Artificial Intelligence Review
https://doi.org/10.1007/s10462-021-10093-1

Natural language processing for Nepali text: a review

Tej Bahadur Shahi[1,2] · Chiranjibi Sitaula[3,4]

Abstract

Because of the proliferation of Nepali textual documents online, researchers in Nepal and overseas have started working towards its automated analysis for quick inferences, using different machine learning (ML) algorithms, ranging from traditional ML-based algorithms to recent deep learning (DL)-based algorithms. However, researchers are still unaware about the recent trends of NLP research direction in the Nepali language. In this paper, we survey different natural language processing (NLP) research works with associated resources in Nepali language. Furthermore, we organize the NLP approaches, techniques, and application tasks used in the Nepali language processing using the comprehensive taxonomy for each of them. Finally, we discuss and analyze based on such assimilated information for further improvement in NLP research works in the Nepali language. Our thorough survey bestows the detailed backgrounds and motivations to researchers, which not only opens up new potential avenues but also ushers towards further progress of NLP research works in the Nepali language.

Keywords Devanagari · Machine learning · Nepali language · Nepali linguistics · Natural language processing · Classification · Sentiment analysis

✉ Tej Bahadur Shahi
tejshahi@cdcsit.edu.np

Chiranjibi Sitaula
c.sitaula@deakin.edu.au

[1] School of Engineering and Technology, Central Queensland University, Rockhampton, QLD 4701, Australia

[2] Central Department of Computer Science and Information Technology (CDCSIT), Tribhuvan University, TU Rd, Kirtipur 44618, Kathmandu, Nepal

[3] School of Information Technology, Deakin University, 75 Pigdons Rd, Waurn Ponds, VIC 3216, Australia

[4] Department of Electrical and Computer Systems Engineering, Monash University, Wellington Rd, Clayton, VIC 3800, Australia

Published online: 27 October 2021

Table 1 Nepali numerals, consonants, where among 36 consonants, 33 are unique consonants and 3 are combination of some unique consonants, and vowels (Sitaula et al. 2021)

Numerals	० (0), १ (1), २ (2), ३ (3), ४ (4), ५ (5), ६ (6), ७ (7), ८ (8), ९ (9)
Consonants	क, ख, ग, घ, ङ, च, छ, ज, झ, ञ ,ट, ठ, ड, ढ, ण, त, थ, द, ध, न, प, फ, ब, भ, म, य, र, ल, व,श, ष, स, ह, क्ष, त्र, ज्ञ
Vowels	अ, आ, इ, ई,उ, ऊ, ऋ, ए, ऐ,ओ, औ, अं, अः

1 Introduction

Nepali is an official language of Nepal and spoken primarily in Nepal, India, Myanmar, and Bhutan as well as by Nepalese diaspora worldwide (Khatiwada 2009). It is estimated that the Nepali language is spoken by 17.6 million people around the world (Khanal 2019) . It is based on the Devanagari script, which comprises thirty-six consonants (where thirty-three are unique consonants and three are the combination of unique consonants), thirteen vowels, and ten numerals (refer to Table 1). Similarly, there are several half alphabets in Devanagari scripts apart from such alphabets. The direction of Nepali writing is from left-to-right order and there are no capital letters in Nepali alphabets (Sitaula et al. 2021). We can write the same word in different ways in the Nepali language. This makes automatic language processing more challenging for the Nepali language in comparison to other languages (Prasain et al. 2008).

Natural language processing (NLP) works in the Nepali language can be traced back to 2004 when authors in Bista et al. (2004) introduced a first Nepali lexicon in various file formats with root word, head word, pronunciation, part of speech, synonyms, and idiom for each word. The main purpose of building this lexicon is attributed to several factors such as building spell checker, and machine translation system. In the same year, the first spell checker and machine translation system prototype was introduced by Madan Puraskar Pustakalaya (MPP)[1]. Later on, with the establishment of the Bhasa Sanchar project[2], Nepali National Corpus (Yadava et al. 2008) was developed, which paved a path for further developments of NLP activities in the Nepali language. In the meantime, a large number of research activities were carried out in Nepali language, such as Nepali spell checker (Bista et al. 2004), Nepali Grammar Checker (Bal et al. 2007), Dobhase- A machine translation system (Bista et al. 2005), online Nepali dictionary (Bal 2009), Nepali text to speech (Shah et al. 2018), Nepali Stemmer and morphological analyzer (Bal and Shrestha 2004), and so on.

The NLP works of Nepali documents have been increasing day by day with the prevalent growth of social media worldwide. Furthermore, the increment of social media and Nepali online portals necessitate the use of automated NLP methods for meaningful information extraction over the bulk amount of textual data with less human effort. While analyzing the NLP research works in Nepali language, it can be grouped into two broad groups: NLP approaches and techniques; and NLP application tasks (refer to Fig. 1). For example, several works have been carried out under these groups such as Classification (Sitaula et al. 2021; Subba et al. 2019; Dangol et al. 2018; Basnet and Timalsina 2018; Kafle et al. 2016; Thakur and Singh 2014; Shahi and Pant 2018), Clustering (Sitaula

[1] http://www.mpp.org.np, (accessed date: 02/07/2021).

[2] www.ltk.org.np, (accessed date: 02/07/2021).

Pipeline.PNG

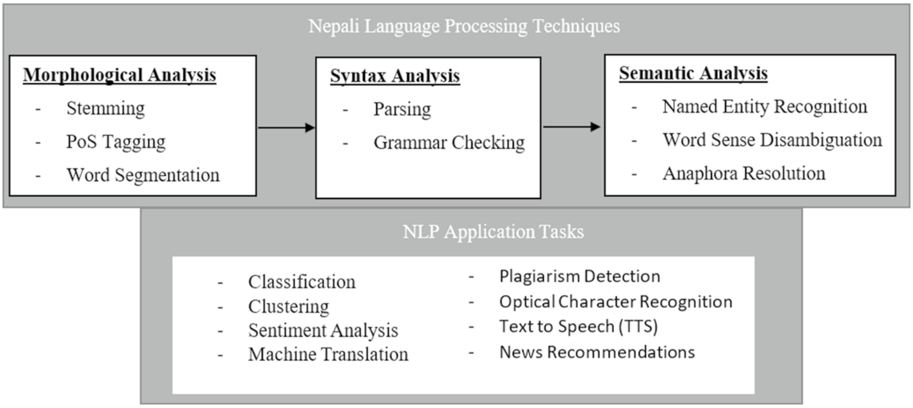


Fig. 1 Overall pipeline of Nepali language processing techniques and applications tasks. Note that PoS denotes Part of Speech tagging

2012; Sarkar et al. 2014), Morphological analyzer and Stemming (Bal and Shrestha 2004; Shrestha and Dhakal 2016), and so on.

Although there are several research works reported in the literature for NLP research in the Nepali language, only the authors in Bal (2009) provided an overview of various NLP resources and applications developed in the Nepali language processing up to 2009. Nevertheless, as the developments of deep learning and machine learning methods significantly raised after 2010, there have been many recent developments in NLP techniques, applications, and approaches in Nepali NLP research works (Singh et al. 2020; Subba et al. 2019; Shahi and Pant 2018; Sitaula et al. 2021) in addition to their work. Thus, there is still a lack of comprehensive survey of existing research with succinct taxonomies so that further research works could be carried out based on them. Given such limitations, this survey paper aims to assimilate existing works with proper taxonomies and presentation, thereby providing the knowledge of previous works and presenting the trends in a systematic way. Also, a thorough study of NLP research in Nepali documents helps understand and identify the gaps in the recent research works to solve the dynamic real life problems in the Nepali linguistic community.

In this paper, we study the detailed NLP research in the Nepali language. This includes the synthesis, analysis, and interpretation of available literature on all levels of Nepali language processing under diversified directions such as NLP approaches, techniques, and applications.

The main **contributions** of our work are as follows:

(i) We provide a survey of the latest developments in Nepali language processing. All stages involved in Nepali language processing such as preprocessing, analysis, synthesis, and applications are comparatively reported in this work.

(ii) We analyze existing advancements, challenges, and limitations in the automated Nepali NLP research works and provide insights on possible solutions.

(iii) We assimilate and categorise the existing research works, which are presented in the form of taxonomies for NLP approaches, techniques, and application tasks in the Nepali language.

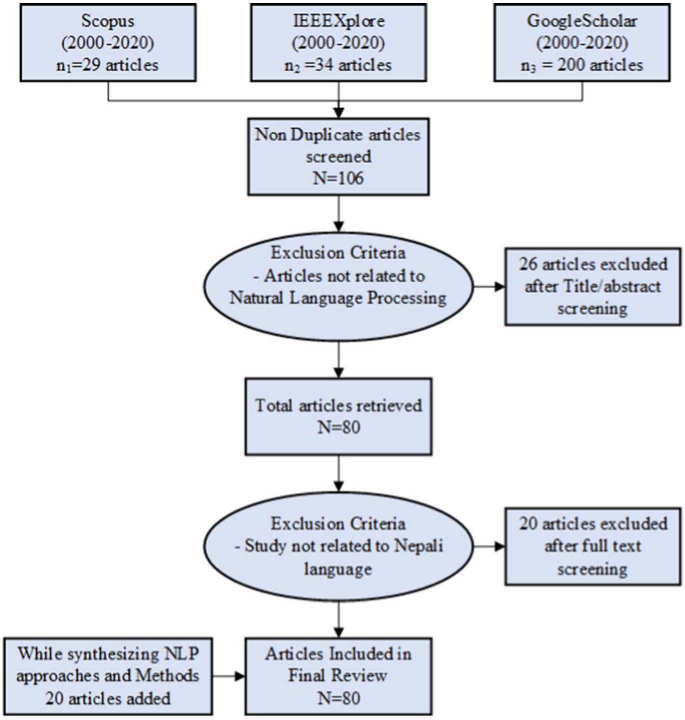

Scopus
(2000-2020)
n_1=29 articles
IEEEXplore
(2000-2020)
n_2 =34 articles
GoogleScholar
(2000-2020)
n_3 = 200 articles
Non Duplicate articles screened
N=106
Exclusion Criteria
- Articles not related to Natural Language Processing
26 articles excluded after Title/abstract screening
Total articles retrieved
N=80
Exclusion Criteria
- Study not related to Nepali language
20 articles excluded after full text screening
While synthesizing NLP approaches and Methods 20 articles added
Articles Included in Final Review
N=80

Fig. 2 Step-wise procedure to retrieve the articles reviewed in this survey

(iv) We suggest the potential avenues of NLP research works for the Nepali language.

The paper is organized as follows. Section 2 explains the methods used to carry out this survey. Furthermore, Sect. 3 presents the publicly available datasets used in the NLP research in the Nepali language. Likewise, Sect. 4 presents the preprocessing techniques used in Nepali NLP research works. Similarly, Sect. 5 details the approaches and techniques that been have reported so far in existing NLP research. Section 6 elaborates different application areas of NLP research in Nepali language. Furthermore, Sect. 7 discusses the overall works and observes the research issues and challenges accordingly. Finally, Sect. 8 concludes the paper with future recommendations.

2 Survey method

To identify the articles for review, we first design a query string using the terms (“Natural Language Processing” OR “NLP”) AND (“Nepali” OR “Nepali Language”) and perform database search on three popular databases (IEEE Xplore, Scopus, and Google Scholar) limiting our search within the title, abstract, and keywords of each article on February 10, 2020. As a result, we achieve 29, 34, and 200+ articles from Scopus, IEEE Xplore, and Google Scholar, respectively for such query string in each of them. After carefully reading the titles, abstracts, keywords, and full-text, we end up selecting only 60 articles related to our work. Later, while synthesizing NLP approaches and techniques in various themes (refer to Sect. 5), we review additional 20 articles that are not directly related to NLP but useful. Hence, 80 articles are considered for the final review in this study. The detailed pipeline of our survey method is presented in Fig. 2.

3 Datasets

In this section, we provide a detailed list of publicly available datasets used in Nepali language processing. These datasets have been used in different tasks, such as Stemming, Part of Speech (PoS) tagging, Named entity recognition, Word sense disambiguation, Anaphora resolution, Classification, Clustering, and Machine translation.

Yadava et al. (2008) in NELRALEC (Nepali Language Resources and Localization for Education and Communication) project, developed a Nepali National Corpus (NNC), containing more than 14 million Nepali words, which was further divided into three subsets– a) Nepali Monolingual Written Corpus, which contains two types of texts: the core sample and general texts. The core sample comes from 15 different genres and contains 802,000 words. The general text has been collected from various sources such as websites, newspapers, and books, which contain 1,400,000 words; b) English Nepali Parallel Corpus, which was designed to enable the machine translation task. It was aligned at sentence-level and document-level manually. There are 27,060 English words aligned with 21,756 Nepali words at sentence level. However, a larger set of words are aligned at the document level (617,340 English words; 596,571 Nepali words); and c) Nepali Spoken Corpus, which contains audio recordings taken in the natural setting of social activities with phonologically transcribed and annotated texts. It was designed for the Speech-to-Text (STT) domain of research in Nepali language[3].

Choudhary and Ramamoorthy (2019) at Central Institute of Indian Language released a gold standard Nepali raw text corpus. This corpus contains nearly 7 million words from six domains: Aesthetics (57.71%), Commerce (0.43%), Mass media (32.18 %), Official documents (0.03%), Science and Technology (1.14%), and Social Science (8.51%). However, this dataset is accessible only after registration[4].

Lamsal (2020) compiled a Nepali text corpus with nearly 90 million running words from various online news websites. It contains text from several sources such as News, Finance, Sports, Entertainment, Health, Technology, and Literature. This is merely a collection of raw texts with a little preprocessing such as removal of HTML tags, English alphabets and numerals, and unnecessary spaces. Additionally, they also provided pre-trained word embedding vectors of around 0.5 million words for further analysis. This dataset can be found at the following repository[5].

16NepaliNews[6] collected 14, 364 documents under 16 classes, where each class contains at least 16 documents. The names of classes in this dataset are as follows: Auto, Bank, Blog, Business Interview, Economy, Education, Employment, Entertainment, Interview,

[3] http://www.elra.info/en/catalogues/free-resources/nepali-corpora/ (accessed date: 17/02/2021).

[4] https://data.ldcil.org/a-gold-standard-nepali-raw-text-corpus (accessed at 17/02/2021).

[5] https://ieee-dataport.org/open-access/large-scale-nepali-text-corpus (accessed date: 16/02/2021).

[6] https://github.com/sndsabin/Nepali-News-Classifier (accessed date: 17/01/2021), Information and Language Processing Research Lab, Kathmandu University, Nepal.

Literature, National News, Opinion, Sports, Technology, Tourism, and World. This dataset is normally used for classification.

NepaliNewsLarge Shahi and Pant (2018) comprises 7, 023 documents under 20 news categories, where each class contains 111 to 700 documents. The names of classes in this dataset are Agriculture, Automobiles, Bank, Blog, Business, Economy, Education, Employment, Entertainment, Health, Interview, Literature, Migration, Opinion, Politics, Society, Sports, Technology, Tourism, and World. This dataset is also used for the classification, which can be found at the following repository[7].

NepaliLinguistic Sitaula et al. (2021) contains 17 news classes, which are Agriculture, Auto, Bank, Blog, Business, Economy, Education, Employment, Entertainment, Health, Interview, Literature, Migration, National News, Opinion, Politics, Society, Sports, Technology, Tourism, and World. This dataset is used for the classification purpose, and can be found at the following repository[8].

The brief summary of publicly available datasets and their associated NLP task is reported in Table 2. There are many other datasets mentioned in the literature, but their sources are not publicly available. So, we exclude them in this section.

4 Preprocessing

The data in a raw corpus contains various unnecessary characters and words that do not contribute much to the NLP research pipeline. Filtering out those noisy data speeds up and simultaneously improves the results (Singh and Gupta 2017). The following prepossessing steps are being commonly used in preparing Nepali dataset/corpus:

Document sanitization: When the dataset is prepared by scrapping the content from the web, the scrapper will also extract unnecessary HTML tags, zero width joiners, punctuation marks, and other unnecessary characters and symbols. These are removed to sanitize the document using a white list containing user-defined characters and symbols.

Tokenization: This step breaks each individual document in the corpus into tokens or words (Shahi and Shakya 2018) that can be used directly. Since Nepali words are separated with a space, it is easy to segment the words in a sentence as shown in Table 3.

Stop words removal: The words that appear with a very high frequency in the text are considered as stop words and normally they neither contribute nor show negligible contribution in knowledge representation and thus they are removed from further consideration. Several rich resource languages (e.g., English) have a standard pre-defined list of stop words. However, for NLP works in the Nepali language, authors have prepared their own stop words list in their work. Thus, there is a lack of the standard stop words list for the Nepali language. The list of commonly used Nepali stop words is shown in Fig. 3 (Sitaula et al. 2021).

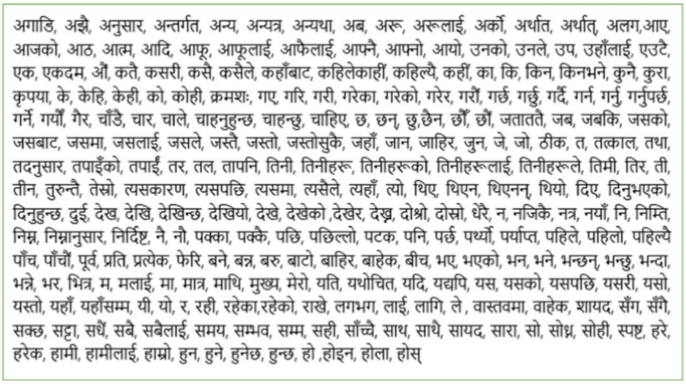
Stemming: Stemming is the process of removing affixes from words. Affixes may be either inflectional or derivational. In Nepali, the meaning of compound words created using derivational affixes are often very different from the root or stem words (refer to Sect. 5.2.1).

[7] https://www.kaggle.com/ashokpant/nepali-news-dataset-large (accessed date :16/02/2021).

[8] https://ieee-dataport.org/documents/nepaliliinguistic (accessed date: 16/02/2021).

Table 2 Available datasets for various NLP research works in Nepali language

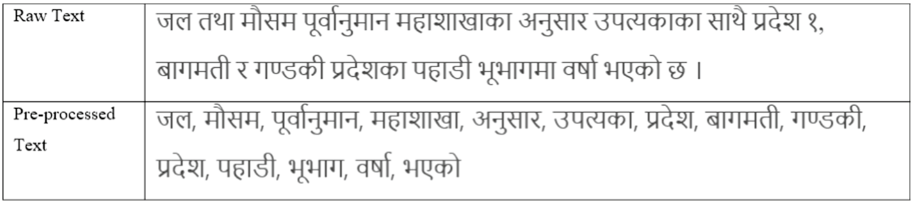
Refs.	Dataset	NLP task	# of samples	Brief summary
Yadava et al. (2008)	Nepali National Corpus	PoS tagging, Parsing and Machine Translation	14 million words	Consist of three sub-corpus: written corpus, spoken corpus and parallel corpus Part of speech annotated Nepali English aligned sentences
Lamsal (2020)	Text Corpus	Word embedding vectors	90 million words	Pre-trained word embedding vectors for deep learning based NLP
Sitaula et al. (2021)	Nepali News dataset	Document Classification/ Clustering	35651 documents	Based on Nepali news documents Large number of documents labeled in 17 news categories
Choudhary and Ramamoorthy (2019)	Nepali Raw text corpus	Various NLP-related tasks	7 million words	Data statistics provided Text from six domains
Senapati et al. (2020)	Annotated anaphoric relation	Anaphora resolution	4,700 words	Well-described data format Text from three domains: short stories, blogs, and news
Dhungana and Shakya (2014)	Nepali WordNet	Word sense disambiguation	348 words	Only 59 words are polysemy words
Singh et al. (2019)	NepaliNER	Named entity recognition	79,087 entities	Large NER dataset till date Divided into train, test and evaluation sets
Singh et al. (2020)	NepSA	Sentiment analysis	4,035 sentences	User comments from YouTube channels were used as data source Inter-annotator agreement was used for labelling



अगाडि, अझै, अनुसार, अन्तर्गत, अन्य, अन्यत्र, अन्यथा, अब, अरू, अरूलाई, अर्को, अर्थात, अर्थात्, अलग,आए, आजको, आठ, आत्म, आदि, आफू, आफूलाई, आफैलाई, आफ्नै, आफ्नो, आयो, उनको, उनले, उप, उहाँलाई, एउटै, एक, एकदम, औं, कतै, कसरी, कसै, कसैले, कहाँबाट, कहिलेकाहीं, कहिल्यै, कहीं, का, कि, किन, किनभने, कुनै, कुरा, कृपया, के, केहि, केही, को, कोही, क्रमशः, गए, गरि, गरी, गरेका, गरेको, गरेर, गरौं, गर्छ, गर्छु, गर्दै, गर्न, गर्नु, गर्नुपर्छ, गर्ने, गर्यौं, गैर, चाँडै, चार, चाले, चाहनुहुन्छ, चाहन्छु, चाहिए, छ, छन्, छु,छैन, छौँ, छौं, जताततै, जब, जबकि, जसको, जसबाट, जसमा, जसलाई, जसले, जस्तै, जस्तो, जस्तोसुकै, जहाँ, जान, जाहिर, जुन, जे, जो, ठीक, त, तत्काल, तथा, तदनुसार, तपाइँको, तपाईं, तर, तल, तापनि, तिनी, तिनीहरू, तिनीहरूको, तिनीहरूलाई, तिनीहरूले, तिमी, तिर, ती, तीन, तुरुन्तै, तेस्रो, त्यसकारण, त्यसपछि, त्यसमा, त्यसैले, त्यहाँ, त्यो, थिए, थिएन, थिएनन्, थियो, दिए, दिनुभएको, दिनुहुन्छ, दुई, देख, देखि, देखिन्छ, देखियो, देखे, देखेको ,देखेर, देख्न, दोश्रो, दोस्रो, धेरै, न, नजिकै, नत्र, नयाँ, नि, निम्ति, निम्न, निम्नानुसार, निर्दिष्ट, नै, नौ, पक्का, पक्कै, पछि, पछिल्लो, पटक, पनि, पर्छ, पर्थ्यो, पर्याप्त, पहिले, पहिलो, पहिल्यै पाँच, पाँचौं, पूर्व, प्रति, प्रत्येक, फेरि, बने, बन्न, बरु, बाटो, बाहिर, बाहेक, बीच, भए, भएको, भन, भने, भन्छन्, भन्छु, भन्दा, भन्ने, भर, भित्र, म, मलाई, मा, मात्र, माथि, मुख्य, मेरो, यति, यथोचित, यदि, यद्यपि, यस, यसको, यसपछि, यसरी, यसो, यस्तो, यहाँ, यहाँसम्म, यी, यो, र, रही, रहेका,रहेको, राखे, लगभग, लाई, लागि, ले , वास्तवमा, वाहेक, शायद, सँग, सँगै, सक्छ, सट्टा, सधैं, सबै, सबैलाई, समय, सम्भव, सम्म, सही, साँच्चै, साथ, साथै, सायद, सारा, सो, सोध्न, सोही, स्पष्ट, हरे, हरेक, हामी, हामीलाई, हाम्रो, हुन, हुने, हुनेछ, हुन्छ, हो ,होइन, होला, होस्

Fig. 3 List of commonly used stop words in Nepali text processing (Sitaula et al. 2021)

Table 3 Illustration of preprocessing tasks of Nepali text



Raw Text	जल तथा मौसम पूर्वानुमान महाशाखाका अनुसार उपत्यकाका साथै प्रदेश १, बागमती र गण्डकी प्रदेशका पहाडी भूभागमा वर्षा भएको छ ।
Pre-processed Text	जल, मौसम, पूर्वानुमान, महाशाखा, अनुसार, उपत्यका, प्रदेश, बागमती, गण्डकी, प्रदेश, पहाडी, भूभाग, वर्षा, भएको

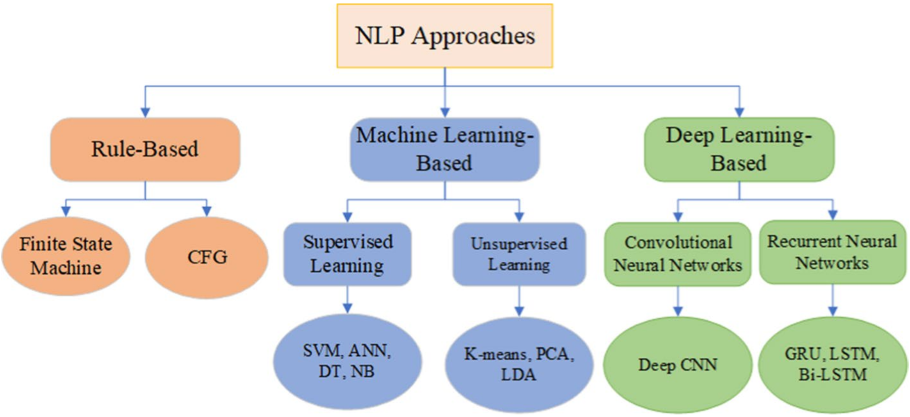


Fig. 4 Various approaches used in Nepali NLP works

5 Approaches and techniques

Here, we discuss approaches and techniques using the corresponding high-level taxonomy.

5.1 Approaches

In Nepali NLP applications, various approaches have been proposed (refer to Fig. 4). Generally, these methods are categorised into three groups: Rule-based, Machine learning-based, and Deep learning-based. Rule-based approaches are the preliminary methods introduced by various computational linguists for basic NLP techniques such as Part of speech tagging (Shahi et al. 2013), Morphological analysis (Bal and Shrestha 2004), Grammar checking (Bal et al. 2007), and so on. Once the resources such as annotated corpora, dictionary, and lexicon were developed, the automated machine learning-based approaches were used for various applications of NLP research such as Classification, Clustering, Machine translation, Information retrieval, Speech to text translation, Text summarization, and so on. Recently, the deep learning-based approaches, especially the variant of recurrent neural network (RNN), have been used in NLP research of the Nepali language.

Rule-based approaches formulate the linguistic rules or patterns to address various NLP tasks. The low-level NLP tasks such as Stemmer, Morphological analyzer, and Parsing were implemented using rule-based techniques. Authors in Bal and Shrestha (2004) implemented the first Morphological analyzer with Nepali Stemmer, which was able to strip morphemes and add their grammatical category. They used two separate sets of rules for striping suffix and prefix from the word. Similarly, Gupta and Bal (2015) proposed a Nepali SentiWordNet, called Bhawanakosh, by pulling all words from English SentiWordNet and translated them into the Nepali language. They looked up each input word into Bhawanakosh to detect its subjectivity. However, rule-based cost for their maintenance. For example, the Stemming rules need to be updated once the corpus changes and few new words are added. Also, linguistic and language-specific knowledge are necessary to formulate such rules, which require language experts mostly.

Machine learning-based approaches are data-driven approaches, where decision making rules are learned from the given dataset. Here, the supervised machine learning methods are dominant to address various NLP tasks. Supervised learning techniques consist of two phases: training and testing. In the training phase, we learn the decision making rules or parameters and apply these rules to new incoming samples during the testing phase. This approach has advantages over rule-based techniques as it can learn required parameters to address NLP tasks from corpus/dataset. However, methods based on it are largely dependent on a large amount of training corpus to achieve satisfactory performance. Under this approach, authors in Shahi et al. (2013) used a Support Vector Machine (SVM) algorithm to classify the words into a different part of speech categories. They used the Nepali Part of Speech-tagged corpus to train the model, and achieved an accuracy of 91.07% for the known words and 89.56 % for the unknown words. This result shows that the accuracy of the proposed method is largely dependent on the training data size.

Recently, the deep learning-based approaches have been successfully implemented for various NLP applications such as Classification, Clustering, Machine translations, Text summarization, and so on in high-resource languages (e.g., English). Following its massive growth, a few works on deep learning-based NLP for the Nepali language have been undertaken. However, these methods are still in infancy in the Nepali language as we do

Table 4 Approaches and techniques of NLP works in Nepali language. Note that the abbreviated notations used in the table are as follows: RB (Rule-based methods), ML (Machine Learning-based methods), DL (Deep learning-based methods), Stem. (Stemming), POS (Part Of Speech Tagging), NER (Named Entity Recognition), MA (Morphological Analyser), WSD (Word Sense Disambiguation), and AR (Anaphora Resolution). Similarly, the symbols ✓ and ✗ denote the presence and absence of the corresponding approach or technique, respectively

Year	Approaches			Techniques					
	RB	ML	DL	Stem.	POS	NER	MA	WSD	AR
2004	✓	✗	✗	✓	✗	✗	✗	✗	✗
2008	✓	✗	✗	✗	✗	✗	✗	✓	✗
2012	✓	✗	✗	✗	✗	✗	✓	✗	✗
2013	✓	✓	✗	✓	✓	✗	✗	✗	✗
2014	✓	✓	✗	✗	✗	✓	✗	✓	✓
2015	✓	✗	✗	✗	✗	✗	✓	✓	✗
2016	✓	✓	✗	✗	✗	✗	✗	✗	✗
2017	✗	✓	✗	✗	✗	✗	✗	✗	✗
2018	✗	✓	✓	✗	✓	✗	✗	✗	✗
2019	✗	✓	✓	✗	✗	✓	✗	✗	✗
2020	✗	✓	✓	✗	✗	✗	✗	✗	✓
2021	✗	✗	✓	✗	✗	✗	✗	✗	✗

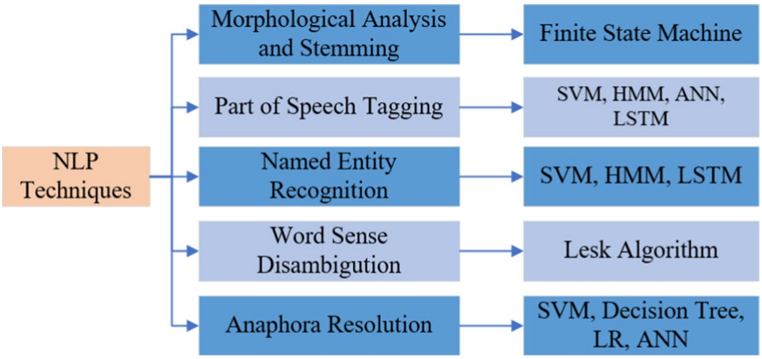


Fig. 5 Taxonomy of NLP techniques with algorithms used in Nepali language processing

not have sufficient annotated resources for training purpose, which is shown in Acharya and Bal (2018).

The brief historical developments of NLP methods along with various NLP techniques used for Nepali language processing are summarized in Table 4. During the years 2004-2015, Rule-based approaches dominated all kind of NLP works, whereas Machine learning and deep learning-based approaches have been a hot topic in recent Nepali NLP literature.

5.2 Techniques

Here, we analyze different NLP techniques used in Nepali textual documents, including Morphological analysis and Stemming, Part of Speech tagging, Named entity recognition, Word sense disambiguation, and Anaphora resolution. The high-level taxonomy representing the NLP techniques along with algorithms is presented in Fig. 5.

5.2.1 Morphological analysis and stemming

The Morphological analyzer and Stemmer are used to break down each word into the basic linguistic unit. The Nepali language is rich in its morphology as it supports various kinds of suffixes, prefixes, and affixes. The first morphological analyzer was developed by Bal and Shrestha (2004), which is based on linguistic rules of the Nepali language. This work has limited applications because it is based on a fixed set of suffixes and prefixes dictionary for both stemming and morphological analysis. Recently, Shrestha and Dhakal (2016) proposed a Stemmer with a large set of suffixes (total 128 suffixes), which leveraged the rule-based stemming algorithm with nearly 5000 words. It only considers suffixes for stemming, which in result leads to over-stemming problem. Furthermore, a computational method for stemming was proposed by Sitaula (2013) to overcome the over-stemming issues, by using a context free grammar and similarity matching technique. After that, a finite state machine-based approach was implemented by Borah et al. (2017) to perform morphological analysis of non-declinable adjectives of Nepali language. In their method, authors reported that the proposed method works with minimal linguistic resources, which was tested with amass of Nepali text documents. However, their work is involving as it first translates the Nepali text into Roman and then, uses a pattern matching approach to find the location of noun in a given sentence. Once the noun is found, they use grammatical rules to tag the word before or after it with an adjective. Similarly, Chhetri et al. (2015) proposed a finite state machine-based method, which splits a noun into its morphemes using a suffix dictionary. Nevertheless, the proposed method only deals with inflectional form of the noun. Another study in Prasain (2008) developed a computational model based on finite state automata to analyze basic verbs in the Nepali language. Last but not the least, a Nepali inflectional morphological analyzer was proposed by Bhat et al. Bhat and Rai (2012) using the finite state approach with Apertium interface[9].

5.2.2 Part of speech (PoS) tagging

PoS tagging is the process of assigning a linguistic tag or meaning to a word such as noun, verb, and pronoun. To automate the tagging process, standard PoS tagset needs to be introduced. PoS tagging can be implemented by three approaches: rule-based approach, statistical approach, and machine learning (or deep learning)-based approach. The rule-based approach is primarily based on manually crafted rules to classify each word into one of the PoS tags such as noun, verb, pronoun, and so on. This process is labour-intensive, which is mostly prone to the human-error. Also, the linguistic expert is required to craft the rules. Similarly, the statistical taggers are based on the estimated probability of tags from the training corpus, whereas the machine learning-based approach estimates the various parameters to classify each word into one of the specified tags using supervised or unsupervised learning.

The first tagset for the Nepali language was introduced by Prasain et al. (2008). Initially, the tagset consisted of 112 tags, which was used to tag Nepali National Corpus (NNC) semi-manually (both human and automated approaches). Later on, due to the larger tagset size, the error rate of annotation surged higher, which let them reduce to 43 tags only. After their pioneering work, scientists started devising the automated tagger in Nepali NLP

[9] http://xixona.dlsi.ua.es/~fran/apertium2-documentation.pdf (accessed date: 13/02/2021).

research. First, the rule-based tagger, also known as Unitag, was introduced by Bal and Shrestha (2004) to automate the PoS tagging for Nepali written corpus, which is rule-based and manually annotated. Rule-based PoS tagger has limitations such as they can't be generalized to other words that don't fit into the specified rule. To this end, it is more static and each time if the model encounters a new word, the new rule needs to be defined. Given such issues, authors in Prajwal et al. (2008) implemented a TnT– a statistical tagger based on the theory of probability and trained on a large annotated corpus of around 82,000 words and 42 PoS tags. The reported accuracy for this tagger is 97% and 56% for known and unknown words, respectively. Further improvement on PoS tagging was reported by Shahi et al. (2013) using the Support Vector Machine (SVM)-based PoS tagger trained on the Nepali national corpus(NNC). They compared the performance of the SVM-based tagger with TnT-a statistical tagger. It shows that the SVM-based tagger performs well on both known and unknown words. . Furthermore, authors in Yajnik (2017) attempted to develop a statistical tagger for Nepali text based on Hidden Markov Model (HMM) and Viterbi algorithm. They achieved an accuracy of 95.43% with their own PoS tag-set. Similarly, another statistical tagger was proposed by Paul et al. (2015), which was trained with a corpus containing 42 tags and 1,50,839 words in the Nepali language. The accuracy of their work for known words was quite impressive, however, the authors didn't report the accuracy for unknown words.

Furthermore, the Neural network-based PoS taggers were also proposed in Yajnik (2018), Prabha et al. (2018). Authors in Yajnik (2018) extracted the features from the marginal probability of Hidden Markov model, which was processed by three different neural network architectures: General Regression ANN, Feed Forward ANN, and Radial Bias ANN. Similarly, authors in Prabha et al. (2018) implemented the variant of recurrent neural networks (RNN): Vanilla RNN, LSTM (Long Short-Term Memory), and GRU (Gated Recurrent Unit). In both works, the accuracy is almost 99%, which outperforms the statistical and rule-based tagger's performance. However, the datasets and models used by Yajnik (2017), Prabha et al. (2018) are not publicly available, therefore, they are difficult to use for practical applications. In contrast, there is a chance of over-fitting while using Neural network-based models when we have insufficient data. An allusive discussion on existing PoS taggers for Nepali text along with their advantage and limitation are reported in Table 5.

5.2.3 Named entity recognition (NER)

Another important task of NLP is Named entity recognition (NER), which attempts to classify specific words of the sentence into pre-defined names such as organization, people, or location (Ekbal and Bandyopadhyay 2008). Such information is useful in information retrieval (IR)-related applications. The first task implemented for the Nepali language was by Bam (2014) using self-created NER corpus. Their preliminary result shows that the proper feature extraction method can be beneficial for automated NER using machine learning methods such as Support Vector Machine. They also listed out the challenges of Nepali NER systems such as no-capitalization in names, agglutinate nature of names, name ambiguity and loan words or imported words in the Nepali language from other languages. Similarly, a hybrid approach combining Hidden Markov Model (HMM) and rule-based system for NER was implemented in Dey et al. (2014). They first assigned the PoS tag to each word using HMM, which was used to find the named entity for a given word from a look-up table. Recently, the deep learning-based NER for the Nepali language was proposed by Singh et al. (2019) with the enhanced NER corpus. They used several concepts

Table 5 Comparative study of different PoS taggers for Nepali text. Note the abbreviation used in table are as follows: SVM(Support Vector Machine), HMM (Hidden Markov Model), TnT (Trigram's and Tag), RNN (Recurrent Neural Network), LSTM (Long Short-Term Memory), CNN (Convolution Neural Network), and NN (Neural Network)

Refs.	Features	Methods	Advantages	Limitations
Shahi et al. (2013)	Word feature	SVM	The effect of data size on performance of tagger is analyzed	Word Context is not considered
	PoS feature	TnT	Both known and unknown words are considered	
	Word bi-gram			
	PoS bi-gram			
Yajnik (2017)	Not specified	HMM	Hybrid approach	Limited performance
		Viterbi		
Paul et al. (2015)	Not specified	HMM	Statistical tagger	Very Limited performance for unknown word
			Consider both known and unknown words	
Prabha et al. (2018)	Word embedding	RNN	Deep Neural network first implemented	Unknown words were not considered
		LSTM		
		CNN		
		CNN and LSTM		
Yajnik (2018)	Emission and transition probability features	Feed Forward NN	Exploited Neural Network	Word context and unknown words were not considered
		Radial bias NN		
		General Regression NN		

Table 6 Comparative study of different NER for Nepali text. Note that the generic Nepali named entity types used as abbreviation in the table are as follows: PER (per person), LOC (locations such as places, cities, countries etc.), ORG (Organizations such as companies, government organizations, agencies, etc.), and MISC (Miscellaneous)

Refs.	NER class	Features	Methods	Advantages	Limitations
Bam (2014)	PER	Gazetter list	SVM	Simple to implement	Word context are not considered
	LOC	First word		First work in Nepali NER	
	ORG	Word length			
	MISC	Digit features			
	Other				
Dey et al. (2014)	PER	POS tag	Rule based + HMM	Hybrid approach	Limited features
	LOC	N-gram		More NER classes	
	ORG				
	Number				
	Currency				
	Quantifier				
	Unknown words				
Singh et al. (2019)	PER	Character embedding	BiLSTM and CNN	Embedding features	Need to expand data size
	LOC	Grapheme embedding		Well-tuned model hyper-parameters	Limited word embedding model used
	ORG	POS encoding		First Neural-based Nepali NER	
	MISC				
	Other				

such as word features, PoS features, and Grapheme features to perform the NER classification, which imparts 86.89 % accuracy. A comparison of existing methods and techniques for Nepali NER with respect to various features considered in each study and their performance is summarized in Table 6.

5.2.4 Word sense disambiguation

In natural language, same word can express more than one meaning when used in a different context. These words create an ambiguity at various levels of language processing. The process of finding or differentiating the meaning of the same word in various contexts is simply known as word sense disambiguation (WSD). WSD is fundamental to many NLP applications such as information retrieval (Zhong and Ng 2012), sentiment analysis (Hung and Chen 2016), and machine translation (Carpuat and Wu 2007). For instance, the meaning of the particular word needs to be identified in the given context for its correct translation to another language. Primarily, WSD can be done in two ways: lexical sample word-based WSD, which focuses on some particular words for disambiguation; and all word WSD, which considers each word in the given input document (Wang et al. 2020). There are two approaches commonly used to tackle the WSD problems: Knowledge-based approach and Machine learning-based approach. The knowledge-based approach uses NLP resources such as dictionary, ontology, and WordNet to find the correct meaning of word in specific context, whereas the machine learning-based approach recasts the disambiguation problem into classification problem (Bhala and Abirami 2014). For this, authors in Dhungana and Shakya (2014), Shrestha et al. (2008) attempted Nepali WSD based on Lesk algorithm (Lesk 1986) under the knowledge-based approach. They considered 348 words in their experiments, where 59 words were Polysemy. They used self-created Nepali WordNet, which contains synset, gloss, example, and hypernym for each target word. Such information was used while collecting the context words and semantic similarity of each pair of words. There are no published works on machine learning-based WSD for Nepali language.

5.2.5 Anaphora resolution

In many applications of NLP works such as machine translation, question answering, and information extraction, it is required to identify the expression referring to the same entity throughout the text, known as co-reference or anaphora resolution. Mostly, the identification of expression that refers to same entity in backward direction was performed by many researchers in literature. The first and recent work of anaphora resolution task in the Nepali language was done by Senapati et al. (2020) with a self-created anaphora-annotated dataset, where they used both language-dependent features such as prefix and suffix inflection in nouns; and language-independent features such as inter-sentence distance, and word distance with four machine learning algorithms: SVM, Decision tree, Logistic regression and Neural network. SVM algorithm is found to have the best-performing model for the anaphora resolution tasks. Also, authors in Shrestha and Bal (2020) proposed an anaphora resolution for named entities such as personal pronouns with hand-crafted rules. This work was based on Lappin and Leass' algorithm, which resolves the co-reference of named entity by calculating recency and degree of saliency for each entity in a sentence (Lappin and Leass 1994). They reported that their method provides an accuracy of 87.45% while testing on 1460 Nepali sentences taken from news articles.

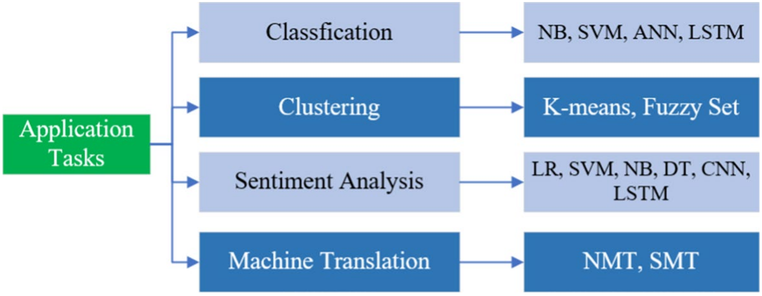


Fig. 6 Taxonomy of NLP Application tasks and algorithms used in Nepali language processing

6 Application tasks

The overall application tasks infer the NLP research works that have been applied in different areas, for example, Classification, Clustering, Sentiment analysis, etc. The high-level taxonomy of application tasks is presented in Fig. 6.

6.1 Classification

Classification is defined as the categorization of documents based on pre-defined labels. This is also called supervised learning. For the classification purpose, we need to represent each document into the mathematical format, called document features. For representation and classification purposes, there have been some works (Sitaula 2014; Subba et al. 2019; Dangol et al. 2018; Basnet and Timalsina 2018; Kafle et al. 2016; Thakur and Singh 2014; Shahi and Pant 2018; Shahi and Yadav 2014) done in Nepali document representation and classification. Initially, Sitaula (2014) identified and classified the semantic orientation of Nepali textual documents using a finite state machine. Although their model exploits the finite state machine to identify and classify the polarity of sentiments, their model has a limited performance. Thakur and Singh (2014) employed Bag-of-Words (BoW) method to represent Nepali documents for the classification using Lexicon pooling with Naive Bayes algorithm. Kafle et al. (2016) did a comparative study of TF-IDF and Word2Vec method (Mikolov et al. 2013) for the classification. Similarly, Shahi and Pant (2018) utilized the TF-IDF method to extract the document features of Nepali news documents and performed classification on news documents using three algorithms: Support Vector Machine (SVM) (Cristianini and Shawe-Taylor 2000), Naive Bayes (NB) method (Lewis 1998), and Artificial Neural Networks (ANN). Moreover, Dangol et al. (2018) leveraged the N-gram model (Brown et al. 1992) of tokens or words in the news documents and established the term-document matrix during the classification. To get the benefits of the recent advantage of deep learning and embedding, Basnet and Timalsina (2018) achieved the document features using Word2Vec model (Mikolov et al. 2013) and classified the documents using the LSTM (Long Short-Term Memory) model. Furthermore, Subba et al. (2019) adopted Bag-of-Words (BoW) (Salton and McGill 1983) method to extract the document features for the Nepali news document representation, which was trained using the deep learning model. Recently, Sitaula et al. (2021) adapted the supervised codebook approach for the representation of textual documents in the Nepali language. It shows that their method outperforms

Table 7 Advantages and disadvantages of different existing methods for Nepali document representation and classification. Note that the abbreviation used in table are as follows: BoW (Bag of words), TF-IDF (Term Frequency-Inverse document frequency), LSTM (Long Short-Term Memory), ANN (Artificial Neural Network), SVM (Support Vector Machine), and RNN (Recurrent Neural Network)

Refs.	Approach	Advantages	Disadvantages
Thakur and Singh (2014)	BoW+Naive Bayes	Easy and simple for the implementation Works for all kinds of documents	Limited classification performance
Kafle et al. (2016)	TF-IDF+Word2Vec	Simple and easy to use for the representation of documents Works for all kinds of documents	Limited classification performance
Basnet and Timalsina (2018)	Word2Vec + LSTM	Follows sequence of tokens that captures semantics of words	Model seems over-fitting Needs extensive experiment to tune hyper-parameters (e.g., LSTM)
Shahi and Pant (2018)	TF-IDF+SVM+ANN	Simple and easy to implement Works for all kinds of text documents	Limited performance
Dangol et al. (2018)	N-gram	Provides the semantics of words using *n*-gram model Provides the improved performance than BoW method	The *n*-gram increases the computational complexity significantly and also difficult to select the optimal *n* parameter
Subba et al. (2019)	BoW+RNN	Shows the semantics of tokens using RNN Outperforms the traditional ANN model	Needs extensive work to tune the best architecture of RNN Provides the limited performance
Sitaula et al. (2021)	Supervised codebook	Imparts higher classification accuracy Computationally efficient for classification	Domain-specific Design of supervised codebook is time consuming

several machine learning algorithms such as TF-IDF (Term Frequency-Inverse Document Frequency), TF (Term Frequency), BoW (Bag of Words), and BERT (Bidirectional Encoder Representations from Transformers).

Furthermore, we compare and summarize these recent methods that have been used in Nepali text document representation and classification in Table 7.

6.2 Clustering

Text clustering is a process of creating groups of semantically similar documents. To perform the text clustering, document representation is essential. Representing documents with their semantics is a challenging problem, especially in NLP research. The most simple and popular document representation is ’bag of words’-based method, which has been used in both Classification and Clustering. For Clustering in Nepali language, Sitaula (2012) leveraged clustering of the Nepali documents using enhanced vector space model and fuzzy set theory. Despite its simplicity, the method is neutral to word semantics. However, it is important to preserve semantic meanings while representing the document. For this, Wordnet (Miller 1995), a collection of words where each word is linked with its synonyms in the form of a graph, is another way of representing document (Sarkar et al. 2014). In Sarkar et al. (2014), Wordnet-based document representation was used along with K-means clustering (MacQueen 1967) and particle swarm optimization (Poli et al. 2007) for Nepali document clustering. Their work preserves the semantics of words in documents to some degree.

6.3 Sentiment analysis

Sentiment analysis is an automated process to classify the opinions expressed into different classes such as positive, negative or neutral. Sentiment classification usually is implemented at three levels: document level, sentence level and aspect level. There have been a few works done on sentiment analysis for the Nepali language. The first sentiment analysis work for Nepali text was proposed by Gupta and Bal (2015) on a dataset of size 25,435 sentences, collected from various online news portals such as Kantipur[10], and Nagariknews[11]. They compared the resource-based bootstrap approach and machine learning-based approach for sentiment analysis, which shows that the machine learning-based approach performs better than other approaches for sentence level sentiment classification. Similarly, the document level sentiment analysis was proposed in Thapa and Bal (2016) for book and movie reviews into positive and negative classes. They used the TF-IDF and bag-of-words (BOW) technique for document representation to train three machine learning-based algorithms, namely, SVM, Multinomial Naive Bayes, and Logistic Regression. Their method shows that the Multinomial Naive Bayes method outperforms all contending methods. However, it has a limited data size of 179 positive and 205 negative samples to train the models. Similarly, the analysis of Nepali subjective text to distinguish between facts and opinion was carried out by Regmi et al. (2017), where features such as bi-gram and TF-IDF were utilized. They used three machine learning models to compare the performance on Nepali subjective text and found that SVM with bi-gram features outperform other

[10] https://ekantipur.com/ (accessed date: 13/02/2021).

[11] https://nagariknews.nagariknetwork.com/ (accessed date: 13/02/2021).

Table 8 Comparative study of different Sentiment analysis approaches for Nepali text. Note that the abbreviation used in table are as follows: SA (sentiment analysis),TF-IDF (Term Frequency-Inverse Document Frequency), BoW (bag of words), SVM (Support Vector Machine), NB (Naive Bayes), LSTM (Long Short-Term Memory), and CNN (Convolution Neural Network)

Refs.	Features	Methods	Advantages	Limitations
Gupta and Bal (2015)	BoW	SentiNetWord	First work on SA for Nepali text	Only preliminary results
		Naive Bayes	Creation of SentiWordNet for further research	
Thapa and Bal (2016)	TF-IDF	SVM	Simple document representation	Word context are not considered.
	BoW	Logistic Regression	First Book and Movie Review dataset for Nepali text	Limited data size
	TF-IDF (without stop-words)	Multinomial NB		Lacking comprehensive analysis
	BoW (without stop-words)			
Regmi et al. (2017)	Word2Vec	SVM	Subjective text considered	Only preliminary result are shown
	Tweet features	Logistic Regression	Fact and Opinion are analyzed first time	Limited linguistic features are used
	TF-IDF	Naive Bayes		
Piryani et al. (2020)	Lexicons	SVM	Deep learning methods are introduced	Model performance evaluated on short twitter text
	Bigram	Decision Tree	Linguistic features improve the model performance	Generalization to other long text is limited
		Naive Bayes		
		LSTM		
		CNN		
		CNN-LSTM		
Tamrakar et al. (2020)	TF-IDF	SVM	Aspect level SA was attempted	Limited machine learning algorithm were experimented
	POS tag	Naive Bayes	PoS tagging as linguistic feature is used	Performance is limited

remaining models. Furthermore, Piryani et al. (2020) explored the deep learning-based methods for sentiment analysis of tweets written in Nepali language. They also proposed a Nepali lexicon-based sentiment analysis with a combination of Nepali SentiWordNet and emotion lexicon. The pre-trained vectors such as word2vec and tweet features were used to train the deep learning as well as traditional machine learning models. In their method, the CNN-LSTM model trained with 600 positive and 600 negative tweets performs best among other models. Machine learning-based methods have been explored to detect sentiment at aspect level in Tamrakar et al. (2020) using PoS tagging, where TF-IDF was used for the representation of document. They utilized two models– Support Vector Machine and Naive Bayes for the sentiment classification. A list of works in Nepali sentiment analysis along with linguistic features used in those studies, their advantages, and limitation are summarized in Table 8.

6.4 Machine translation

Machine translation is one of the most studied and successful applications in other languages such as English. However, in the case of Nepali language, there have been a few works done coupling with other languages' knowledge. For example, the Hindi-Nepali translation was carried out by Laskar et al. (2019) using a Neural network approach, where the authors did experiments in two directions: Hindi to Nepali and Nepali to Hindi. Since both languages belong to similar classes of language and use Devanagari scripts, the Bilingual Evaluation Understudy (BLEU) (Papineni et al. 2002) score of the proposed system is high, which is 24.6. Another study in Acharya and Bal (2018) compared two approaches: Neural machine translation (NMT) and Statistical machine translation (SMT) for Nepali-English language pair, where SMT outperforms the NMT. Specifically, the reported BLEU for NMT is 3.28, where it is 5.27 for SMT. Similar result was shown in Paul and Purkayastha (2018) for English-Nepali language translation using SMT. The proposed SMT system was built using three language tool kits: a) Moses[12] for decoding; b) Giza++[13] for translation model; and c) IRSTLM[14] for language model. Considering the Nepali language as a low-resource language for machine translation, authors Guzmán et al. (2019) released an English-Nepali pair dataset by taking help from Wikipedia. In their method, authors also experimented with state-of-the-art methods in both SMT and NMT domains on the proposed benchmark dataset. Their experimental result reveals that the result is poor for the English-Nepali pair compared to other language pairs.

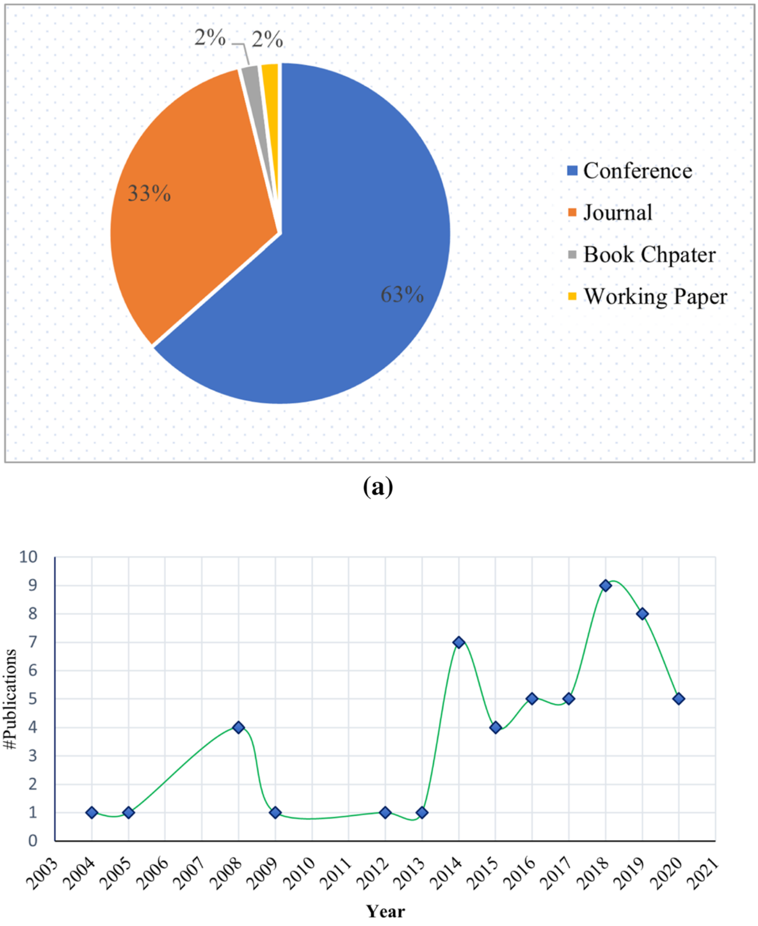
6.5 Others

Apart from well-established applications of NLP research works, there have been several other interdisciplinary applications such as text-to-speech (Shah et al. 2018), recommendations (Basnet and Timalsina 2018), image captioning (Adhikari and Ghimire 2019), optical character recognition (OCR) (Pant and Bal 2016; Acharya et al. 2015), sentence similarity (Sitaula and Ojha 2013), and plagiarism detection (Bachchan and Timalsina 2018).

[12] http://www.statmt.org/moses/.

[13] https://github.com/moses-smt/giza-pp.

[14] https://sourceforge.net/projects/irstlm/.



(a)

(b)

Fig. 7 Distribution of Nepali NLP research works in terms of four publication venues: journal, conference, book chapter, and working paper (**a**) and distribution of Nepali NLP research works by publication year (**b**)

First, the related application to NLP is text to speech (TTS), which translates written text into a speech format. Some attempts towards Nepali TTS are reported in Shah et al. (2018). Mostly, these system adopts the open-source text to speech translation framework such as Festival[15]Taylor et al. (1998) and its variants to perform Nepali TTS. However, these systems still need an improvement on various speech synthesis aspects. Similarly, Optical character recognition and image captioning are at the intersection of image processing and NLP-related works.

A few works on Nepali OCR (optical character recognition) can be found in Pant and Bal (2016), Pant et al. (2012), Acharya et al. (2015). The main challenge in Nepali OCR comes from the complex nature of the Devanagari script, which has a special orthography

[15] https://www.cstr.ed.ac.uk/projects/festival/.

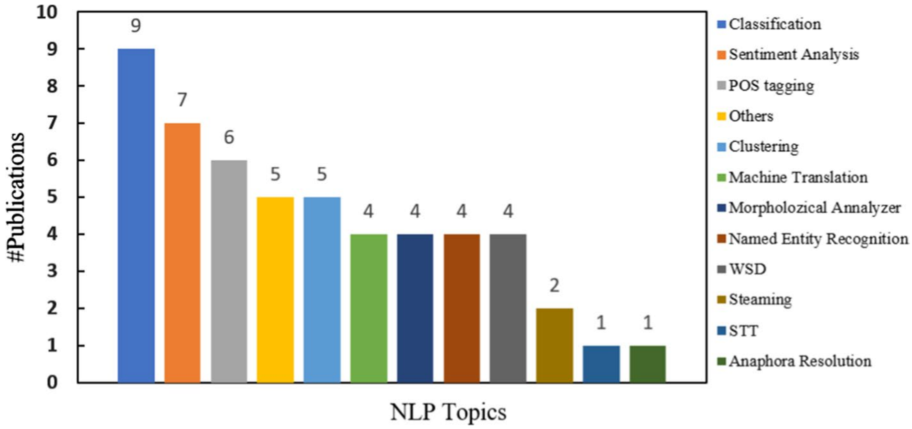


Fig. 8 Distribution of research works in terms of NLP topics for Nepali language. Note that abbreviation of WSD refers to Word sense disambiguation and STT refers to Speech to Text

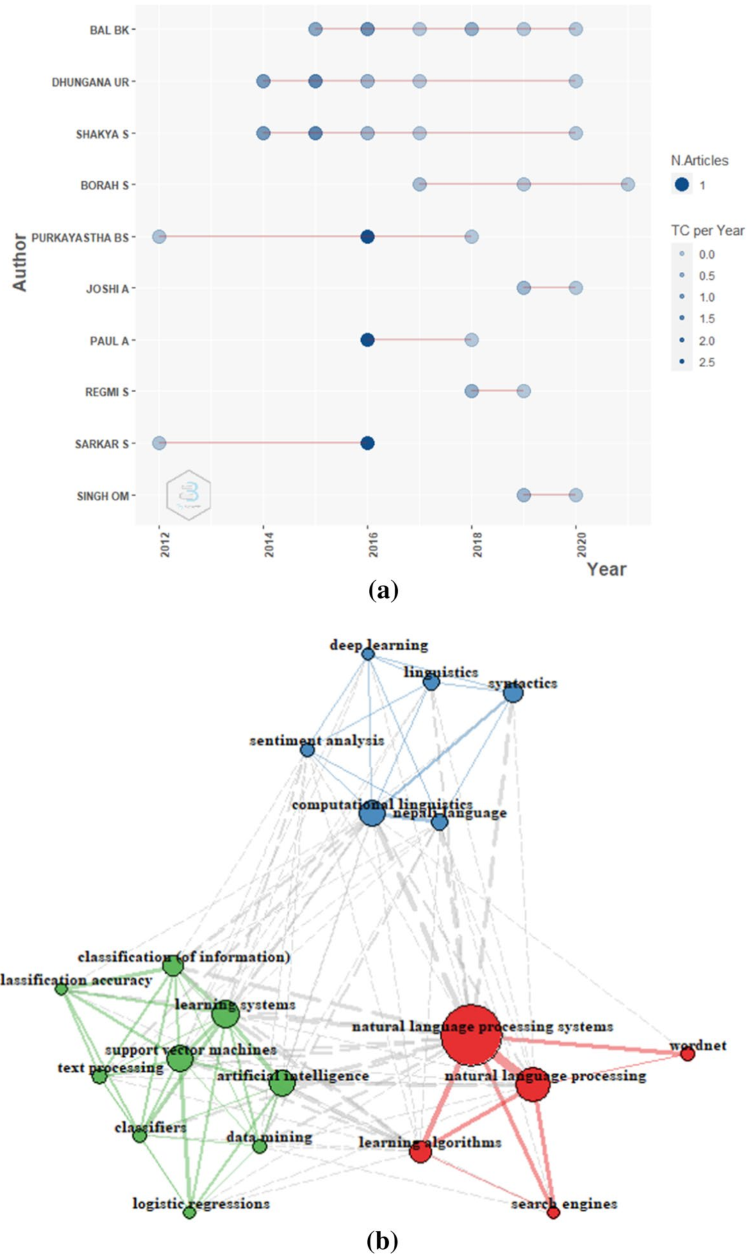
such as half forms of consonants, vowel modifiers, and touching characters. Additionally, another research work related to Nepali sentence similarity was proposed by Sitaula and Ojha (2013). Here, the authors utilize an algorithm based on the finite state machine technique for the similarity of sentences. Nevertheless, there is still a lack of enough datasets in their work and simulations show the limited performance.

Plagiarism detection for Nepali text using a Neural network trained on the corpus of thesis/dissertation written in the Nepali language was developed by Bachchan and Timalsina (2018). They represented the document using TF-IDF and used the cosine similarity score for suspicious text detection. However, their work is limited to thesis documents, thus, their model needs to be further validated on other textual documents.

7 Discussion

In this section, we primarily summarize the NLP survey works carried out in the Nepali language under three different aspects: types of research output publication, direction of research output publication, and volumes of research output publication. Also, we narrate the recent research progresses with the help of bibtex analysis, list out the existing challenges, and suggest some future parades in Nepali language processing.

Nepali language, which is a low-resource language compared to other languages, is still in infancy as there are very few research outcomes as shown in Fig. 8. The majority of these research outcomes are disseminated by conference proceedings (63%) and journal articles (33%) while a few are published as book chapters and working papers as shown in Fig. 7. During the period of 2003 and 2008, most of the researches were focused on the design of major NLP resources such as Nepali National corpus, and Part of speech (PoS) tagset. As of 2010, the natural language processing (NLP) resource transits from computational linguistic prospective to machine learning techniques worldwide; however, the Nepali NLP research works have been progressing very slowly compared to other languages such as Daud et al. (2017). Nevertheless, the widespread of online news portals in Nepali language has expedited the Nepali NLP research after 2015 (see in Fig. 7).



(a)

(b)

Fig. 9 Bibliometric analysis of articles retrieved from Scopus database: **a** top-ten authors' productivity over time, and **b** co-occurrence of main keywords used in NLP research works in Nepali language. Note that the size of each circle is proportional to the frequency of the keyword and color of circle represents the cluster of keywords that belongs to similar topics in NLP research works

Over the last decades, a few attempts have been made on various aspects of Nepali language processing such as NLP approaches, techniques, and applications tasks. Under NLP approaches, initially, rule-based approaches were mostly used but the development of machine learning approaches to tackle various NLP topics such as PoS tagging, Stemming, and Named entity recognition have been noticed recently. The machine learning algorithms such as SVM, ANN, and LSTM are the most popular methods (refer to Figs. 4 and 5). Comparing the NLP techniques and applications such as POS tagging, Stemming, Named entity recognition, Classification, and Clustering, a high percentage of works have been done in high-level analysis of Nepali documents such as Classification, Sentiment analysis, and Clustering. These tasks pay attention a little in terms of the linguistic aspects of the language. Most of such tasks utilize the news content from online news portals for their research works. The reason behind it might be the free and easy accessibility of web content worldwide. On the other hand, linguistics-based tasks such as Morphological analysis, Named entity recognition, Word sense disambiguation, and Stemming are less-studied (refer to Fig. 8). By taking reference to high-resource languages such as English, German, and Spanish, we believe that we should first focus towards developing the linguistics resource, which could, thereafter, expedite the development of NLP tasks such as Machine translation, Classification, Clustering, and so on.

To observe the recent trends and advancements in Nepali NLP research works, we perform a bibtex analysis based on the published articles retrieved from the Scopus database. As a result, it shows that the author's performance among published articles related to Nepali NLP works in recent years are a very few in numbers, who are working constantly in this field, such as "Bal BK" and "Dhungana UR" as shown in Fig. 9. In the meantime, the thematic analysis of the main keywords presented in those published articles reveals that people working in Nepali NLP works are mainly focused on three main domains: Natural Language Processing system, Machine learning approaches, and Computational linguistic perspective (please refer to Fig. 9). Considering the results as seen in two diagrams (Fig. 9), we believe that Nepali NLP research works, which has been relying on a quite few researchers, are primarily progressing towards recent technologies such as machine learning in addition to the computational linguistics.

Here, we summarize the findings of this survey into four groups: existing well-developed NLP works in the Nepali language, current focus, challenges and/or bottlenecks, and suggestions. With respect to linguistic resource development, Nepali national corpus, PoS tagset, Nepali lexicon, has been available as well-developed resources. Based on these resources, the downstream NLP tasks such as Named entity recognition, Word sense disambiguation, Nepali spell checker, Classification and Clustering, are in good progress. As an example, Machine translation (e.g.Dobhase), Speech to Text (STT) and Nepali computational grammar have been built as a prototype (Bal 2009). Recently, many NLP researchers in the Nepali language are more focused on the application of machine learning and deep learning to address the various high-level NLP tasks such as Sentiment analysis, Classification, and Clustering of news articles with higher accuracy. However, the major challenges in Nepali NLP are the lack of gold standard datasets, high-performing methods and approaches, and well-established text representations techniques.

In summary, we detail these challenges and strategies as follows:

(i) The first challenge is the lack of gold standard datasets for each Nepali NLP task. For instance, there are three news classification datasets (refer to Sect. 3 for the list of datasets) used in different research works but a common heavy-weight gold standard news dataset is still lacking for Nepali NLP research works. Dataset preparation is a challenging and labour-intensive task as it requires domain experts' input and manual annotation, which is the main limitation or barrier of Nepali language research. Also, the funding to develop such datasets is limited as there are no dedicated agencies for the development of linguistic resources. With the help of dedicated funding and monitoring body, we would be able to standardize, collect, and prepare the gold-standard dataset in the future.

(ii) The second challenge is towards improving the performance of fundamental NLP tasks such as Stemming, Morphological analyzer, and PoS tagging tasks. In the existing literature, these tasks are mostly accompanied by rule-based systems; however, the statistical and machine learning methods have been proved to be more accurate in other resource rich languages such as English. The main reason that researchers couldn't explore the machine learning and deep learning-based methods for such tasks in the Nepali language is again related to the insufficient training/annotated datasets. This problem could be overcome by using the Transfer learning (TL) approach, that is, transferring the knowledge from the model trained on large datasets to the target domain. As an example, we can transfer the knowledge from the models pre-trained in Hindi or other languages to the Nepali text.

(iii) The third challenge is on the enhancement of document representation techniques. While doing the literature review, we notice that existing works prefer using traditional methods such as TF-IDF based representations mostly for Classification, Clustering, and Sentiment analysis. However, such representation methods ignore the contextual information of the textual documents. To this end, the word embedding technique such as word2vec, which has been proved to be very useful in English and other language document representation, is yet to be explored fully for Nepali document representation.

8 Conclusion

We have provided a detailed analysis and discussion of Nepali NLP research works in various aspects, such as approaches, techniques, and application tasks. We notice that the datasets and language resources such as Wordnet are cardinal in conducting NLP research. We unwrap the available Nepali datasets through description and their associated tasks, which ultimately help the future Nepali language research to augment in this field. It is clearly seen that machine learning methods outperform the rule-based methods in each NLP application task when sufficient data is available. Thus, this paper emphasizes to use the machine learning methods for such tasks and suggests to develop the large annotated datasets in parallel.

To sum up, we have reviewed the NLP research works on Nepali textual documents, where we notice that most of the researches in Nepali linguistic community are focusing on high-level NLP applications such as Classification, Sentiment analysis, and Machine translation; however, a few works have been done in building resources for fundamental NLP approaches and techniques such as Stemming, Morphological analysis, PoS tagging, Parsing and grammar checking, and Word sense disambiguation. Our study does not consider the computational aspects of NLP research works in the Nepali language. Given such limitations in this study, we underscore that extensive study of computational aspect of Nepali language needs to be carried out in this domain for further developments.

9 List of abbreviations

The list of abbreviations used in this work are reported in Table 9.

Table 9 List of abbreviations

Abbreviations	Full form
ANN	Artificial Neural Network
AR	Anaphora Resolutions
BERT	Bidirectional Encoder Representations from Transformers
BoW	Bag of Words
BLEU	Bilingual Evaluation Understudy
BiLSTM	Bidirectional Long Short Term Memory
CFG	Context Free Grammar
CNN	Convolution Neural Network
DL	Deep Learning
DT	Decision Tree
GRU	Gated Recuurent Unit
HTML	Hyper Text Markup Language
HMM	Hidden Markov Model
IR	Information Retrieval
IDF	Inverse Document Frequency
LDA	Latent Dirichlet Allocation
LSTM	Long Short Term Memory
LR	Logistic Regression
MA	Morphological Analyzer
ML	Machine Learning
MPP	Madan Puruskar Pustakalaya
NB	Naive Bayes
NN	Neural Network
NER	Named Entity Recognition
NNC	Nepali National Corpus
NMT	Neural Machine Translation
NLP	Natural Language Processing
OCR	Optical Character Recognition
PCA	Principal Component Analysis
POS	Part of Speech
RNN	Recurrent Neural Network
RB	Rule Based
SVM	Support Vector Machine
SMT	Statistical Machine Translation
TnT	Tri-gram Tagger
TF	Term Frequency
SA	Sentiment Analysis
STT	Speech to Text
TTS	Text to Speech
WSD	Word Sense Disambiguation

References

Acharya P, Bal BK (2018) A comparative study of SMT and NMT: case study of English-Nepali language pair. In: SLTU, pp 90–93

Acharya S, Pant AK, Gyawali PK (2015) Deep learning based large scale handwritten devanagari character recognition. In: 2015 9th International conference on software, knowledge, information management and applications (SKIMA). IEEE, pp 1–6

Adhikari A, Ghimire S (2019) Nepali image captioning. In: 2019 artificial intelligence for transforming business and society (AITB), IEEE 1:1–6

Bachchan RK, Timalsina AK (2018) Plagiarism detection framework using monte carlo based artificial neural network for Nepali language. 2018 IEEE 3rd international conference on computing. Communication and security (ICCCS). IEEE, pp 122–127

Bal BK (2009) Towards building advanced natural language applications–an overview of the existing primary resources and applications in Nepali. In: Proceedings of the 7th workshop on Asian language resources (ALR7), Association for Computational Linguistics, Suntec, Singapore, pp 165–170

Bal BK, Shrestha P (2004) A morphological analyzer and a stemmer for Nepali. PAN Localization, Working Papers 2007:324–331

Bal BK, Shrestha P, Pustakalaya MP, PatanDhoka N (2007) Architectural and system design of the Nepali grammar checker. PAN Localization Working Paper

Bam S, Shahi T (2014) Named entity recognition for Nepali text using support vector machines. Intell Inf Manag 6(2):21–29. https://doi.org/10.4236/iim.2014.62004

Basnet A, Timalsina AK (2018) Improving Nepali news recommendation using classification based on LSTM recurrent neural networks. In: 2018 IEEE 3rd international conference on computing. Communication and Security (ICCCS), IEEE, pp 138–142

Basnet A, Timalsina AK (2018) Improving Nepali news recommendation using classification based on lstm recurrent neural networks. In: Proceedings of international conference on computing, Communication and Security (ICCCS), pp 138–142

Bhala RV, Abirami S (2014) Trends in word sense disambiguation. Artif Intell Rev 42(2):159–171

Bhat SM, Rai R (2012) Building morphological analyzer for Nepali. J Modern Lang 22(1):45–58

Bista S, Keshari B, Bhatta J, Parajuli K (2005) Dobhase: online English to Nepali machine translation system. In: The proceedings of the 26th Annual conference of the Linguistic Society of Nepal

Bista S, Khatiwada L, Keshari B (2004) Nepali lexicon development. PAN Localization, Working Papers 2007:311–15

Borah S, Choden U, Lepcha N (2017) Design of a morph analyzer for non-declinable adjectives of nepali language. In: Proceedings of the 2017 international conference on machine learning and soft computing, pp 126–130

Brown PF, Della Pietra VJ, Desouza PV, Lai JC, Mercer RL (1992) Class-based n-gram models of natural language. Comput Linguist 18(4):467–480

Carpuat M, Wu D (2007) Improving statistical machine translation using word sense disambiguation. In: Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL), pp 61–72

Chhetri I, Dey G, Das SK, Borah S (2015) Development of a morph analyser for Nepali noun token. In: 2015 international conference on advances in computer engineering and applications. IEEE, pp 984–987

Choudhary N, Ramamoorthy L (2019) LDC-IL raw text corpora: an overview. Linguistic resources for AI/NLP in Indian languages. Central Institute of Indian Languages, Mysuru pp 1–10

Cristianini N, Shawe-Taylor J (2000) An introduction to support vector machines and other kernel-based learning methods. Cambridge University Press, Cambridge

Dangol D, Shrestha RD, Timalsina A (2018) Automated news classification using n-gram model and key features of Nepali language. SCITECH Nepal 13(1):64–69

Daud A, Khan W, Che D (2017) Urdu language processing: a survey. Artif Intell Rev 47(3):279–311

Dey A, Paul A, Purkayastha BS (2014) Named entity recognition for Nepali language: a semi hybrid approach. Int J Eng Innov Technol (IJEIT) 3:21–25

Dhungana UR, Shakya S (2014) Word sense disambiguation in Nepali language. In: 2014 Fourth international conference on digital information and communication technology and its applications (DICTAP). IEEE, pp 46–50

Ekbal A, Bandyopadhyay S (2008) Bengali named entity recognition using support vector machine. In: Proceedings of the IJCNLP-08 workshop on named entity recognition for south and south east Asian Languages

Gupta CP, Bal BK (2015) Detecting sentiment in Nepali texts: a bootstrap approach for sentiment analysis of texts in the Nepali language. In: 2015 international conference on cognitive computing and information processing (CCIP). IEEE, pp 1–4

Guzmán F, Chen P, Ott M, Pino J, Lample G, Koehn P, Chaudhary V, Ranzato M (2019) Two new evaluation datasets for low-resource machine translation: Nepali-English and Sinhala-English. CoRR abs/1902.01382. http://arxiv.org/abs/1902.01382

Hung C, Chen SJ (2016) Word sense disambiguation based sentiment lexicons for sentiment classification. Knowl-Based Syst 110:224–232

Kafle K, Sharma D, Subedi A, Timalsina AK (2016) Improving Nepali document classification by neural network. In: Proceedings of IOE graduate conference, pp 317–322

Khanal R (2019) Linguistic geography of nepalese languages. Third Pole J Geogr Educ 18:45–54. https://doi.org/10.3126/ttp.v18i0.27994

Khatiwada R (2009) Nepali. J Int Phon Assoc 39(3):373–380

Lamsal R (2020) A large scale Nepali text corpus. IEEEdataport. https://doi.org/10.21227/jxrd-d245

Lappin S, Leass HJ (1994) An algorithm for pronominal anaphora resolution. Comput Linguist 20(4):535–561

Laskar SR, Pakray P, Bandyopadhyay S (2019) Neural machine translation: Hindi-Nepali. In: Proceedings of the fourth conference on machine translation (Volume 3: Shared Task Papers, Day 2), pp 202–207

Lesk M (1986) Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In: Proceedings of the 5th annual international conference on Systems documentation, pp 24–26

Lewis DD (1998) Naive (bayes) at forty: the independence assumption in information retrieval. In: European conference on machine learning. Springer, pp 4–15

MacQueen J et al (1967) Some methods for classification and analysis of multivariate observations. In: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, Oakland, CA, USA 1:281–297

Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space. arXiv:13013781

Miller GA (1995) Wordnet: a lexical database for english. Commun ACM 38(11):39–41

Pant AK, Panday SP, Joshi SR (2012) Off-line nepali handwritten character recognition using multilayer perceptron and radial basis function neural networks. In: 2012 third Asian Himalayas international conference on internet, IEEE, pp 1–5

Pant N, Bal BK (2016) Improving Nepali ocr performance by using hybrid recognition approaches. In: 2016 7th international conference on information, intelligence, systems & applications (IISA). IEEE, pp 1–6

Papineni K, Roukos S, Ward T, Zhu WJ (2002) BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics, pp 311–318

Paul A, Purkayastha BS (2018) English to Nepali statistical machine translation system. In: Proceedings of the international conference on computing and communication systems. Springer, pp 423–431

Paul A, Purkayastha BS, Sarkar S (2015) Hidden Markov model based part of speech tagging for Nepali language. In: 2015 international symposium on advanced computing and communication (ISACC). IEEE, pp 149–156

Piryani R, Piryani B, Singh VK, Pinto D (2020) Sentiment analysis in Nepali: exploring machine learning and lexicon-based approaches. J Intell Fuzzy Syst (Preprint):1–12

Poli R, Kennedy J, Blackwell T (2007) Particle swarm optimization. Swarm Intell 1(1):33–57

Prabha G, Jyothsna P, Shahina K, Premjith B, Soman K (2018) A deep learning approach for part-of-speech tagging in nepali language. In: 2018 international conference on advances in computing. Communications and informatics (ICACCI). IEEE, pp 1132–1136

Prajwal R, Prasad KL, Bal BK (2008) Report on Nepali computational grammar. Madan Puraskar Pustakalaya https://www.academia.edu/2414578/Report_on_Nepali_Computational_Grammar

Prasain B (2008) Computational analysis of Nepali basic verbs (written forms). NepaleseLinguistics 23:262–270

Prasain B, Khatiwada L, Bal B, Shrestha P (2008) Part-of-speech tagset for Nepali. Madan Puraskar Pustakalaya, Unpublished

Regmi S, Bal BK, Kultsova M (2017) Analyzing facts and opinions in Nepali subjective texts. In: 2017 8th international conference on information, intelligence, systems & applications (IISA). IEEE, pp 1–4

Salton G, McGill MJ (1983) Introduction to modern information retrieval. Mcgraw-Hill, New York

Sarkar S, Roy A, Purkayastha B (2014) A comparative analysis of particle swarm optimization and K-means algorithm for text clustering using Nepali wordnet. Int J Nat Lang Comput (IJNLC) 3(3):83–92. http://www.airccse.org/journal/ijnlc/papers/3314ijnlc08.pdf

Senapati A, Poudyal A, Adhikary P, Kaushar S, Mahajan A, Saha BN (2020) A machine learning approach to anaphora resolution in Nepali language. In: 2020 international conference on computational performance evaluation (ComPE). IEEE, pp 436–441

Shah KB, Chaudhary KK, Ghimire A (2018) Nepali text to speech synthesis system using FreeTTS. SCITECH Nepal 13(1):24–31

Shahi TB, Dhamala TN, Balami B (2013) Support vector machines based part of speech tagging for Nepali text. Int J Comput Appl 70(24):38–42. https://doi.org/10.5120/12217-8374

Shahi TB, Pant AK (2018) Nepali news classification using naïve bayes, support vector machines and neural networks. In: 2018 International conference on communication information and computing technology (ICCICT). IEEE, pp 1–5

Shahi TB, Shakya S (2018) Nepali SMS filtering using decision trees, neural network and support vector machine. In: 2018 international conference on advances in computing. Communication Control and Networking (ICACCCN). IEEE, pp 1038–1042

Shahi TB, Yadav A et al (2014) Mobile sms spam filtering for Nepali text using naïve bayesian and support vector machine. Int J Intell Sci 4(01):24–28

Shrestha BB, Bal BK (2020) Named-entity based sentiment analysis of Nepali news media texts. In: Proceedings of the 6th workshop on natural language processing techniques for educational applications, pp 114–120

Shrestha I, Dhakal SS (2016) A new stemmer for Nepali language. In: 2016 2nd international conference on advances in computing, communication, & automation (ICACCA). IEEE, pp 1–5

Shrestha N, Hall PA, Bista SK (2008) Resources for nepali word sense disambiguation. In: 2008 international conference on natural language processing and knowledge engineering. IEEE, pp 1–5

Singh OM, Padia A, Joshi A (2019) Named entity recognition for nepali language. In: 2019 IEEE 5th international conference on collaboration and internet computing (CIC). IEEE, pp 184–190

Singh OM, Timilsina S, Bal BK, Joshi A (2020) Aspect based abusive sentiment detection in Nepali social media texts. In: 2020 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM). IEEE, pp 301–308

Singh J, Gupta V (2017) A systematic review of text stemming techniques. Artif Intell Rev 48(2):157–217

Sitaula C (2012) Semantic text clustering using enhanced vector space model using Nepali language. Comput Sci Telecommun 4:41–46

Sitaula C (2013) A hybrid algorithm for stemming of Nepali text. Intell Inf Manag. https://doi.org/10.4236/iim.2013.54014

Sitaula C (2014) Semantic orientation of texts using iterative finite state machine. J Comput Sci Control Syst 7(1):51

Sitaula C, Ojha YR (2013) Semantic sentence similarity using finite state machine. Intell Inf Manag 5(6):171–174

Sitaula C, Basnet A, Aryal S (2021) Vector representation based on a supervised codebook for nepali documents classification. PeerJ Comput Sci 7:e412

Subba S, Paudel N, Shahi TB (2019) Nepali text document classification using deep neural network. Tribhuvan Univ J 33(1):11–22

Tamrakar S, Bal BK, Thapa RB (2020) Aspect based sentiment analysis of Nepali text using support vector machine and naive bayes. Tech J 2(1):22–29

Taylor P, Black AW, Caley R (1998) The architecture of the festival speech synthesis system. In: The third ESCA/COCOSDA workshop (ETRW) on speech synthesis

Thakur SK, Singh VK (2014) A lexicon pool augmented Naive Bayes classifier for Nepali text. In: Proceedings of seventh international conference on contemporary computing (IC3), pp 542–546

Thapa LBR, Bal BK (2016) Classifying sentiments in Nepali subjective texts. In: 2016 7th international conference on information, intelligence, systems & applications (IISA). IEEE, pp 1–6

Wang Y, Wang M, Fujita H (2020) Word sense disambiguation: a comprehensive knowledge exploitation framework. Knowl-Based Syst 190(105):030. https://doi.org/10.1016/j.knosys.2019.105030

Yadava YP, Hardie A, Lohani RR, Regmi BN, Gurung S, Gurung A, McEnery T, Allwood J, Hall P (2008) Construction and annotation of a corpus of contemporary Nepali. Corpora 3(2):213–225

Yajnik A (2017) Part of speech tagging using statistical approach for Nepali text. World Acad Sci Eng Technol Int J Comput Electr Autom Control Inf Eng 11(1):76–79

Yajnik A (2018) Ann based pos tagging for nepali text. Int J Nat Lang Comput 7:13–18

Zhong Z, Ng HT (2012) Word sense disambiguation improves information retrieval. In: Proceedings of the 50th annual meeting of the association for computational linguistics (Volume 1: Long Papers), pp 273–282

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.