Language Resources and Evaluation (2025) 59:1613–1629
https://doi.org/10.1007/s10579-024-09782-4

PROJECT NOTES

Exploratory Analysis of Rinconada Bikol Language-Nabua Text Corpus

Joseph Jessie S. Oñate[1,2] · **Tiffany Lyn O. Pandes**[1,2]

Accepted: 27 September 2024 / Published online: 15 October 2024

Abstract

Any text corpus is an essential aspect of a language in which study is carried out or continues since the corpus of that language can be used to perform functional language study. Hence, this study focused on the development of the Rinconada Bikol Language-Nabua (RBL-Nabua) text corpus. RBL-Nabua texts were collected in social media, learning modules and other sources that will serve as the dataset. Corpus Linguistics tools and techniques were used to analyze and evaluate the developed RBL-Nabua corpus, such as tokenization, term frequency, parts-of-speech tagging, collocation, and visualization. This study managed to collect five thousand seven-hundred twenty (8754) RBL-Nabua words, and manually POS and Polarity tagged. A bag-of-words sentiment analyzer was also implemented by training five hundred RBL—Nabua sentences labeled with positive, negative, and neutral using different machine learning algorithms. Experimental result shows that it obtained an F1-Score of 57.83%, 52.08%, 39.98%, and 36.71% for Logistic Regression, Support Vector Machine, Decision Tree Classifier, and Gaussian Naive Bayes, respectively. This study could help develop natural language processing applications such as sentiment analysis, text classification, morphological analysis, language translation, and others in future studies. This study will also help in the preservation of the RBL-Nabua language as well as its affiliated culture.

Keywords Rinconada Bikol Language-Nabua · Corpus linguistic · Local language · Low-resourced language

✉ Joseph Jessie S. Oñate
josephjessie97@cspc.edu.ph

Tiffany Lyn O. Pandes
tiffpandes@cspc.edu.ph

[1] College of Computer Studies, Camarines Sur Polytechnic Colleges, National Highway, 4434 Nabua, Camarines Sur, Philippines

[2] AI Research Center for Community Development, Camarines Sur Polytechnic Colleges, National Highway, 4434 Nabua, Camarines Sur, Philippines

1 Introduction

Over the past years, corpus linguistics has grown faster, which addresses different varieties of linguistic issues. Any text corpus is an essential aspect of a language in which study is carried out or continues since the corpus of that language can be used to perform functional language study. Corpus-based techniques have also been developed to analyze how terms, such as commonality and systematic connection with other words, are used. Many people have been rethinking the nature of language as a result of work in corpus analysis and linguistics. Language has been influenced by expressions found in text corpora (Kübler and Zinsmeister, 2015).

However, while Natural Language Processing (NLP) research tackles a mere 20 out of the world's 7,000 languages, the vast majority remain under-explored. These languages, often imprecisely labeled as "low-resource," deserve more attention like Rinconada Bikol Language (RBL) - Nabua. Low-Resourced Language can be interpreted in a variety of ways, including as less studied, resource-scarce, computerized, privileged, rarely taught, or low density (Magueresse et al., 2020). Until now, very little to no effort has been made in developing text and corpora for low-resourced languages in this direction. No significant research work on Rinconada Bikol Language (RBL) text corpus has been found. No research on this language's data retrieval, syntactic parsing, sentiment analysis, and machine translation has been performed. RBL is widely used by six municipalities and cities in the province of Camarines Sur, Philippines. RBL is divided into two (2) dialects, "Sinabukid," as it refers to the highland dialect, and "Sinaranew," which refers to the lakeside dialect. These two dialects are composed of different variants. Sinabukid is composed of two variants, the Agta variant and the Iriga Variant, while Sinaranew is composed of four variants, the Baao variant, the Bula-Pili variant, the Bato variant, and the Nabua variant will be the focus of this study. Another problem being faced by this low-resourced local language is that the menace of being extinct with the rise of non-local language books, movies, stories and other multimedia being delivered through the internet. Also, The decision made consciously or unconsciously by families to prioritize regional and national languages over their own language and culture is the biggest threat to the survival of the Philippine languages Rubi and Molina (2020).

Hence, the main focus of this study is to construct a corpus for Rinconada Bikol Language (RBL) - Nabua variant. This study aims to develop an RBL-Nabua corpus as an initial phase of developing natural language applications on the RBL-Nabua variant. RBL-Nabua corpus is a collection of words or spoken language of Nabueños in Rinconada Bikol language, which could be used in different natural language processing applications. Moreover, researchers in numerous disciplines, including language learning and psycholinguistics, often turn to corpora - vast collections of text - as a natural foundation for building word lists. Hence, this study also intends to establish a wordlist of RBL - Nabua.

This study could also help develop natural language processing applications such as sentiment analysis, text classification, morphological analysis, language translation, and others in future studies. These applications could help in terms

of tourism and promoting the local language of Rinconada, especially in Nabua, Camarines Sur, Philippines. This study will also open the possibilities of more researches focusing on local languages, and the corpus will also serve as a way of preserving the local language of Nabua, Camarines Sur, Philippines as well as its affiliated culture.

2 Related literature

2.1 Corpus and its developments

Today, technologies are now being used in corpus development by collecting large amounts of text from different sources and being stored in a computer (Biber et al., 1998). In this case, the computer will be able to conduct analysis on patterns on word associations to the stored data in a far more complex way by using different tools and techniques such as natural language toolkit (Loper and Bird, 2002) in a more reliable and easy way.

A text corpus is an electronic database for language researchers or natural language processing applications. Corpora are a critical resource in corpus linguistics to research the language represented in samples or real-world text. There are many tools in the written text that can be used to build a corpus (Ghayoomi et al., 2010). Corpus linguistics gives a more accurate assessment of language than the picture, perception, and anecdotes. Almost every language pattern can be analyzed through a corpus-based analysis (Krieger, 2003). Aside from serving as a valuable resource for NLP, the synchronous character of language corpora has allowed tracking not just significant linguistic changes but also the possible evolving cultural trends that underpin such changes (Tsou and Chin, 2010).

Medium in acquiring data in the development of a corpus was also introduced. Adolphs and Knight (2010) stipulated that spoken corpus is a unique resource for exploring natural conversation and that the increasing trend in spoken corpus growth bears witness to the significance it brings to a wide variety of scientific communities. Nevertheless, it is also said that a written corpus is more manageable than creating a spoken corpus. However, there are still many difficulties involved in the creation of written corpus, such as gaining access to these texts (Nelson, 2010).

In terms of tools, techniques, and methodology in the analysis of the corpus, Hasko (2012) presented a qualitative analysis of the corpus. Qualitative analysis is a technique for the study of linguistic phenomena that is focused on authentic contact circumstances which are stored digitally in a language corpora, which can be retrieved, reclaimed, and analyzed through a computer. This was also supported by the book of Schäfer and Bildhauer (2013) which introduced some tools and techniques in developing a web corpus.

Several pieces of study on corpus development evolved, such as in Urdu Website (Becker and Riaz, 2002), Japanese parsed corpus (Kuroashi and Nagao, 1998), Sindi, one of the major languages spoken in Pakistan (Rahman and Mutee, 2015), Brazilian Portuguese Schramm et al. (2000), Igbo, an African Language (Onyenwe

et al., 2014), Ireland (Kilgarriff et al., 2007), Thai (Boriboon et al., 2009), and Sindhi (Dootio and Wagan, 2021).

In the Philippines, there are several studies about the corpora of Philippine languages. One of which is the study of Dita et al. (2009) that focuses on an online repository called "Palito" was developed as a tool for data collection and storage of Philippine corpora, which are automatically indexed by the repository so they may be found quickly. However, most of the recent studies are focused on Tagalog corpus also, limited researches and studies were conducted concerning other Philippines languages, and no researches are aiming for the corpora development of Rinconada Bikol Language, especially the Nabua variant, which is a low-resourced language. In the next section, studies focusing on low-resourced Philippine languages will be discussed.

2.2 Corpus studies on Philippine low-resourced local languages

In the Philippines, there are over 120 spoken languages, all considered low-resourced, even the two (2) most common spoken languages, Tagalog and Cebuano (Fernandez and Adlaon, 2022) and these two languages can be seen in recent studies. The study of Magueresse et al. (2020) published the first large-scale, openly available, preprocessed unlabeled text corpora in the low-resource Filipino language, which named as "WikiText-TL-39," and demonstrated how transfer learning strategies like BERT and ULMFiT could be used to train robust classifiers in low-resource settings. In language translation, Adlaon and Marcos (2019) developed a Cebuano to Tagalog translator implemented using the OpenNMT framework and a recurrent neural network using a Bible dataset that contains monolingual corpus of different languages. Nevertheless, a wider scope of study on the acquisition of corpora for Philippine languages was done by Dimalen and Roxas (2007) using a technique called AutoCor which is used to automatically collect and categorize corpora of texts in related languages which, includes Bicolano. Other than that, no other studies focus on Bicolano corpus, especially the Rinconada Bikol Language.

3 Methodology

This study follows the development process consisting of three (3) steps: data gathering, development of text corpus, and evaluation and analysis are shown in Fig. 1.

3.1 Data gathering and pre-processing

Rinconada Bikol Language data was collected through social media platforms from Facebook posts and groups of Nabueños and Mother Tongue Based Language (MTBL) Learning Modules of Primary Schools. After data collection, basic

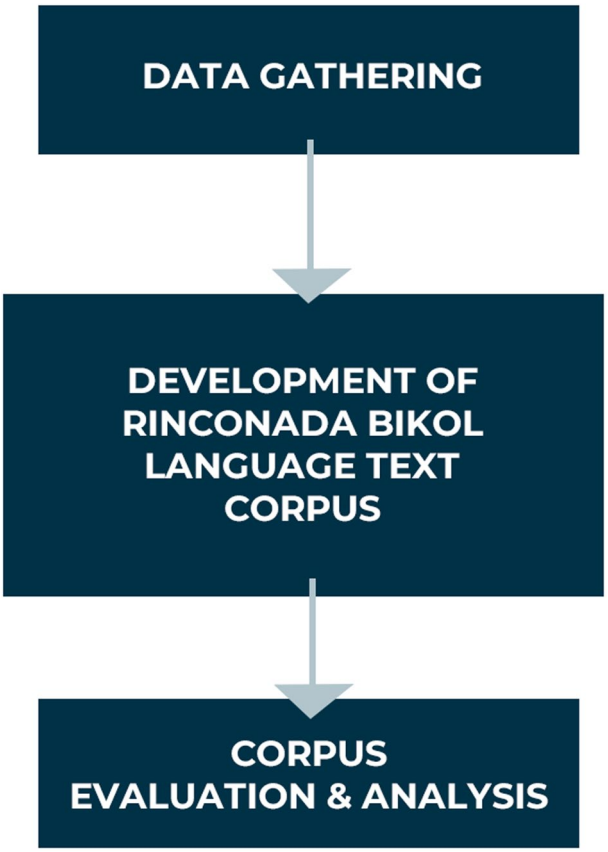

Fig. 1 Development process

Table 1 Summary of collected data

Source	Number of text	Percentage (%)
Social media	52	16.35
Learning modules	266	83.65
Total	318	100

cleanups and duplicate detection were conducted. These involve cleanups within the documents and the removal of documents that do not meet specific criteria. Shown in the Table 1 below is the summary of the collected data.

3.1.1 Social media

The researcher gathered data through social media posts from Facebook, and Facebook groups of Nabueños. A total of 52 texts were collected by manually

```
0   Ipost na po ang mga a mga nagbabalgar sa healt...
1   Di po kita mag ngata ngata ka dita nababayad n...
2   MAIRAK MAN PO KAMU SA MGA KABABAYAN TA ADING L...
3   ISI PO NAMU KAIPUHAN NINYO MAGKITA PERO ONOWON...
4   MGA KABABAYAN TA MGA PASSENGERS KIN PWEDE PO T...
5   SINABI NANG SENIOR CITIZENS and MINORS bawal l...
6   ATANG NAG POSITIVE SU KAPAMILYA EXPOSE IYA sig...
7   ATANG LSI positive sa antigen nagpauli pa sa p...
8   1 YEAR RESIDENCY naman prior to election naman...
9    ( "Malilingawan Mo man" ) Malilingawan mo man...
```

Fig. 2 Texts social media posts

scraping posts written in RBL. Each text contains words ranging from 50 to 500. Shown in Fig. 2 is a sample of collected texts from social media posts. One of the major problems in data gathering over the internet is that there is a limited number of written Rinconada-Nabua language texts available. Social media as an RBL source is only 16.35% of the total collected data in this study shown in Table 1. The majority of the data comes from the learning modules discussed in the next subsection.

3.1.2 Learning modules

Due to the limited amount of RBL texts in social media, the researchers requested copies of the learning module for the subject Mother Tongue Language, written in the RBL-Nabua variant, from the public school teachers. A total of 220 learning modules were acquired, which covers 83.65% of the total collected data in this study shown in Table 1. The learning modules were in PDF format, converted into a text file using the python conversion library, and saved as a txt file. Each learning module has 10–20 pages of 1000–2000 words (Fig. 3).

```
0    MGA GIGIBUWUN SA PAGKATU-UD  SA MTB-MLE  Kuwar...
1    Para sa mga Kag-igin o Tagagabay   Tanganing l...
2    Learners Packet  sa  MTB-MLE  Kwarter   Bilang...
3    Learners Packet  sa  MTB-MLE  Kwarter  Bilang ...
4              Dios maray na alduw kanimo Siguro ma...
5                 Ngaran GradoSeksyon          Maes...
6    SMILE Learners Packet Ngaran GradoSeksyon   Ma...
7    MGA GIGIBUWUN SA PAGKATU-UD SA MTB-MLE  Kwarte...
8    MGA GIGIBUWUN SA PAGKATU-UD SA MTB-MLE  Kwarte...
9    Ngaran   GradoSeksiyon    Maestro Maestra   Pe...
```

Fig. 3 Texts from learning modules

After saving the data into a text file, data pre-processing was conducted by removing the unnecessary words from the data, such as URL and invalid characters, using regular expression (regex) conditions in python.

3.2 Development of text corpus: linguistic process

At this stage, linguistic post-processing was conducted, such as tokenization and part-of-speech tagging. Each sub-section below discusses the steps of each linguistic process. Natural Language Toolkit (NLTK) was used to conduct the linguistic processes. NLTK is a suite of open-source program modules, tutorials, and problem sets, providing ready-to-use computational linguistics courseware. NLTK covers symbolic and statistical natural language processing and is interfaced with annotated corpora.

3.2.1 Tokenization

Tokenization splits the raw text or phrases into words called tokens. These tokens can contribute to the meaning comprehension or creation of the NLP model. The tokenization helps to clarify the text's context by evaluating the word series. RBL texts were tokenized into sentences. Sentences that are written in pure non-RBL such as English and Tagalog were manually checked and removed to avoid any errors in the analysis. However, RBL texts containing a small amount of English or Tagalog words were not removed to preserve the context of the sentences.

After sentence tokenization, word tokenization was also conductedto extract the words from the sentences. Shown below is the sample RBL-Nabua text word tokenized using the Python Natural Language Processing Toolkit (NLTK) Regular Expression Tokenizer. After the tokenization, these tokens were saved in a dataframe as a word list.

"Sa mga nagoonga tabi kun kuno migbalik a kuryente, oda pa malinaw na impormasyon na ipinaabot a NGCP kanamo kun kuno ninda ibabalik a kuryente."

[*'Sa', 'mga', 'nagoonga', 'tabi', 'kun', 'kuno', 'migbalik', 'a', 'kuryente', 'oda', 'pa', 'malinaw', 'na', 'impormasyon', 'na', 'ipinaabot', 'a', 'NGCP', 'kanamo', 'kun', 'kuno', 'ninda', 'ibabalik', 'a', 'kuryente'*]

3.2.2 Part-of-speech tagging

Part-of-the-speech tagging is a mechanism for a word in a text (corpus), based on both its meaning and its context, to refer to a specific part of the speech. Since there is no existing POS Tagged Rinconada-Nabua Language, the processed words were manually tagged by the researchers to ensure the integrity of the tag list. The researcher used the Penn Treebank POS tags(Santorini, 1990) for the manual tagging of collected words. Each word in the word list is tagged with its corresponding parts of speech. Table 2 is the example of POS-tagged words with their corresponding POST tag, English, and Tagalog meaning.

Table 2 Example of POS tagged words

Words	POS tag	Tagalog meaning	English meaning
kita	PRP$	tayo	we
ngata	WP	bakit	why
nababayad	VB	nakikita	seen
kalaban	NN	kalaban	opponent
mairak	JJ	kawawa	pitiful
kamu	PRP$	kayo	you
isi	VB	alam	know
sintabo	NN	pera	money
ading	TO	eto	this
alagad	CC	subalit	however

3.3 Evaluation and analysis

After the construction of the corpus, the overall technical quality of the corpus was determined and assessed, which includes:

- Term frequency is the measurement of how frequently a term occurs within a document. It is being calculated as the number of times a word appears in the text corpus.
- Frequency of Part-of-Speech tag, which indicates how often a POS tag appears in the POS tagged list.
- Collocation which are the phrases or statements that comprise numerous words and are likely to occur together. This is measured using the Pointwise Mutual Information (PMI) by getting the BiGram churned out by NLTK Collocation library to quantify the likelihood of two words appearing together, taking into consideration the possibility that it is caused by the frequency of the single terms. As a result, the algorithm calculates the (log) chance of co-occurrence, scaled by the product of the single probability of occurrence:

$$PMI(a, b) = \log(\frac{P(a, b)}{P(a)P(b)}) \tag{1}$$

- Visualization by using different graphical representations such as graphs and charts.
- A Bag-of-Words (BOW) model with Sentiment Analysis was also developed. Five hundred (500) sentences extracted from the text corpus were labeled as ‘Positive’, ‘Negative’, and ‘Neutral’ shown in Fig. 4. Sentences were labelled based on the whole context of the sentence not per word. This is to ensure that the label captures the whole context/tone of the sentence rather than each word which may result to poor classification. Moreover, these sentences were labeled by the researchers who are grassroots Nabueños. These sentences were cleaned by converting words to lowercase, keeping only letters by removing numbers and other symbols.

	texts	label
0	pirming magkaiba a magkabarkada	positive
1	nanrarakup iya sa isura sa salug	positive
2	tapos nang simbagan a modyul na adi	positive
3	iparumrum sa mga eskwela na ibalik tulus sa ma...	positive
4	ginibu a modyul na adi para kanimo	positive
5	kaipuwan mong sunudun saka simbagan na solo a ...	neutral
6	magayat sa tabang ka kanimong kagigin kung kin...	positive
7	isi ko kayangkaya mo di	positive
8	sigurado kong makatutuud saka mamumuya ka	positive
9	ingatan mo a modyul na adi	neutral

Fig. 4 Polarity-labeled sentences

After which, these sentences were fed into a sklearn TF-IDF Vectorizer to convert texts into a meaningful representation of numbers that can be utilized to fit into machine prediction algorithms such as Support Vector Machine, Decision Tree Classifier, Gaussian Naive Bayes, and Logistic Regression using the sklearn library. The dataset was split into eighty (80) percent training and twenty (20) percent testing data.

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} = \frac{2 * TP}{2 * TP + FP + FN} \tag{4}$$

F1 Score (equation 4) was used to evaluate the performance of the models, which combines the Precision (equation 2) and Recall (equation 3) scores. Recall evaluates how many of the positive class samples included in the dataset were properly recognized by the model, whereas Precision measures how many of the "positive" predictions produced by the model were accurate.

Fig. 5 Wordcloud of the collected Rinconada Bikol Language—Nabua variant

Table 3 Top ten words

Rank	Words	Frequency	Tagalog meaning	English meaning
1	*sa*	7741	*sa*	*in*
2	*a*	7017	*ang*	*the*
3	*na*	6685	*na*	*already*
4	*ka*	3611	*ikaw*	*you*
5	*mga*	3389	*mga*	*those*
6	*mo*	2413	*mo*	*you*
7	*saka*	1725	*at*	*and*
8	*o*	1495	*o*	*or*
9	*su*	1105	*yung*	*that*
10	*adi*	1057	*ito*	*this*

4 Experimental results

4.1 Linguistic processing

After data processing, a total of 12,292 sentences were generated with an average of 9 words per sentence. These sentences consists of 118,243 tokens (words). A total of eight thousand seven hundred fifty four (8754) unique RBL—Nabua words were collected and were visualized in a word cloud according to the frequency shown in Fig. 5.

Shown in Table 3 are the top ten words with their corresponding frequency. Also shown on the table are the corresponding English and Tagalog meaning of the top ten words; on the top is *'sa'*, which means *'in'* in English and *'sa'* in Tagalog, second in the rank is *'a'* which means *'the'* in English and *'ang'* in Tagalog, the third in rank is *'na'* which means *'already'* in English and *'na'* in Tagalog and others. These common words are grammatical functions such as demonstrative pronouns, adverbs, prepositions and others.

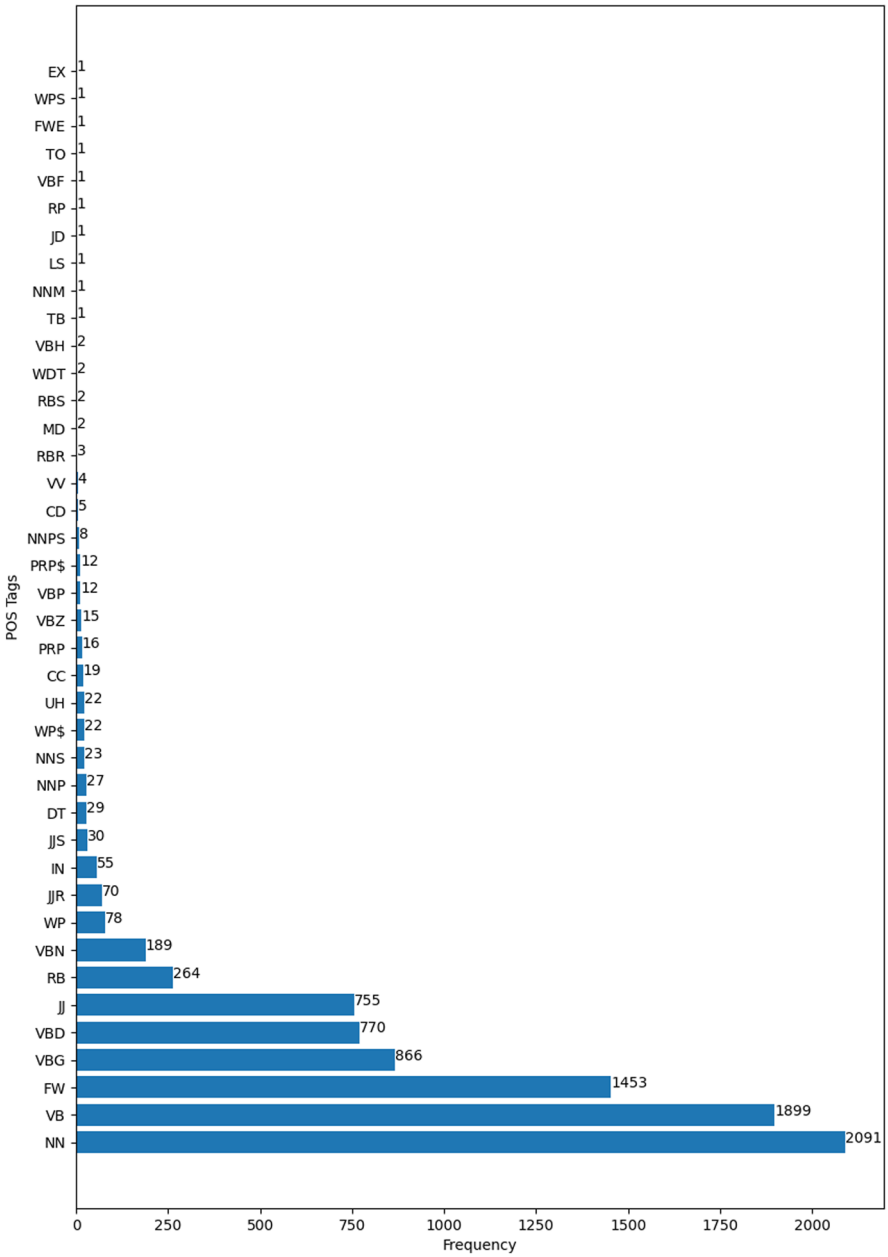

Fig. 6 POS tag frequency

Table 4 Collocation of RBL-Nabua corpus

Bigram	PMI Score
('usad', 'na')	2678.14
('a', 'mga')	2566.14
('modyul', 'na')	2321.45
('answer', 'sheet')	2315.36
('na', 'adi')	1789.04
('key', 'to')	1671.01
('to', 'correction')	1655.51
('a', 'modyul')	1306.91
('ka', 'mga')	1053.64
('tamang', 'simbag')	1038.58

Table 5 Performance of the sentiment analyzer

Model	Precision	Recall	F1 Score
Logistic Regression	58.69%	64.70%	57.83%
Support Vector Machine	53.49%	57.06%	52.08%
Decision Tree Classifier	43.56%	48.40%	39.98%
Gaussian Naive Bayes	40.59%	47.41%	36.71%

4.2 Evaluation and analysis

Figure 6 displays the frequency of POS tags in the tagged list. Noun (NN) leads with 2091 words; next in rank is Verb (VB) with 1899 words; Foreign word (FW) has 1453 words; 886 words for Verb, gerund or present participle (VBG); 770 words for Verb, past tense (VBD); Adjective (JJ) has 755 words; Adverb (RB) with 264 words; and others.

The researchers also conducted a collocation of the corpus. Collocation allows the extraction of multiword units of corpus data that can be utilized for lexicography, particularly technical translation. Collocation groups together comparable terms to distinguish the many senses of the word. Shown in Table 4 is the bigram of the RBL-Nabua text corpus with its Pointwise Mutual Information (PMI) Score.

The bigram with the highest PMI Score of 2678.14 is (*'usad'*, *'na'*) which means referring to *one which/one who* followed by (*'a'*, *'mga'*) with 2566.14 PMI Score which means *the* and the rest are words that are common in the learning modules since the majority of our data was obtained from learning modules written in RBL Nabua. However, majority of the bigrams are common words and grammatical functions that are being used by the Nabueños in daily life.

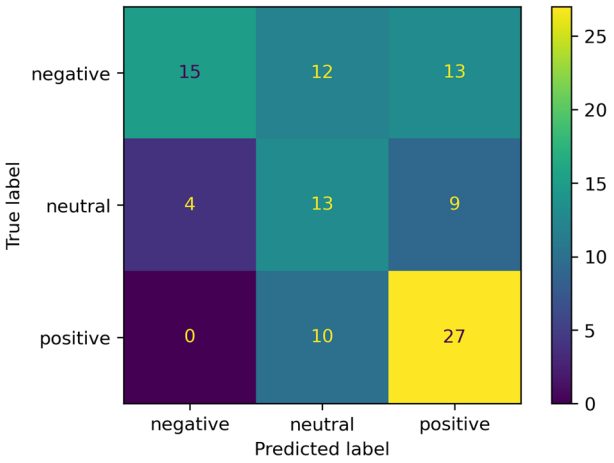

Fig. 7 Confusion matrix for support vector machine

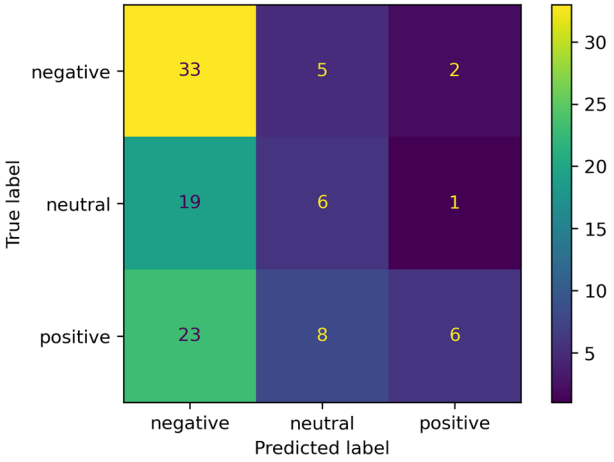

Fig. 8 Confusion matrix for Gaussian Naive bayes

4.2.1 BOW sentiment analysis

Shown in the Table 5 is the performance of the Bag-of-Words Sentiment Analyzer using the different machine learning models. In terms of Precision, the Logistic Regression model achieved the highest score of 58.69%, followed by the Support Vector Machine with a 53.49% score, then the Decision Tree Classifier with a score of 43.56%, and Gaussian Naive Bayes which has the lowest score of 40.59%.

In terms of Recall, Logistic Regression attained again the highest score of 64.70%, followed by Support Vector Machine with a 57.06% recall score. Subsequently, the

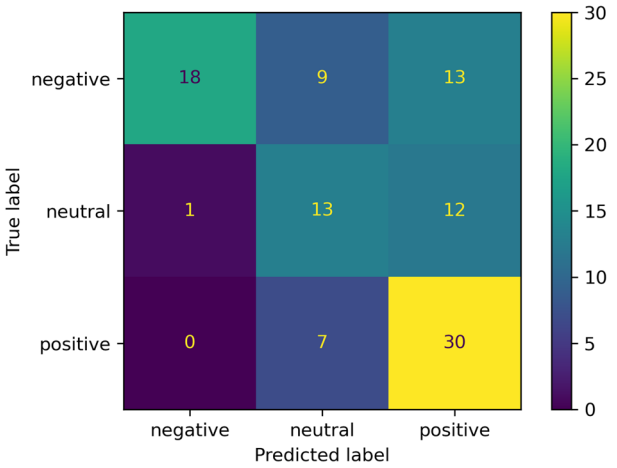

Fig. 9 Confusion matrix for logistic regression

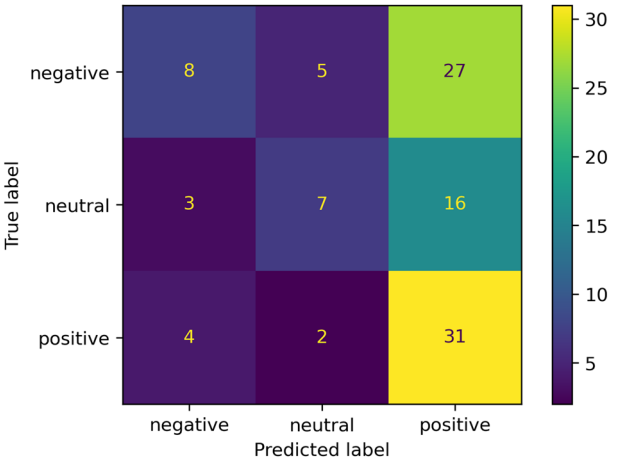

Fig. 10 Confusion matrix for decision tree classifier

Decision Tree Classifier with a score of 48.40%, and Gaussian Naive Bayes has the lowest score of 47.41%.

As to the F1 Score, Logistic Regression outscored all of the other models with a score of 57.83%. Support Vector Machine ranked second with 52.08%. Followed by Decision Tree Classifier with a 39.98% score. Last in rank is Gaussian Naive Bayes with a score of 36.71%.

Each model prediction was also visualized in a confusion matrix shown in Figs. 7, 8, 9, and 10. In Support Vector Machines and Logistic Regression, it can be seen that the majority of the classes are classified correctly. However, in

Gaussian Naive Bayes and Decision Tree Classifier only the negative and positive classes, respectively earned a high correct prediction.

In general, the highest F1 Score of 57.83% was attained by the Support Vector Machine. Even though the score is greater than 50%, this is still considered low compared to other studies since the baseline acceptable performance of a sentiment analyzer should be at around 80% above (Nafis and Awang, 2021). Nevertheless, this is still considered a breakthrough since this is the first time to test the RBL - Nabua in a Bag-of-Words Sentiment Analysis model and considering that the trained model has only five hundred (500) labeled texts.

5 Conclusion

This study primarily focused on the development of a text corpus for Rinconada Bicol Language—Nabua variant. RBL Nabua texts were collected in social media, learning modules, and other platforms. The researchers managed to gather 318 texts which contain tokens ranging from 50 to 2000. These texts were pre-processed by cleaning the data, tokenization. A total of 12,292 sentences were collected which consists of 118,243 tokens. These tokens were analyzed using term frequency. One of the top words are *'sa'*, *'a'*, *'na'*, and others which are mostly word used in grammatical functions. This study collected 8754 unique RBL—Nabua words. Furthermore, the words were also tagged with parts of speech. The highest number of words on the POS tagged list is the noun (NN) with 2091 words followed by a verb (VB) with 1899 words and others. Collocation analysis was also conducted to group together comparable terms to distinguish the many senses of the word. Collocation analysis shows that (*'usad'*, *'na'*) is the bigram with the highest PMI Score of 2678.14. Majority of the bigrams are associated with the grammatical functions and words commonly used in the learning modules. Moreover, a Bag-of-Words sentiment analyzer was also implemented using different machine learning algorithms. 500 RBL - Nabua sentences were labeled with polarity (positive, neutral, negative). The sentiment analyzer obtained an F1-Score of 57.83%, 52.08%, 39.98%, and 36.71% for Logistic Regression, Support Vector Machine, Decision Tree Classifier, and Gaussian Naive Bayes, respectively. The performance could be improved by labeling and retraining more RBL-labeled texts.

Given a relatively small number of RBL Nabua texts, this study managed to establish a text corpus and wordlist that could be a starting point for more natural language processing studies such as sentiment analysis, text classification, morphological analysis, language translation, and others. The researcher recommends further improving the text corpus by collecting more texts and adding more techniques in collecting and analyzing RBL Nabua. This is also a way of preserving the low-resourced local language of Nabua, Camarines Sur, Philippines.

Acknowledgements We thank everyone behind this study's success, especially our family, colleagues, the teachers from Santiago Old Elementary School, our interns, and friends, for the unending support.

Funding No Funding was acquired on the conduct of this study.

Declarations

Compliance with ethical standards This study complied with the rules of good scientific practice stated on the Ethical Responsibilities of Authors.

Conflict of interest On behalf of all authors, the corresponding author states that there is no Conflict of interest.

Ethical approval Data that has been used in this study are offered free and without restriction. Data providers was acknowledged in the acknowledgement part.

Informed consent Informed Consent is not applicable on the conduct of this study.

Competing interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Adlaon, K. M. M., & Marcos, N. (2019). Building the language resource for a cebuano-filipino neural machine translation system. In *Proceedings of the 2019 3rd International Conference on Natural Language Processing and Information Retrieval*. Association for Computing Machinery, New York, NY, USA, NLPIR (2019), pp. 127–132, https://doi.org/10.1145/3342827.3342833

Adolphs, S., & Knight, D. (2010). Building a spoken corpus: what are the basics? The Routledge Handbook of Corpus Linguistics URL https://eprints.ncl.ac.uk

Becker, D., & Riaz, K. (2002). A study in Urdu corpus construction. In *Proceedings of the 3rd workshop on Asian language resources and international standardization - COLING '02*, vol 12. Association for Computational Linguistics, Not Known, pp. 1–5, https://doi.org/10.3115/1118759.1118760, URL http://portal.acm.org/citation.cfm?doid=1118759.1118760

Biber, D., Douglas, B., Biber, P. D., et al. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge University Press, google-Books-ID: 2h5F7TXa6psC.

Boriboon, M., Kriengket, K., Chootrakool, P., et al. (2009). BEST Corpus Development and Analysis. In *2009 International conference on asian language processing*. IEEE, Singapore, pp. 322–327, https://doi.org/10.1109/IALP.2009.76, URL https://ieeexplore.ieee.org/document/5380726/

Dimalen, D. M. D., & Roxas, R. E. O. (2007). Autocor: A query based automatic acquisition of corpora of closely-related languages. In *Proceedings of the Korean society for language and information conference*, Korean Society for Language and Information, pp. 146–154.

Dita, S. N., Roxas, R., & Inventado, P. (2009). Building online corpora of philippine languages. In *PACLIC*.

Dootio, M. A., & Wagan, A. I. (2021). Development of Sindhi text corpus. *Journal of King Saud University - Computer and Information Sciences, 33*(4), 468–475. https://doi.org/10.1016/j.jksuci.2019.02.002

Fernandez, J. L., & Adlaon, K. M. M. (2022). Exploring word alignment towards an efficient sentence aligner for Filipino and Cebuano languages. In *Proceedings of the fifth workshop on technologies for machine translation of low-resource languages (LoResMT 2022)*. Association for Computational Linguistics, Gyeongju, Republic of Korea, (pp. 99–106), URL https://aclanthology.org/2022.loresmt-1.13

Ghayoomi, M., Momtazi, S., & Bijankhan, M. (2010). A study of corpus development for Persian. *International Journal of Asian Language Processing, 20*, 17–34.

Hasko, V. (2012). Qualitative Corpus Analysis. In C.A. Chapelle, (ed.), *The encyclopedia of applied linguistics*. Blackwell Publishing Ltd, Oxford, UK, p. wbeal0974, https://doi.org/10.1002/9781405198431.wbeal0974, URL https://onlinelibrary.wiley.com/doi/10.1002/9781405198431.wbeal0974

Kilgarriff, A., Rundell, M., & Uí Dhonnchadha, E. (2007). Efficient corpus development for lexicography: Building the New Corpus for Ireland. *Language Resources and Evaluation, 40*(2), 127–152. https://doi.org/10.1007/s10579-006-9011-7

Krieger, D. (2003). Corpus Linguistics: What It Is and How It Can Be Applied to Teaching. The Internet TESL Journal IX(3). URL http://iteslj.org/Articles/Krieger-Corpus.html

Kuroashi, S., & Nagao, M. (1998). Building a japanese parsed corpus while improving the parsing system. In *LREC*.

Kübler, S., & Zinsmeister, H. (2015). Corpus linguistics and linguistically annotated corpora. *Bloomsbury Academic*. https://doi.org/10.5040/9781472593573, URL https://doi.org/10.5040/9781472593573

Loper, E., & Bird, S. (2002). NLTK: the Natural Language Toolkit. In *Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics*, vol 1. Association for Computational Linguistics, Philadelphia, Pennsylvania, pp. 63–70, https://doi.org/10.3115/1118108.1118117, URL http://portal.acm.org/citation.cfm?doid=1118108.1118117

Magueresse, A., Carles, V., & Heetderks, E. (2020). Low-resource languages: A review of past work and future challenges. CoRR arXiv:abs/2006.07264. URL https://arxiv.org/abs/2006.07264,

Nafis, N. S. M., & Awang, S. (2021). An enhanced hybrid feature selection technique using term frequency-inverse document frequency and support vector machine-recursive feature elimination for sentiment classification. *IEEE Access, 9*, 52177–52192. https://doi.org/10.1109/access.2021.3069001

Nelson, M. (2010). Building a written corpus. In *The Routledge handbook of corpus linguistics*. Routledge, https://doi.org/10.4324/9780203856949.ch5, URL https://www.taylorfrancis.com/books/9780203856949

Onyenwe, I., Uchechukwu, C., & Hepple, M. (2014). Part-of-speech Tagset and Corpus Development for Igbo, an African Language. In *Proceedings of LAW VIII - The 8th linguistic annotation workshop*. Association for Computational Linguistics and Dublin City University, Dublin, Ireland, pp. 93–98, https://doi.org/10.3115/v1/W14-4914, URL http://aclweb.org/anthology/W14-4914

Rahman, U., & Mutee, (2015). Towards Sindhi Corpus Construction. SSRN Scholarly Paper ID 3820418, Social Science Research Network, Rochester, NY, URL https://papers.ssrn.com/abstract=3820418

Rubi, R. B., & Molina, M. C. C. (2020). Rinconada: The people and its language explored. *Asia Pacific Journal of Education, Arts and Sciences, 7*(4), 37–43.

Santorini, B. (1990). Part-of-speech tagging guidelines for the penn treebank project.

Schramm, M. C., Freitas, L. F. R., & Barone, A. Z. D. (2000). A Brazilian Portuguese Language Corpus Development. In *6th international conference on spoken language processing*. ISCA Archive, Bejing, China, URL https://www.isca-speech.org/archive/archive_papers/icslp_2000/i00_2579.pdf

Schäfer, R., & Bildhauer, F. (2013). Web corpus construction. *Synthesis Lectures on Human Language Technologies, 6*(4), 1–145. https://doi.org/10.2200/S00508ED1V01Y201305HLT022

Tsou, B. K., & Chin, A. C. (2010). A large synchronous corpus as monitoring corpus: Some comparative content analysis of Chinese and japanese language developments. In *4th international universal communication symposium*, IEEE. https://doi.org/10.1109/iucs.2010.5666763

Terms and Conditions

Springer Nature journal content, brought to you courtesy of Springer Nature Customer Service Center GmbH ("Springer Nature").

Springer Nature supports a reasonable amount of sharing of research papers by authors, subscribers and authorised users ("Users"), for small-scale personal, non-commercial use provided that all copyright, trade and service marks and other proprietary notices are maintained. By accessing, sharing, receiving or otherwise using the Springer Nature journal content you agree to these terms of use ("Terms"). For these purposes, Springer Nature considers academic use (by researchers and students) to be non-commercial.

These Terms are supplementary and will apply in addition to any applicable website terms and conditions, a relevant site licence or a personal subscription. These Terms will prevail over any conflict or ambiguity with regards to the relevant terms, a site licence or a personal subscription (to the extent of the conflict or ambiguity only). For Creative Commons-licensed articles, the terms of the Creative Commons license used will apply.

We collect and use personal data to provide access to the Springer Nature journal content. We may also use these personal data internally within ResearchGate and Springer Nature and as agreed share it, in an anonymised way, for purposes of tracking, analysis and reporting. We will not otherwise disclose your personal data outside the ResearchGate or the Springer Nature group of companies unless we have your permission as detailed in the Privacy Policy.

While Users may use the Springer Nature journal content for small scale, personal non-commercial use, it is important to note that Users may not:

1. use such content for the purpose of providing other users with access on a regular or large scale basis or as a means to circumvent access control;
2. use such content where to do so would be considered a criminal or statutory offence in any jurisdiction, or gives rise to civil liability, or is otherwise unlawful;
3. falsely or misleadingly imply or suggest endorsement, approval , sponsorship, or association unless explicitly agreed to by Springer Nature in writing;
4. use bots or other automated methods to access the content or redirect messages
5. override any security feature or exclusionary protocol; or
6. share the content in order to create substitute for Springer Nature products or services or a systematic database of Springer Nature journal content.

In line with the restriction against commercial use, Springer Nature does not permit the creation of a product or service that creates revenue, royalties, rent or income from our content or its inclusion as part of a paid for service or for other commercial gain. Springer Nature journal content cannot be used for inter-library loans and librarians may not upload Springer Nature journal content on a large scale into their, or any other, institutional repository.

These terms of use are reviewed regularly and may be amended at any time. Springer Nature is not obligated to publish any information or content on this website and may remove it or features or functionality at our sole discretion, at any time with or without notice. Springer Nature may revoke this licence to you at any time and remove access to any copies of the Springer Nature journal content which have been saved.

To the fullest extent permitted by law, Springer Nature makes no warranties, representations or guarantees to Users, either express or implied with respect to the Springer nature journal content and all parties disclaim and waive any implied warranties or warranties imposed by law, including merchantability or fitness for any particular purpose.

Please note that these rights do not automatically extend to content, data or other material published by Springer Nature that may be licensed from third parties.

If you would like to use or distribute our Springer Nature journal content to a wider audience or on a regular basis or in any other manner not expressly permitted by these Terms, please contact Springer Nature at

onlineservice@springernature.com