neDIOM: Dataset and Analysis of Nepali Idioms

Rhitabrat Pokharel and Ameeta Agrawal
Department of Computer Science
Portland State University, USA
{pokharel,ameeta}@pdx.edu

Abstract

Idioms, integral to any language, convey nuanced meanings and cultural references. However, beyond English, few resources exist to support any meaningful exploration of this unique linguistic phenomenon. To facilitate such an inquiry in a low resource language, we introduce a novel dataset of Nepali idioms and the sentences in which these naturally appear. We describe the methodology of creating this resource as well as discuss some of the challenges we encountered. The results of our empirical analysis under various settings using four distinct multilingual models consistently highlight the difficulties these models face in processing Nepali figurative language. Even fine-tuning the models yields limited benefits. Interestingly, the larger models from the BLOOM family of models failed to consistently outperform the smaller models. Overall, we hope that this new resource will facilitate further development of models that can support processing of idiomatic expressions in low resource languages such as Nepali.

1 Introduction

Idioms are inherent linguistic phenomena in all languages, comprising a collection of words that, when combined, convey a unique and distinct meaning not achievable by the individual words within the phrase. Neglecting idioms would lead to a significant loss of meaning and context, given their tendencies to carry nuances and cultural references. Properly identifying and processing idioms is essential for machine translation, sentiment analysis, information retrieval, and several other tasks. Large language models (LLMs), such as GPT and LLaMa are designed to mimic human language understanding and generation. A comprehensive grasp of idioms is crucial to ensure that these models generate text that is not only linguistically accurate but also contextually meaningful.

There are plenty of idiom resources for high resource languages such as English (Korkontzelos et al., 2013; Tayyar Madabushi et al., 2021), Chinese (Tan and Jiang, 2021b), Japanese (Tedeschi et al., 2022), Italian (Tedeschi et al., 2022; Moussallem et al., 2018), and German (Tedeschi et al., 2022; Moussallem et al., 2018; Fadaee et al., 2018) to name a few. However, idioms in low resource languages have received less attention.

When resources are available, several interesting tasks involving idioms have been studied. Measuring semantic similarity between idioms (Korkontzelos et al., 2013), classification between idiomatic and literal usages of idioms (Tayyar Madabushi et al., 2021), translation (Moussallem et al., 2018) and language generation (Chakrabarty et al., 2022b; Pokharel and Agrawal, 2023) are some of the main focuses. However, we still do not know how LLMs process idioms in a low resource language like Nepali.

There might be a perception that plenty of methodologies and resources are readily available for the compilation of analogous linguistic phenomena, i.e., Multiword Expressions (MWEs). It is important to clarify that, MWEs constitute a broader linguistic category including not only idiomatic expressions but also other linguistic phenomena like noun compounds and sentence fragments, and the customary collection processes for MWEs, such as part-of-speech tagging (Farahmand et al., 2015) and statistical co-occurrence analysis (Kunchukuttan and Damani, 2008), prove to be insufficient in effectively distinguishing idiomatic expressions.

In this work, we introduce a novel dataset – neDIOM – of almost 200 Nepali idioms and more than 500 sentences of their contextual usage, making this, to our knowledge, the first such dataset in Nepali[1]. By contributing this resource, we hope to

[1]The dataset will be made available for further research.

Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2025), pages 160–171
January 19, 2025.

facilitate exploration of LLMs' performance in handling idiomatic expressions and documenting linguistic phenomena in this low-resource language.

Depending on the language and the scenario, the idiom dataset creation job can be more or less challenging. For instance, in English, the idiom "under the weather" can be directly used in a sentence without alteration. However, "pull someone's leg" undergoes inflection, posing a significant challenge for automated idiom identification (Pasquer et al., 2020), especially in non-Latin languages. We enumerate further challenges related to dataset creation in the subsequent sections.

Our experiments with LLMs reveal that their performance with respect to idioms in low-resource languages leaves a big room for improvement.

The main contributions of our work are:

- The introduction of a new dataset in Nepali, which includes idioms, their contextual usage, and the marking of idiom positions.
- An extensive benchmarking of this dataset using several state-of-the-art LLMs.

2 Related Work

In this section, we review existing work in creating idiom resources and related tasks.

2.1 Idiom Datasets

The development of datasets focusing on idioms has seen some diversity across multiple languages, with English being the predominant language (Peng et al., 2015; Haagsma et al., 2020; Chakrabarty et al., 2022b). Attention has also extended to well-resourced languages like French, Dutch, Italian, Portuguese, Chinese, Polish, and Japanese (Korkontzelos et al., 2013; Moussallem et al., 2018; Tan and Jiang, 2021b; Tedeschi et al., 2022; Qiang et al., 2023). In contrast, languages with fewer resources, including Gujarati, Telugu, and Malayalam, have received less investigation (Agrawal et al., 2018). Notably, for Nepali, there is only one small dataset with 42 samples and without context sentences (Neupane, 2018).

Some datasets were created by translating idioms from English to other languages (Moussallem et al., 2018; Neupane, 2018; Fadaee et al., 2018; Tang, 2022), but the translated idioms are not always an idiom in the target language (Agrawal et al., 2018). In the cases when datasets have been created from scratch, the idioms are typically collected from one source and the sentences containing the idioms from another source (Korkontzelos et al., 2013; Peng et al., 2015; Fadaee et al., 2018; Zheng et al., 2019; Haagsma et al., 2020; Tan and Jiang, 2021b; Tedeschi et al., 2022). For the former step, most idioms are collected from sources where the idioms are already listed as such, precluding the need to identify the idiom from a sentence/paragraph (Korkontzelos et al., 2013; Fadaee et al., 2018; Zheng et al., 2019; Haagsma et al., 2020; Tedeschi et al., 2022). For the latter step (i.e., collecting the context where the idioms have been used), Haagsma et al. (2020) used automatic method which was later checked by manual reviewers, while Tayyar Madabushi et al. (2021) manually collected both the idioms and the contexts manually from the internet. Annotations for idiom-related tasks are also often obtained manually (Agrawal et al., 2018; Neupane, 2018; Haagsma et al., 2020).

2.2 Idiom Tasks

Korkontzelos et al. (2013) presented work on **semantic similarity**, encompassing idioms across English, French, German, and Italian. (Salehi et al., 2018) investigate the compositionality of idiomatic expressions by leveraging multilingual lexical resources, focusing on English and Germanic languages. Tan and Jiang (2021b) focused on gauging the similarity between idioms, concentrating specifically on the Chinese language. Chakrabarty et al. (2022b) studied natural language inference with a focus on idiomatic expressions in English.

Numerous studies have tackled the challenge of **distinguishing between idiomatic and literal language usage**, classifying expressions into idiomatic and literal categories (Peng et al., 2015; Tayyar Madabushi et al., 2021; Tan and Jiang, 2021a). Haagsma et al. (2020) classified idiomatic versus literal usages while also annotating their genre. Tedeschi et al. (2022) explored idiom identification across multiple languages, including Chinese, Dutch, French, Japanese, Polish, Portuguese, Spanish, and more. Other studies have also contributed to idiom classification, cloze tasks, and usage recognition across various languages (Zheng et al., 2019; Tian et al., 2023; Fenta and Gebeyehu, 2023; Zhou et al., 2023).

Translation of idiomatic expressions is another key area of investigation (Moussallem et al., 2018; Fadaee et al., 2018; Neupane, 2018; Agrawal et al., 2018; Tang, 2022). However, in several cases the translated idioms were not necessarily idioms in

Idiom	S1	S2	S3	Label	Idiom's Position
डाँडो काट्नु *daando kaatnu* 'to cover a considerable distance'	जे छ त्यसमा नै चित्त बुझाएर दसैं कटाउने जोहो मिलाउनुहोस्। *j cha tesmaa nai, chitta bujhayera dashain kataune joho milaaunuhos* 'Find contentment in whatever you have to celebrate Dashain.'	पाहुना आफन्त बोलाउँदा खर्चले डाँडो काट्न सक्छ। *paahunaa aafanta boolaaundaa kharchale daando kaatna sakcha* 'Inviting guests and relatives could exceed the budget.'	फेरि सानो रकम टीका लाएर दिँदा चित्त नबुझ्न सक्छ। *feri saano rakam tikaa liyera dinda chitta nabujhna sakcha* 'Given the circumstance, providing a small offering with Tika may not suffice.'	*I*	पाहुना आफन्त बोलाउँदा खर्चले ###डाँडो काट्न### सक्छ। *paahunaa aafanta boolaaundaa kharchale ###daando kaatna### sakcha*
इन्तु न चिन्तु हुनु *intu na chintu hunu* 'to get overly anxious'	तर, यी युवाको उपचार नहुँदा शरीर कुहिन थालेको छ। *tara, yi yuwale upachaar nahundaa shareer kuhina thaaleko cha* 'However, due to the lack of treatment of this young man, the body has started to rot.'	उनका बुवा बलबहादुर छोराको यो अवस्था देखेर इन्तु न चिन्तु छन्। *unkaa buwaa balbahaadur choraako yo abasthaa dekhera intu na chintu chan* 'His father, Bal Bahadur, is deeply distraught upon witnessing his son's condition.'	अस्पतालले भनेको तीन लाख रुपैयाँ जुटाउन नसकेपछि बस्नेतले सबैसँग हारगुहार गरे। *aspataalle vaneko teen laakh rupainyaa jutaauna nasakepachhi basnetle sabaisanga haarguhaar garey* 'After failing to arrange the three lakh rupees as demanded by the hospital, Basnet has now turned to everyone.'	*L*	उनका बुवा बलबहादुर छोराको यो अवस्था देखेर ###इन्तु न चिन्तु### छन्। *unkaa buwaa balbahaadur choraako yo abasthaa dekhera ###intu na chintu chan###*
आकाशको फल *aakhaashko fal* 'a pie in the sky'	तर, त्यसै बस्नुभन्दा म्युजिक भिडिओमा काम गर्दा पनि नयाँ नयाँ कुरा जान्न र अनुभव गर्न मिल्ने उनले बताए। *tara, tyasai basnubhandaa music videoma kaam gardaa pani nayaan nayaan kuraa jaanna ra anubhaab garna milne unle bataaye* 'However, he said that instead of sitting there, you can learn and experience new things while working on a music video.'	आज आकाशको फ्यान फलोर्स हजारौं छन्। *aaja aakaahko fyan fallowers hajaaroun chhan* 'Today Akash has thousands of fan followers.'	फुर्सदमा सामाजिक सञ्जालमा आफ्नो कामबारे सर्वसाधारणले गरेका कमेन्ट पनि पढ्ने गरेको उनले बताए। *fursadmaa saamaajik sanjaalmaa aafno kaambaare sarwasaadharanle gareko kament pani padhne gareko unle bataaye* 'He said that in his spare time, he also reads the comments made by public about his work on social media.'	NA	आज आकाशको फ्यान फलोर्स हजारौं छन्। *aaja aakaahko fyan fallowers hajaaroun chhan*

Table 1: Samples from the dataset, each associated with a distinct label.

the target language.

The **generation** of idiomatic expressions and their paraphrases has also attracted much attention. Chakrabarty et al. (2022a) investigated generating plausible continuations for idiomatic sentences in English, meanwhile Zhou et al. (2022); Qiang et al. (2023) focused on generating literal paraphrases. (Pokharel and Agrawal, 2023) evaluated language models' ability to generate contextually relevant continuations for narratives with idiomatic expressions in English and Portuguese.

Needless to say, yet important to highlight, is the fact that the exploration of idiomatic expressions in low-resource languages has received much less attention.

3 `neDIOM`: Nepali Idiom Dataset

We introduce `neDIOM`, a dataset of Nepali idioms along with their naturally-occurring contexts. The dataset comprises 526 carefully selected samples

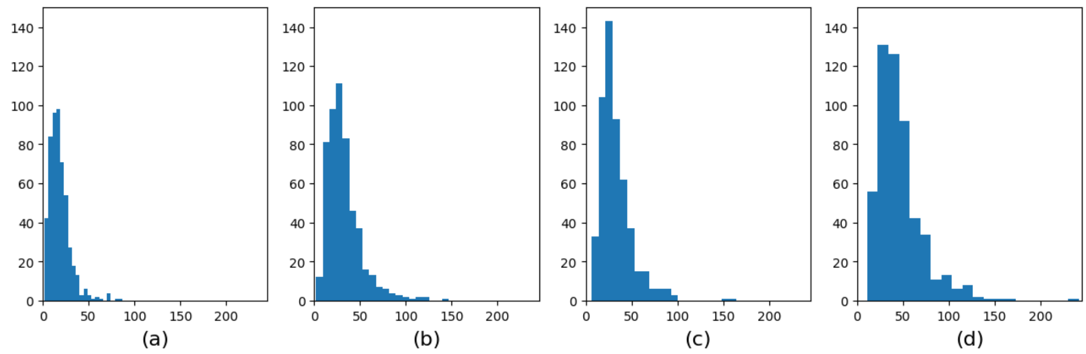

Figure 1: Distribution of sentence lengths for (a) *S2* (b) *S1* + *S2* (c) *S2* + *S3*, and (d) *S1* + *S2* + *S3*. On the x-axis is the number of words in a sentence, and on the y-axis is the frequency.

containing 191 unique idioms. Each sample includes the idiom, the sentence in which it appears, and the preceding and following sentences in the context. A selection of samples from the curated dataset is presented in Table 1. This dataset contains six attributes:

- *Idiom* is a multiword expression, whose overall meaning cannot be derived directly from the meanings of its individual words;
- *S2* is the sentence in which the idiom appears;
- *S1* is the sentence that precedes *S2* in the contextual sequence;
- *S3* is the sentence that follows *S2* in the context;
- *Label* indicates the annotation, whether the idiom is being used in an idiomatic sense or a literal sense; and
- *Idiom's position* specifies the exact location of the idiom within *S2*.

3.1 Data Collection

Next, we outline the methodology used for creating this dataset.

Collecting idioms A total of 296 idioms were manually collected from across the internet and from the reference (आचार्य, ऋषिकेश, 2020). These idioms were subsequently used to extract contextual usage. One might wonder why the context was not collected simultaneously along with these expressions. The reason is that the sources from which we obtained the idioms mostly provided definitions or descriptions without the associated contexts.

Collecting contexts with idioms The next step focuses on collecting naturally-occurring sentences and contexts in which these idioms appear. We use the OSCAR corpus[2], an expansive multilingual collection with over 152 languages, which is a result of the language-wise classification of content from the Common Crawl corpus[3]. We chose this corpus because of the abundance of Nepali text it offered, approximately 392K documents, which is particularly significant for Nepali, a language with considerable resource constraints.

The idioms originally appear in gerund form which changes its grammatical structure when used in a sentence. To identify relevant sentences containing these idioms, we adopted a strategy of using partial segments of the idioms. For instance, for the idiom नाक खुम्च्याउनु (*naak khumchyaaunu*, 'to turn up one's nose'), we employed the truncated version नाक ख (*naak kha*) to expand our search scope. In this context, खु (*khu*) represents the initial syllable of the word खुम्च्याउनु (*khumchyaaunu*) and ख (*kha*) stands as the first grapheme, which maintains consistency regardless of the word's usage. We utilized a similar technique for idioms containing more than two words.

After extracting documents from the OSCAR corpus, we tokenized them at sentence-level to obtain *S1*, *S2*, *S3* for each idiom instance using the *indic_tokenize* module[4]. This process resulted in a total of 1,216 samples (idioms and their surrounding contexts). It is worth noting that of the 296 idioms we had used in our search, we were able to collect contexts for 271 idioms. These were further reduced to 191 idioms after manual annotation.

Figure 1 plots the context sentence lengths of *S2*, *S1*+*S2*, *S2*+*S3*, and *S1*+*S2*+*S3*. We observe that most sentences with idioms (*S2*) consist of 50

[2] https://huggingface.co/datasets/oscar-corpus/OSCAR-2201
[3] https://commoncrawl.org/
[4] https://indic-nlp-library.readthedocs.io/en/latest/indicnlp.tokenize.html

Idiom	Lemma	S2	Lemmatized S2
आलु खानु	['**आलु**', '**खान**']	छोराले परीक्षामा दुई ओटा विषयमा **आलु खाएछ** ।	'छोरा', 'परीक्षामा', 'दुई', 'ओटा', 'विषय', '**आलु**', '**खा**'
aalu khaanu	*aalu khaana*	*chhoraale pareekchyaamaa dui otaa bishayamaa aalu khaayechha*	*'chhoraa', 'parikchyaamaa', 'dui', 'otaa', 'bishaya', 'aalu', 'khaa'*
'come up empty-handed'		'My son got a zero in two subjects in the exam.'	

Table 2: An example showing the challenges with lemmatization: खानु and खाएछ have different lemmatized forms.

Figure 2: Word cloud showing the most common to the least common idioms in the dataset.

tokens or less. Adding the preceding or following sentences usually doubles the length. These details enable us to estimate the average input length when passing the data to the LLMs for efficient processing. . The three most frequent idioms are ठीक पार्नु (*theek paarnu*, 'to make right'), डाँडो काट्नु (*daando kaatnu*, 'to go far away'), and हात लाग्नु (*haat laagnu*, 'to get hold of something') as shown in the word cloud in Figure 2.

3.2 Challenges

To build our dataset, we needed a list of idioms along with the contexts in which they were used. This presented a bit of a "chicken-and-egg" problem. If we can obtain idioms from a source, why not also gather their contextual sentences from the same source? In many cases, such as ours, this is simply not available. Alternatively, if we have sentences with idioms, could we not simply extract the idioms automatically? Idioms appear infrequently in text, and identifying them within a specific context is challenging. This made the dataset creation process more complex than it might initially seem. Some of the issues encountered during the development of `neDIOM` are discussed below.

Lack of language resources: The availability of comprehensive idiom repositories, especially in low-resource languages, is scarce. While some sources offer lists of idioms[5], these compilations are often quite small. Furthermore, even after a list of idioms has been collected, identifying pertinent contexts that include these idioms is quite a challenge. For instance, within a corpus of 392K Nepali documents, we were able to extract contexts for only 271 (91%) of the idioms in our list. Of those, many were filtered out during manual annotation (described in the next section), leaving only about 191 (64%) of the idioms in the final `neDIOM` dataset.

Idiom Detection and Context Creation Challenges: Even with a list of idioms, the variation in how idiomatic expressions appear in text adds to the challenge. For instance, consider the idiom नाक खुम्च्याउनु (*naak khumchyaunu*, 'to turn up your nose') and the sentence कोठा सफा नगरेको देखेर आमाले **नाक खुम्च्याउनु** भयो । (*kothaa safaa nagareko dekhera aamaale naak khumchyaunu bhayo*, 'Mom turned up her nose upon seeing the room was not cleaned.'). The idiom's form has been altered in the sentence, making it difficult to identify its usage by a simple pattern-based search. On the other hand, using the complete expression for searching mostly resulted in null matches. Our attempts at employing similarity-based methodologies to identify sentences containing idioms also proved to be unsatisfactory.

Prior work used lemmatization to deconstruct idioms and locate them within sentences (Tedeschi et al., 2022; Fadaee et al., 2018). Nevertheless, this approach proved to be insufficient for Nepali where an existing Nepali lemmatizer[6] failed to identify the idioms within contexts, primarily because the lemmatized idioms within the contexts differed from the lemmatized version of the idiom, leading to no match. For example, in Table 2, खानु (*khaanu*) is lemmatized to खान *khaana* and खाएछ (*khaayecha*)

[5] Idioms fall under the category of multiword expressions https://aclweb.org/aclwiki/Multiword_Expressions

[6] https://github.com/dpakpdl/NepaliLemmatizer

to खा (*khaa*) although both of those words have the same uninflected form. There is a need for developing better lemmatization tool for Nepali's typology.

3.3 Data Annotation

In the data annotation phase, given an idiom along with sentences *S1*, *S2*, and *S3*, we asked the annotators to assess the coherence and relevance of the provided contextual sentences. This step was crucial to address any potential noise in the data collection step. The annotation process can be summarized as follows:

1. If the annotators considered the sentences to be coherent, the sample was labeled as *(I)diomatic* if the idiom was used in its idiomatic sense or labeled as *(L)iteral* if the idiom was used in a literal sense. Then the position of the idiom within *S2* was marked using tokens "###".
2. If the sentences were not deemed coherent, the sample was labeled as *NA*.

To ensure high-quality annotations, we engaged three annotators, all native Nepali speakers with a minimum of higher secondary education. Initially, each annotator annotated 10 sample sentences and their methodology and results were discussed in order to establish a consistent baseline for annotation. Then, the entire set of 1,216 samples was annotated separately by two annotators. Next, the annotations underwent a final review by the third expert annotator. It was discovered that there were discrepancies in 26 annotations between the two annotators. Overall, Annotator #1 had 2 incorrect annotations, while Annotator #2 had 24 incorrect annotations. These discrepancies were rectified in the final version of the dataset. Discarding the 'NA' samples (about 56% of the data) helped to filter out noisy or irrelevant samples, and collectively, the process yielded 408 'I' samples and 118 'L' samples, a total of 526 samples with 191 unique idioms. The higher ratio of 'I' labels in the dataset suggests that most of these idioms are typically used in idiomatic senses rather than a literal sense.

4 Experiments

4.1 Task Formulation

The new `neDIOM` dataset can facilitate several idioms-related tasks such as idiom identification, idiomaticity detection, generating continuations in idiomatic contexts, or with some additional annotations, idiom translation, and sentiment analysis. We explore the dataset further in the classic yet challenging task of idiomaticity detection. Given the context and/or the associated idiom, the task is to identify whether the idiom has been used in a literal or idiomatic sense in the context. This task can provide insights into a model's ability to distinguish between non-compositional figurative and literal meanings.

4.2 Experimental Setup

We used four different multilingual language models: XLM-R-279m[7] (Conneau et al., 2020), BLOOM-560m[8] (Scao et al., 2023), BLOOM-1b1[8], BLOOM-3b[8], BLOOM-7b[8], LlaMa2-7b[9] (Touvron et al., 2023), and GPT-3.5[10].

Out of the 526 samples in our dataset, 506 were used for testing and the remaining 20 for fine-tuning the models under two settings: 5-shot setting where the training data consisted of a total of 10 samples (5 from each label); and 10-shot setting where the training data consisted of 20 samples (10 from each label). We also report results of experiments under the zero-shot setting where no training data is used. The inputs were prepared in 8 ways: *S2* only, *S1+S2* only, *S2+S3* only, *S1+S2+S3* only, with each of these four variants used with or without idioms.

In zero-shot setting, since the models were not originally fine-tuned for our classification task, we applied a log-likelihood method, calculating the likelihood for each label based on the model's next-word predictions, and selected the label with the highest likelihood. For the classification task, the results are reported in terms of macro-averaged F1 scores across all the models. Additional implementation details are included in Appendix A.

5 Results and Discussion

Idiomatic vs. Literal Classification: Table 3 presents the results of our classification experiment. A mediocre F1 score indicates that the model's performance in distinguishing between literal and idiomatic labels was subpar, implying that it struggled

[7] https://huggingface.co/xlm-roberta-base

[8] https://huggingface.co/docs/transformers/model_doc/bloom

[9] https://huggingface.co/docs/transformers/v4.34.1/model_doc/llama2

[10] https://platform.openai.com/docs/models/gpt-3-5

Models	Zero Shot (w/ idioms)	Zero Shot (w/o idioms)	5-shot (w/ idioms)	5-shot (w/o idioms)	10-shot (w/ idioms)	10-shot (w/o idioms)
S2						
XLM-R	0.44	0.18	0.50	0.44	0.44	0.19
BLOOM-560m	0.50	**0.52**	0.44	**0.52**	0.47	0.18
BLOOM-1b1	0.50	0.50	0.44	0.19	0.47	0.19
BLOOM-3b	0.46	0.37	0.19	0.44	0.44	0.20
BLOOM-7b	0.44	0.44	-	-	-	-
Llama2-7b	0.44	0.19	-	-	-	-
GPT-3.5	0.50	0.47	0.47	0.51	**0.52**	0.50
S1+S2						
XLM-R	0.44	0.44	0.23	**0.47**	0.44	0.46
BLOOM-560m	0.29	0.35	0.18	**0.47**	0.33	0.44
BLOOM-1b1	0.48	0.51	0.46	0.44	0.18	0.19
BLOOM-3b	0.38	0.34	0.46	0.45	0.20	0.20
BLOOM-7b	0.19	0.32	-	-	-	-
Llama2-7b	0.44	0.18	-	-	-	-
GPT-3.5	**0.53**	0.48	0.44	0.4	**0.56**	0.47
S2+S3						
XLM-R	0.44	0.44	0.18	**0.51**	0.44	0.46
BLOOM-560m	0.28	0.33	0.42	0.44	0.2	0.44
BLOOM-1b1	**0.47**	**0.47**	0.46	0.18	0.44	0.22
BLOOM-3b	0.34	0.33	0.43	0.43	0.18	0.18
BLOOM-7b	0.44	0.44	-	-	-	-
Llama2-7b	0.17	0.44	-	-	-	-
GPT-3.5	0.46	**0.47**	0.44	**0.51**	0.44	**0.50**
S1+S2+S3						
XLM-R	0.44	0.44	0.18	0.53	0.44	0.50
BLOOM-560m	0.25	0.31	0.20	0.18	0.31	0.54
BLOOM-1b1	0.47	0.48	0.44	0.18	0.45	0.32
BLOOM-3b	0.31	0.30	0.18	0.18	-	-
BLOOM-7b	0.19	0.44	-	-	-	-
Llama2-7b	-	-	-	-	-	-
GPT-3.5	**0.51**	0.49	**0.55**	0.53	**0.55**	0.54

Table 3: F1 score results of the experiments run on various models under zero shot, 5-shot, and 10-shot settings. (w/ idioms) refers to the settings where idioms are present in the input, while (w/o idioms) indicates inputs without idioms. The models in bold represent the best performance for the corresponding setting.

to accurately classify both types of expressions. This suboptimal performance stresses the need for further refinement and investigation into enhancing the model's capabilities in this particular classification task.

Effect of With Idioms vs. Without Idioms: To assess the potential impact of explicitly informing the models about the presence of idiomatic expressions, we conducted each experiment in two distinct setups. In the "with idioms" setup, the input consisted of the context sentence(s) along with the associated idiom phrase, while in the "without idioms" setup, we presented the context without specifying the idiom.

As illustrated in Figure 3, the results revealed that the presence or absence of idiomatic expressions obtained mixed results. In certain instances, it led to performance enhancements, while in others, it did not yield significant improvements. This fluctuation in outcomes can likely be attributed to the models' limited familiarity with Nepali idiomatic expressions, which consequently constrained them to limited classification decisions.

Zero-shot vs Few-shot: The results of our experiments investigating whether fine-tuning led to improved predictions are plotted in Figure 4. We observe that the benefits of fine-tuning are rather limited, with only a few notable exceptions. Our initial assumption was that the LLMs, having been trained on extensive corpora, would adapt well to low-resource languages after some fine-tuning. Additionally, LLMs trained on substantial datasets from the same language family, even if they lack significant data from the low-resource language, would bring about cross-lingual benefits. However, our results show that few-shot fine-tuning did not

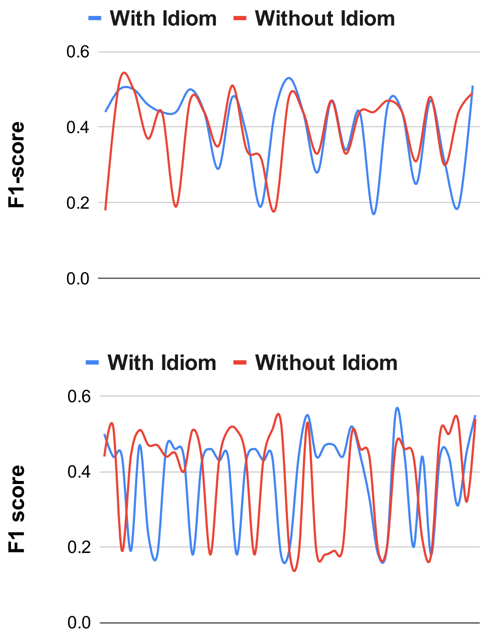

Figure 3: The line charts showing the averaged F1 scores under zero-shot setting (above) and both few-shot settings (below). Each data point on the x-axis represents a specific combination of model and context size.

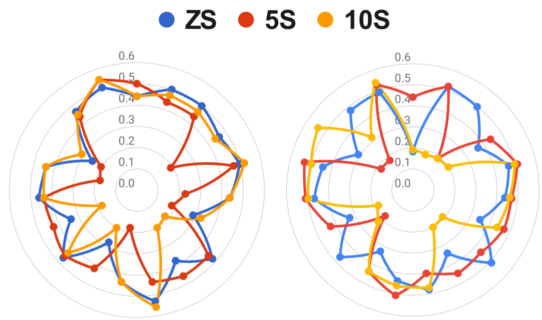

Figure 4: Plot showing the performance of the models under zero-shot (ZS), 5-shot (5S), and 10-shot (10S) settings with idiom (left) and without idiom (right).

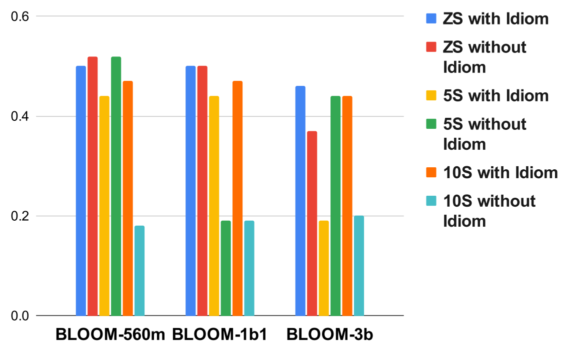

Figure 5: F1 scores of various sizes of BLOOM models when *S2* is used as input for idiom classification in zero-shot (ZS), 5-shot (5S), and 10-shot (10S) settings.

bring any additional gains, leaving significant room for enhancing LLM performance in low-resource language scenarios.

Impact of Model Size: In our experiments, we included several models of different sizes from the BLOOM family of models which allows us to draw insights regarding the comparable performance of smaller vs. larger models. The results are plotted in Figure 5. Curiously, contrary to the expectation that larger models within the same architecture would yield improved performance, the results do not consistently support this hypothesis. While there are minor enhancements in the 10-shot setting when idioms are not explicitly provided, the performance across other cases exhibits inconsistency. This phenomenon may be attributed to the shared training data for all three model variations (Scao et al., 2023). With an increase in model parameters, it appears that the available training data for low-resource languages may not be sufficient to adequately inform the expanded model capacity.

Effect of Surrounding Context: To evaluate the impact of the surrounding context on the comprehension of both idiomatic and literal scenarios, we conducted experiments in four distinct contexts: *S2*, *S1+S2*, *S2+S3*, and *S1+S2+S3*. Table 3 indicates that the sole instance of improved performance, associated with an increase in context, was observed with GPT-3.5. Performance saw a boost when all three context components – *S1*, *S2*, and *S3* – were provided, in comparison to scenarios where only *S1*, *S1+S2*, or *S2+S3* were presented. For all other models, it appears that using just *S2* is satisfactory and strikes a good balance between performance and efficiency.

6 Conclusion

In this study, we introduced a novel dataset, `neDIOM`, designed to facilitate research on idioms in low-resource languages, with a focus on Nepali. The dataset boasts high-quality content, as it was meticulously evaluated through manual assessment. Despite LLMs being extensively trained on data from high-resource languages within the same language family, their performance in low-resource language contexts fell short of expectations, even after fine-tuning. This highlights the urgency of making LLMs more inclusive to ensure their benefits are accessible to a broader population.

Limitations

We conducted only zero-shot experiments for some models due to resource limitations. Moreover, the data used was sourced from the internet, which may not fully represent all domains. As a low-resource language, we face challenges in finding abundant and high-quality online resources, such as literature books.

Our research identified several avenues for further exploration.

- First, there is a need for additional resources to create a more extensive and representative collection of Nepali idioms, with more fine-grained annotations.
- Second, it is important to refine the lemmatization methods to ensure consistency across various contexts when processing Nepali text, that will eventually help in automatic collection of idioms.
- Moving forward, our future plans also involve leveraging the positional information of idioms within the dataset to investigate how well LLMs can detect idiom positions.
- Additionally, we aim to develop techniques to enhance the models' performance when dealing with input containing idiomatic expressions in Nepali.

Ethical Considerations

Given that the dataset is sourced from a corpus comprising internet articles, it is possible that the texts may include content that could be potentially offensive to certain groups of people. Language models may inadvertently interpret idioms in ways that were not intended, as these idioms often express multiple meanings. Additionally, there are instances where specific idioms are closely tied to a particular culture's worldview, and this perspective may not necessarily align with the beliefs of other groups. The annotators received fair compensation for their work.

References

Ruchit Agrawal, Vighnesh Chenthil Kumar, Vigneshwaran Muralidharan, and Dipti Misra Sharma. 2018. No more beating about the bush: A step towards idiom handling for indian language nlp. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Tuhin Chakrabarty, Yejin Choi, and Vered Shwartz. 2022a. It's not rocket science: Interpreting figurative language in narratives. *Transactions of the Association for Computational Linguistics*, 10:589–606.

Tuhin Chakrabarty, Arkadiy Saakyan, Debanjan Ghosh, and Smaranda Muresan. 2022b. Flute: Figurative language understanding through textual explanations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7139–7159.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Marzieh Fadaee, Arianna Bisazza, and Christof Monz. 2018. Examining the tip of the iceberg: A data set for idiom translation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Meghdad Farahmand, Aaron Smith, and Joakim Nivre. 2015. A multiword expression data set: Annotating non-compositionality and conventionalization for english noun compounds. In *Proceedings of the 11th Workshop on Multiword Expressions*, pages 29–33.

Anduamlak Abebe Fenta and Seffi Gebeyehu. 2023. Automatic idiom identification model for amharic language. *ACM Transactions on Asian and Low-Resource Language Information Processing*.

Hessel Haagsma, Johan Bos, and Malvina Nissim. 2020. Magpie: A large corpus of potentially idiomatic expressions. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 279–287.

Ioannis Korkontzelos, Torsten Zesch, Fabio Massimo Zanzotto, and Chris Biemann. 2013. Semeval-2013 task 5: Evaluating phrasal semantics. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 39–47.

Anoop Kunchukuttan and Om Prakash Damani. 2008. A system for compound noun multiword expression extraction for hindi. In *6th International. Conference on Natural Language Processing*, pages 20–29. Citeseer.

Diego Moussallem, Mohamed Ahmed Sherif, Diego Esteves, Marcos Zampieri, and Axel-Cyrille Ngonga Ngomo. 2018. LIdioms: A multilingual linked idioms data set. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Nabaraj Neupane. 2018. Translating idioms from nepali into english. *Translation Today*, 12:83.

Caroline Pasquer, Agata Savary, Carlos Ramisch, and Jean-Yves Antoine. 2020. Verbal multiword expression identification: Do we need a sledgehammer to crack a nut? In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3333–3345, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Jing Peng, Anna Feldman, and Hamza Jazmati. 2015. Classifying idiomatic and literal expressions using vector space representations. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 507–511, Hissar, Bulgaria. INCOMA Ltd. Shoumen, BULGARIA.

Rhitabrat Pokharel and Ameeta Agrawal. 2023. Generating continuations in multilingual idiomatic contexts. In *Proceedings of the 3rd Workshop on Multi-lingual Representation Learning (MRL)*, pages 292–301, Singapore. Association for Computational Linguistics.

Jipeng Qiang, Yang Li, Chaowei Zhang, Yun Li, Yi Zhu, Yunhao Yuan, and Xindong Wu. 2023. Chinese idiom paraphrasing. *Transactions of the Association for Computational Linguistics*, 11:740–754.

Bahar Salehi, Paul Cook, and Timothy Baldwin. 2018. Exploiting multilingual lexical resources to predict mwe compositionality. In *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, volume 2, page 343. Language Science Press.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina Mcmillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, Dragomir Radev, Eduardo González Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Bar Natan, Francesco de Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady Elsahar, Hamza Benyamina, Hieu Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jörg Frohberg, Joseph Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro von Werra, Leon Weber, Long Phan, Loubna Ben Allal, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, María Grandury, Mario Šaško, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad A. Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nurulaqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Villegas, Peter Henderson, Pierre Colombo, Priscilla Amuok, Quentin Lhoest, Rheza Harliman, Rishi Bommasani, Roberto Luis López, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma, Shayne Longpre, Somaieh Nikpoor, Stanislav Silberberg, Suhas Pai, Sydney Zink, Tiago Timponi Torrent, Timo Schick, Tristan Thrush, Valentin Danchev, Vassilina Nikoulina, Veronika Laippala, Violette Lepercq, Vrinda Prabhu, Zaid Alyafeai, Zeerak Talat, Arun Raja, Benjamin Heinzerling, Chenglei Si, Elizabeth Salesky, Sabrina J. Mielke, Wilson Y. Lee, Abheesht Sharma, Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, Debajyoti Datta, Eliza Szczechla, Gunjan Chhablani, Han Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jos Rozen, Leo Gao, Lintang Sutawika, M Saiful Bari, Maged S. Al-Shaibani, Matteo Manica, Nihal Nayak, Ryan Teehan, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H. Bach, Taewoon Kim, Tali Bers, Thibault Fevry, Trishala Neeraj, Urmish Thakker, Vikas Raunak, Xiangru Tang, Zheng-Xin Yong, Zhiqing Sun, Shaked Brody, Yallow Uri, Hadar Tojarieh, Adam Roberts, Hyung Won Chung, Jaesung Tae, Jason Phang, Ofir Press, Conglong Li, Deepak Narayanan, Hatim Bourfoune, Jared Casper, Jeff Rasley, Max Ryabinin, Mayank Mishra, Minjia Zhang, Mohammad Shoeybi, Myriam Peyrounette, Nicolas Patry, Nouamane Tazi, Omar Sanseviero, Patrick von Platen, Pierre Cornette, Pierre François Lavallée, Rémi Lacroix, Samyam Rajbhandari, Sanchit Gandhi, Shaden Smith, Stéphane Requena, Suraj Patil, Tim Dettmers, Ahmed Baruwa, Amanpreet Singh, Anastasia Cheveleva, Anne-Laure Ligozat, Arjun Subramonian, Aurélie Névéol, Charles Lovering, Dan Garrette, Deepak Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey Schoelkopf, Jan-Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Jordan Clive, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat, Miruna Clinciu, Najoung Kim, Newton Cheng, Oleg Serikov, Omer Antverg, Oskar van der Wal, Rui Zhang, Ruochen Zhang, Sebastian Gehrmann, Shani Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdeněk Kasner, Alice Rueda, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ana Santos, Anthony Hevia, Antigona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh Hajihosseini, Bahareh Behroozi, Benjamin Ajibade, Bharat Saxena, Carlos Muñoz Ferrandis, Danish Contractor, David Lansky, Davis David, Douwe Kiela, Duong A. Nguyen, Edward Tan, Emi Baylor, Ezinwanne Ozoani, Fatima Mirza, Frankline Ononiwu, Habib Rezanejad, Hessie Jones, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jesse Passmore, Josh Seltzer, Julio Bonis Sanz, Livia Dutra, Mairon Samagaio, Maraim Elbadri, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael Mckenna, Mike Qiu, Muhammed Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nour Fahmy, Olanrewaju Samuel, Ran An, Rasmus Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas Wang, Sourav Roy, Sylvain Viguier, Thanh Le, Tobi Oyebade, Trieu Le, Yoyo Yang, Zach Nguyen, Abhinav Ramesh Kashyap, Alfredo Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Singh, Benjamin Beilharz, Bo Wang, Caio Brito, Chenxi Zhou, Chirag Jain, Chuxin Xu, Clémentine Fourrier, Daniel León Periñán, Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrimann, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U. Vrabec, Imane Bello, Ishani Dash, Jihyun Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthik Rangasai Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pàmies, Maria A Castillo, Marianna Nezhurina, Mario Sänger, Matthias Samwald, Michael Cullan, Michael Weinberg, Michiel de Wolf, Mina Mihaljcic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patrick Haller, Ramya Chandrasekhar, Renata Eisenberg, Robert Martin, Rodrigo Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shlok S Deshmukh, Shubhanshu Mishra, Sid Kiblawi, Simon Ott, Sinee Sang-Aroonsiri, Srishti Kumar, Stefan Schweter, Sushil Bharati, Tanmay Laud, Théo Gigant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak, Yash Shailesh Bajaj, Yash Venkatraman, Yifan Xu, Yingxin Xu, Yu Xu, Zhe Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. 2023. BLOOM: A 176B-Parameter Open-Access Multilingual Language Model. Working paper or preprint.

Minghuan Tan and Jing Jiang. 2021a. Does bert understand idioms? a probing-based empirical study of bert encodings of idioms.

Minghuan Tan and Jing Jiang. 2021b. Learning and evaluating chinese idiom embeddings. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1387–1396.

Kenan Tang. 2022. Petci: A parallel english translation dataset of chinese idioms. *arXiv preprint arXiv:2202.09509*.

Harish Tayyar Madabushi, Edward Gow-Smith, Carolina Scarton, and Aline Villavicencio. 2021. AStitchInLanguageModels: Dataset and methods for the exploration of idiomaticity in pre-trained language models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3464–3477, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Simone Tedeschi, Federico Martelli, and Roberto Navigli. 2022. Id10m: Idiom identification in 10 languages. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2715–2726.

आचार्य, ऋषिकेश. 2020. सरल नेपाली व्याकरण बोध र अभिव्यक्ति. JBD Publication Pvt. Ltd., Kathmandu. Available in Nepali language.

Ye Tian, Isobel James, and Hye Son. 2023. How are idioms processed inside transformer language models? In *Proceedings of the The 12th Joint Conference on Lexical and Computational Semantics (* SEM 2023)*, pages 174–179.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Chujie Zheng, Minlie Huang, and Aixin Sun. 2019. ChID: A large-scale Chinese IDiom dataset for cloze test. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 778–787, Florence, Italy. Association for Computational Linguistics.

Jianing Zhou, Ziheng Zeng, and Suma Bhat. 2023. Clcl: Non-compositional expression detection with contrastive learning and curriculum learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 730–743.

Jianing Zhou, Ziheng Zeng, Hongyu Gong, and Suma Bhat. 2022. Idiomatic expression paraphrasing without strong supervision. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11774–11782.

A Implementation Details

Fine-tuning experiments were run on an A100 Tensor Core GPU, employing the AdamW optimizer for three epochs in each case. Due to resource limitations, fine-tuning was carried out for all models except for BLOOM-7b and LlaMa2-7b, for which only zero-shot experiments were conducted. We determined the maximum token length for each context based on the tokens generated by the models, ensuring that all context was encompassed in the model experiment. This length ranged from 200 subword tokens for S2 in the BLOOM-560m model to 1300 tokens for the combined context of S1, S2, and S3 in the LlaMa2 model. This approach ensured an efficient use of computational resources.

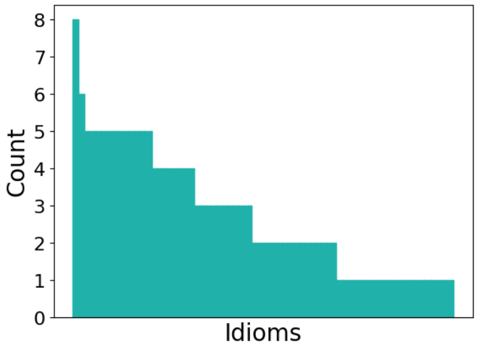

Figure 6: Histogram of idioms present in `neDIOM`.

B Exploratory Analysis

There are 191 unique idioms in the neDIOM dataset, with the minimum idiom length of 2 words and a maximum length of 4 words. Figure 6 presents the histogram of the idioms. While one idiom appears in 8 contexts, most idioms appear only once or twice in the dataset.