Exclusion and Inclusion principles for SR

<table>
<thead>
<tr>
<th>inclusion criteria</th>
<th>exclusion criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>I1: The paper should be peer-reviewed.</td>
<td>E1: Papers that do not focus on the body stress-related study.</td>
</tr>
<tr>
<td>I2: The paper should be in the English language.</td>
<td>E2: Grey literature</td>
</tr>
<tr>
<td>I3: No time frame limit for publication</td>
<td>E3: Duplicate research and publications</td>
</tr>
<tr>
<td>I4: Papers should be published in research or full-article publication.</td>
<td>E4: Ph.D. theses, working papers, and project deliverables</td>
</tr>
<tr>
<td>I5: Standard of paper was blind to impact factor</td>
<td></td>
</tr>
</tbody>
</table>

Studies that were ruled out at the outset of the screening process:

Based on the title and the abstract. = 2000

Studies that were excluded = 227, due to duplication of Conference publications. Certain studies were removed. = 125

Studies excluded due to insufficient data. = 85

Studies excluded because of unpublished papers, Projects deliverables and Ph.D. = 53
2.3 Objectives 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 which is to make use of both labeled data and relevant information from unlabeled 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.

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 \( s = \langle s_1, s_2, ..., s_T \rangle \) given an examination sequence \( o = \langle o_1, o_2, ..., o_T \rangle \) is calculated as:

\[
p(s|o) = \frac{1}{Z_o} \exp \left( \sum_{t=1}^{T} \sum_{k=1}^{K} \lambda_k X f_k(s_{t-1}, s_t, o_t) \right)
\]

Where, \( f_k st_{t-1}, st, o_t \) is a feature function whose weight \( \lambda_k \), is to be learned via training. The values of the feature function may range between \(-\infty \ldots \ldots + \infty \) 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 work. Research on 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 designing 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 Kumaoni 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 Literature of Review

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

Rawat et al., (2022) [29] examined the opportunities and constraints of developing a chatbot for the Kumaon language. It also links the Kumaon language to certain other languages to facilitate its industrial processing and offer a comprehensive overview of the language. The Kumaon language is supported by this chatbot so that a variety of services can be provided. An investigation of the Kumaon language is utilized as a case study to explore the issue of language loss. The new aspect of this study is an encrypted Kumaon-speaking chatbot for providing peace of mind to users. Artificial intelligence relies heavily on NLP.

Renkovskaya E. (2020) [30] emphasized that the focus of the article is on the apara-derived associative plurals seen in modern IA languages. The grammaticalization of these markers into other grammatical units, such as honorific particles, standard plural markers, definiteness markers, inalienable ownership
markers, etc., is common throughout NIA languages. Contacts with languages outside the Indo-Aryan family are a key component in this grammatical evolution.

Saxena V. et al., (2020) [31] investigated the prenatal phase and infant care customs in Uttarakhand. It used a qualitative descriptive research approach to learn more about 990 traditional birth attendants’ personal experiences in maternal care and deliveries over a substantial number of years. The information was gathered via in-depth, one-on-one interviews. Participants expressed themselves via narratives, picturizations, songs, and role-plays to demonstrate their skills; these expressions were translated from the interviewers’ native language into English before analysis. Many of the attendees came from far-flung regions of Uttarakhand and spoke only their native Garhwali or Kumaon language. Data was translated into descriptions/verbatims with the assistance of a local language specialist.

Tewari et al., (2020) [32] investigate the traditional plant applications and ethnopharmacological data in the Kumaun Himalaya. A semi-structured questionnaire was used to conduct in-depth interviews with a total of 26 traditional healers and seasoned locals, ranging in age from 30 to 85. Quantitative analysis was performed on the data by considering its practical use value. In addition, for species with UV greater than 0.15, the informant consensus factor (ICF) and fidelity level (FL) were calculated. The authors know of 56 plant species from 34 different families. The Asteraceae and Lamiaceae families, which include the most species, were the most heavily sampled, followed by the Rosaceae family. The primary functions of plants were grouped into 29 disease groups. Most of the reported species’ use was for GI issues (11.21%), then for immuno-modulation (10.2%), adaptogens (10.2%), anti-stress (10.2%), nervous system issues (6.54%), analgesics (7.47%), and antimicrobials (6.54%).

Imran Md et al., (2018) [33] declared that the objective of this study is to investigate the feasibility of homestay programs in Uttarakhand and the openness of the local population to participate in programs. Homestay programs are unique in the tourist industry of India. Specifically, only the Munsiyari and Sar-moli villages in the Johar Valley and the Ladakh Autonomous Region in Jammu and Kashmir fall under this category. Household surveys, an average treatment effect analysis, an analysis of community economic development, in-field observations, analytical interpretation, and an ethnographic understanding of life in K village, Nainital District, Uttarakhand are all part of this study’s mix of qualitative and quantitative methods. The findings indicate that families working in the tourist business have not benefited economically; this could be due to the local community not being actively engaged in tourism development. Most locals also believed the homestay program could pave the way for new employment and economic prospects.

Sharma et al., (2018) [34] evaluated this study to examine how gender and regional culture influence entrepreneurship-related goals and perceived obstacles in two geographically distinct areas of a state. Information was mostly collected from individuals enrolled in graduate-level programs. The data was collected using a stratified random sample design with proportional sampling as the primary strategy. Validity and reliability tests were conducted on the instruments. Three different statistical methods were employed to determine significant correlations: Fisher’s exact test, the chi-square test, and the Mann-Whitney U test. There were found to be substantial disparities in the way young men and women saw obstacles and how likely they were to take the entrepreneurial plunge. The results also show that the perceptions of barriers and the entrepreneurial aspirations of both genders differ according to cultural shifts at the regional level.

Cynthia et al., (2017) [35] computed that an examination of the linguistic variety in India with a focus on the policies that affect the country’s linguistic minority is presented here. Uttarakhand’s Kumaun area speaks for the minority languages of the country when it comes to language legislation at the national level. The Indian Constitution goes into the more apparent status planning included in the designation of official languages. To better understand how status and acquisitions planning impact India’s linguistic landscape, it is helpful to have a look at the country’s language-in-education rules for languages to be studied and languages to be utilized as a medium of instruction. Finally, the discussion of linguistic minority education outside the application of official language and education regulations reveals the prevalence of pluralistic language practices in multilingual environments.

Ertopcu et al., (2017) [36] found that a great number of phrases generate impacts on individuals. These thoughts provide context for the reader by way of certain things in the phrase. In NLP, this operation is known as NER. NLP algorithms can identify many different types of things in a text, including names, places, dates, times, and amounts of money. One of the primary issues in these processes is determining
if a word represents the name of a person, a place, or an organization, or whether an integer represents a
time, date, or money. Here, the authors create a new paradigm for NER algorithms. The performance of the
model, the authors have trained it using a custom dataset and compared the findings to those of competing
models. Significant results are achieved using a dataset of 1400 phrases.

**Makhija S. (2016)** [37] stated that internet access now plays a crucial part in people’s lives all around the
world. Data must be obtained by search engines in the blink of an eye. Internet search engines are always
experimenting with new methods to improve their results and response times. Stemming is a method for
breaking down a word into its smallest possible component, known as the root word or stems word and it
is also known as conflation. The process of stemming is an integral part of many NLP programs. This study
explores the numerous stemmers available, each tailored to a certain language. Additionally, a stemmer
for the Sindhi language, which uses the Devanagari script, is an objective of this study. Inflected words
are reduced to their base forms by removing the prefixes and suffixes. The study of stemmers for several
different languages that’s why it’s important to recognize the over and under-stemming problems that often
plague stemmers.

**Leavitt et al., (2014)** [38] evaluated that this study provides a brief overview of the development of
interpretation theory in the contemporary Western world and claims that the two fields are mutually de-
pendent. Anthropologists are well positioned to carry out interpretations that take context seriously and to
conduct ethnographies that focus on texts because they are not constrained by the expectation that profes-
sional translators create texts that are readily consumable for the intended audience. Such “ugly” (Ortega
y Gasset) translations can require the reader to try to reposition themselves, to transcend a border into
what could be another world, first another linguistic world. The conclusion of this study is based on these
arguments and text artifacts culled from Central Himalayan oral traditions.

**Das A. et al., (2013)** [39] stated that authors analyze the efficiency of CRF-based NER systems for
the Indian language. Authors have considered a set of universal characteristics shared by all languages.
Capitalization is a language-specific characteristic that has been applied only to the English language. The
next step is to investigate the Bengali, Hindi, and English gazetteer. These encyclopedias are compiled using
the information found on Wikipedia and other websites. The results of the studies suggest that the system
performs best in English (F = 88%) and worst in both Tamil and Telugu (F = 69%). It should be noted
that no gazetteer was utilized for the two lowest-performing languages. Accuracy (F measure) for NER in
Bengali and Hindi is 87% and 79%, respectively.

**Pallavi et al., (2013)** [40] found that the process of locating and categorizing names within a dataset
is called NER, which is a sub-task of IE. Most of the initial research focused on handwritten rules, while
today’s NER systems are built using Machine Learning models like the HMM, Maximum Entropy Markov
model (MEMM), Maximum Entropy (MaxEnt), SVM, and CRFs. The purpose of this study is to provide
Indian language NERs that have been constructed utilizing machine learning techniques. NER work has
only been completed for a few Indian languages, including Tamil, Hindi, Bengali, Telugu, Oriya, Punjabi,
Urdu, and Kannada, among the 22 official languages. This research goes into the methods used and the
results of these NERS in terms of recall, accuracy, and the F-measure.

**Sasidhar et al., (2011)** [41] described two-phase NER for the Telugu language. The initial part of the
process involves identifying nouns with the use of a Telugu dictionary, a stemmer, and a set of suffixes.
In the second stage, the NE is recognized via the use of transliterated gazetteer listings associated with
Named Entity tags, various NE suffix characteristics, context features, and morphological features. The
authors found that Rule Based methods might be effective with enough gazetteer lists, language-dependent
characteristics, and rules for a single language.

**Morwal S. et al., (2012)** [42] analyzed that NLP is a subfield of AI that deals with processing written
language to extract meaningful information, and one of its subtasks is NER. It finds use mostly in machine
translation, text-to-speech synthesis, natural language comprehension, IE, Information retrieval, question
answering, etc. The purpose of NER is to sort words into several classes, such as places, people, businesses,
times, dates, etc. In this study, we provide a comprehensive overview of the HMM-based method of NER.
HMM, models are used to construct NER systems with the primary motivation being flexibility and gener-
alization across language domains. The NER system is not static so users could tailor the experience to their
own goals. Similarly, the NER system’s corpus is not industry specific.
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.

<table>
<thead>
<tr>
<th>Authors</th>
<th>Technique Used</th>
<th>Outcomes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rawat et al., (2022) [29]</td>
<td>NLP</td>
<td>This study indicates that the Kumaoni language learning tool could be utilized to make chatbots more efficient in usage at once, which would ultimately aid in preventing the extinction of the language.</td>
</tr>
<tr>
<td>Renkovskaya E. (2020) [30]</td>
<td>Machine Learning</td>
<td>These indicators are included in many NIA languages, although in many of these languages they have been grammaticalized into various grammatical units, such as standard plural markers, honorific particles, definiteness markers, markers of irreversible property, and so on.</td>
</tr>
<tr>
<td>Saxena V. et al., (2020) [31]</td>
<td>Machine Learning</td>
<td>The data were described and produced with the aid of a native speaker of the target language.</td>
</tr>
<tr>
<td>Tewari et al., (2020) [32]</td>
<td>ICF and FL</td>
<td>Most species were reported to be utilized for gastrointestinal illnesses (11.21%), then for immuno-modulation (10.2%), anti-stress (10.2%), adaptogens (10.2%), analgesics (7.47%), nervous system-related disorders (6.54%), and antimicrobials (6.54%).</td>
</tr>
<tr>
<td>Imran Md et al., (2018) [33]</td>
<td>Propensity Score Matching (PSM) methods</td>
<td>The tourist sector has not provided economic advantages to the households engaged; however, this might be due to the community's lack of involvement in tourism development.</td>
</tr>
<tr>
<td>Sharma et al., (2018) [34]</td>
<td>chi-square test</td>
<td>The results demonstrated that cultural shifts at the regional level affect how both men and women see barriers to entrepreneurship and how likely they are to start their businesses.</td>
</tr>
<tr>
<td>Ertopcu et al., (2017) [36]</td>
<td>NLP</td>
<td>The model uses the custom-built dataset and evaluates it alongside other models. Significant results are obtained on a dataset of 1400 sentences.</td>
</tr>
<tr>
<td>Makhija S. et al., (2016) [37]</td>
<td>NLP</td>
<td>The study of stemmers for several different languages that's why it's important to recognize the over and under-stemming problems that often plague stemmers.</td>
</tr>
<tr>
<td>Leavitt et al., (2014) [38]</td>
<td>Machine Learning</td>
<td>The research analyzes some of these ideas using text artifacts based on Central Himalayan oral traditions.</td>
</tr>
<tr>
<td>Das A. et al., (2013) [39]</td>
<td>CRF</td>
<td>The system's F-measure is 88% for English and 69% for each Tamil and Telugu.</td>
</tr>
<tr>
<td>Pillai et al., (2013) [40]</td>
<td>HEMM, SVM, and CRF</td>
<td>Telugu was able to get a high f-measure using the CRF method. Accuracy for bilingual NERs is enhanced using the hybrid method.</td>
</tr>
<tr>
<td>Sasidhar et al., (2011) [41]</td>
<td>CRF and HMM</td>
<td>They found that rule-based methods can perform well given adequate gazetteer sets, language-dependent characteristics, and rules for a single language.</td>
</tr>
<tr>
<td>Morwal S. et al., (2012) [42]</td>
<td>HMM</td>
<td>Its dynamic nature and lack of defined state boundaries make the NER system adaptable to a wide variety of applications.</td>
</tr>
</tbody>
</table>

13
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 otherwise 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 always 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 Indic 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 expansive 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).

- 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, Devnagri (Indian language), and Kumaoni language [50].

<table>
<thead>
<tr>
<th>S. No</th>
<th>English</th>
<th>Devanagari</th>
<th>Kumaoni</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>I</td>
<td>Mein</td>
<td>Mi/mei/mai</td>
</tr>
<tr>
<td>2</td>
<td>He</td>
<td>Who/Usne</td>
<td>Wou/ull/wu</td>
</tr>
<tr>
<td>3</td>
<td>She</td>
<td>Who/Usne</td>
<td>Wu</td>
</tr>
<tr>
<td>4</td>
<td>You</td>
<td>Tum/Aap (respect)</td>
<td>Tu/ter</td>
</tr>
<tr>
<td>5</td>
<td>It</td>
<td>Yeh</td>
<td>Ya</td>
</tr>
<tr>
<td>6</td>
<td>This</td>
<td>Yaeh</td>
<td>Yo</td>
</tr>
<tr>
<td>7</td>
<td>That</td>
<td>Vaeh</td>
<td>Wo</td>
</tr>
<tr>
<td>8</td>
<td>A</td>
<td>EK</td>
<td>A</td>
</tr>
<tr>
<td>9</td>
<td>Yes</td>
<td>Haan</td>
<td>Hoye</td>
</tr>
<tr>
<td>10</td>
<td>Come</td>
<td>Aao/Aaiye</td>
<td>Aoh</td>
</tr>
<tr>
<td>11</td>
<td>Came</td>
<td>Aaya (he)/ Aayee (she)/ Aaye (plural)</td>
<td>Uul</td>
</tr>
<tr>
<td>12</td>
<td>Open</td>
<td>Kholo/Kjoliye (respect)/ Kholna (plural)</td>
<td>Khulul/Kholna</td>
</tr>
<tr>
<td>13</td>
<td>Opened</td>
<td>Khola (he)/ Kholee (she)/ Khole (plural)</td>
<td>Khol halo</td>
</tr>
<tr>
<td>14</td>
<td>Sit</td>
<td>Baitho/ Baithiye (respect)/ Baithna (to sit)</td>
<td>Baithnou</td>
</tr>
<tr>
<td>15</td>
<td>Walk</td>
<td>Chalo/ Chaliye (respect)/ Chalna (to walk)</td>
<td>Chalul/ hituul</td>
</tr>
<tr>
<td>16</td>
<td>Eat</td>
<td>Khao/ Khaive (respect)/ Khana (to eat)</td>
<td>Khan</td>
</tr>
<tr>
<td>17</td>
<td>Go</td>
<td>Jao/ Jaaiye (respect)/ Jaana (to go)</td>
<td>Jaan/ janai</td>
</tr>
<tr>
<td>18</td>
<td>Went</td>
<td>Gaye</td>
<td>Jaan/janai</td>
</tr>
<tr>
<td>19</td>
<td>Run</td>
<td>Daudna</td>
<td>Bhagun/daudun</td>
</tr>
<tr>
<td>20</td>
<td>He ate an apple</td>
<td>Usne sev Khaya</td>
<td>Ullsaibkheihaali</td>
</tr>
</tbody>
</table>

8 Conclusion and Future Scope

The field of Named Entity Recognition 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. There are a lot of different applications for natural language processing, such as automatic translation, text processing, information retrieval, voice recognition, and so on. In each of these natural language processing applications, named entity recognition is an essential step in the process. An examination of the relevant literature suggests that hybrid models, which include both rules and a machine learning algorithm, provide superior results when applied to Kumaoni languages. 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. In the future, it would obtain machine learning-based kumaoni information effectively and quickly.

References

22. Allen James, [2003], “Natural Language Understanding”, Pearson Education.
23. Javier Artiles, Enrique Amigo, Julio Gonzalo [2009], The role of named entities in Web People, SIGDAT, a Special Interest Group of the ACL. pages 534-542.


