CHUNKER BASED SENTIMENT ANALYSIS FOR NEPALI TEXT

Archit Yajnik and Sabu Lama Tamang

Department of Mathematics, Sikkim Manipal Institute of Technology, Rangpo, Sikkim

ABSTRACT

The article represents the Sentiment Analysis (SA) of a Nepali sentence. Skip-gram model is used for the word to vector encoding. In the first experiment the vector representation of each sentence is generated by using Skip-gram model followed by the Multi-Layer Perceptron (MLP) classification and it is observed that the F1 score of 0.6486 is achieved for positive-negative classification with overall accuracy of 68%. Whereas in the second experiment the verb chunks are extracted using Nepali parser and carried out the similar experiment on the verb chunks. F1 score of 0.6779 is observedfor positive -negative classification with overall accuracy of 85%. Hence, Chunker based sentiment analysis is proven to be better than sentiment analysis using sentences.

KEYWORDS

skip –gram model, MLP classification, parser, verb chunks

1. INTRODUCTION

Sentiment Analysis is a well-known text classification technique in computer science that examines people's ideas and perspectives and categorises them according to their nature into various classifications. [1] However, the terms sentiment analysis and opinion mining are sometimes used synonymously because opinion mining is a tool that contextualises polarity ratings in terms of topics, facets, and objectives. Sentiment detection is the process of categorising a sentence or text into classifications that are neutral, positive, or negative (sometimes) depending on the polarity of the sentences [2]. There are 45 million native speakers of the Nepali language throughout the world, notably in Nepal, Bhutan, Myanmar, and various regions of India, including Sikkim, West Bengal (Darjeeling district), Uttaranchal, and Assam [3]. Despite the popularity, the Nepali language continues to remain understudied. There are yet more languages, like Nepali, that can be taken into consideration for the sentiment analysis assignment. Until now, numerous languages, including English, Chinese, Persian, and Arabic, have been used for the task. Due to the lack of a well-annotated corpus, SA in the Nepali language is not as straightforward as it might first appear. This research work can be extended to tense classification along with another type of review, like social –media comments, election reviews, movie reviews, book reviews, etc. In future, the proposed approach can also be applied to other native Indian languages.

1.1. Contribution

This research aims to improve sentiment analysis by overcoming challenges in handling Nepali text data. With a vast amount of user-generated content in Nepali, Sentiment Analysis can help us

David C. Wyld et al. (Eds): AI&FL, SCM, NLPTT, DSCC -2023
pp. 47-55, 2023. CS & IT - CSCP 2023 DOI: 10.5121/csit.2023.131406

in generating valuable information. This corpus was collected from Kaggle and sentiment analysis. The approach can be extended to various domains as in tense classification, social media comments, election reviews, movie reviews, and book reviews, and can be applied to other native Indian languages in the future.

1.2. Skip- Gram Model

The Skip-gram model is a method for vector representation of words from a huge amount of unstructured text data, which was introduced by Mikolov et al. [4]. This model tries to maximize word classification which is based on another word in the same sentence.

Furthermore, this method predicts the context words using the main word and predict words within a certain range before and after the current word. Suppose a text is composed by a sequence of words wo, w1, w2, w3…... wN. Then for word w, context of w is given by its left and right neighbourhood. Then to each word w a vector representation v is assigned, and the probability that w0 is in the context of wi is defined as the SoftMax of their vector product.

$$p(w_O|w_i) = \frac{\exp\left(v_{w_i} . v_{w_O}^{\top}\right)}{\sum_{w=1}^{V} \exp\left(v_{w_i} . v_w^{\top}\right)} \quad \text{……………………………..…. (1)}$$

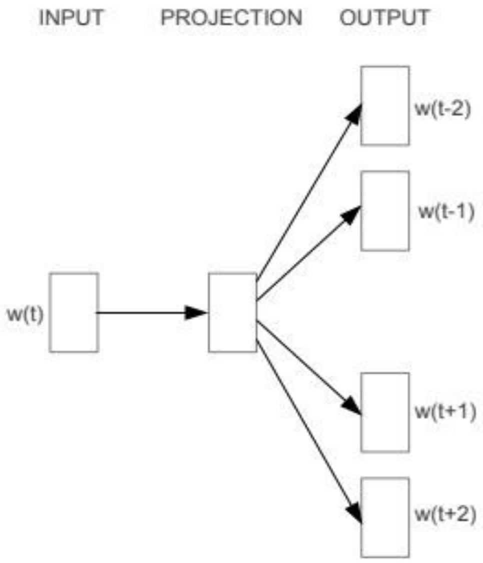

Fig. 1 Skip- gram model

In this work, Skip gram model (see Fig. 1) is employed because it does not involve dense matrix multiplications, resulting in extremely efficient training. An optimized single-machine implementation can train more than 100 billion words in a day [5]

The main objective of skip-gram model is to predict the context of central words. So therefore, training the model means maximizing the objective function. If given a sequence of training words w1, w2, w3,……….,wN.

The **objective function** is given by

$$1/\mathrm{N}\sum_{t=1}^{N}\sum_{-a\leq i\leq a,i\neq 0} logP(w\mathrm{i+t/wi}) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2)$$

Where,
a =size of the training context and wi is the central word.
Wi-a,………,wi-1,wi,wi+1,…………..,wi+a

Higher accuracy can be achieved by using larger values of because they generate more training instances, but doing so also adds to the computational complexity. This is equivalent to minimising the cross-entropy average over the corpus, which appears to be the loss function.

Loss-function,

$$E= -1/\mathrm{N}\sum_{t=1}^{N}\sum_{-a\leq i\leq a,i\neq 0} logP(w\mathrm{i+t/wi}) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots... (3)$$

Cross –Entropy Loss function is also called logarithmic loss, log loss or logistic loss. The predicted class probability is compared to the actual desired output, and a logarithmic penalty is calculated. A large score is given for large differences close to 1, and a small score for small differences tending to 0. [6] The aim is to minimize the loss, i.e., the smaller the loss the better the model. A perfect model has a cross-entropy loss of 0. [6]

1.3. Verb Chunker

Parsing is a crucial step in Natural Language Processing, extracting meaning from text by identifying speech parts, phrases, and clauses. It is essential for machine translation and preserving semantic and syntactic knowledge of a natural language [7]. The task of parsing is firstly initiated by POS tagging [8] and the taggers used reduces the ambiguity of the parser’s input sentence which results in less ambiguous results. [9]. Particularly, Nepali POS tagging [8] has many applications such as it gives information about the word and its neighbouring words which can be further useful for higher level NLP tasks such as semantic analysis, machine translation and so on. (MacKinlay., 2005.) . In the given sentence भाईलाई खेल्न लगाइयो, the verb chunker is खेल्न लगाइयो.

2. STATE OF THE ART TECHNIQUES IN SENTIMENT ANALYSIS FOR NEPALI TEXT

A. Machine learning algorithms for sentiment classification SVM and Naive Bayes are promising sentiment classification algorithms, but face challenges in Nepali language adaptation due to language characteristics and limited datasets.

B. Lexicon-based approaches for sentiment analysis Develop Nepali sentiment lexicons, use machine learning for accurate analysis, explore rule-based methods for sentiment classification, and integrate lexicon-based and machine learning approaches for improved accuracy and efficiency [12].

C. Deep learning methods for sentiment analysis Convolutional Neural Networks revolutionize sentiment analysis by capturing local patterns and dependencies.
Deep learning models have been shown better than conventional machine learning models and lexicon-based approach in sentiment analysis [12].

3. NEPALI DATASET

In this research the dataset available in kaggle [11] for Nepali language Sentiment Analysis has been used to perform the experiment. A total of 4,200 sentences were used during the experiment. Out of which 30% (1260) were test sentences and 70% (2940) were training sentences.

A few significant challenges associated with Nepali languages are discussed below.

a. A slight variation of words in Nepali language text can influence the polarity of the word. For example, Line 1) भाई खेल्दैछ (“brother is playing”) and Line 2) भाई खेल्दैन (“brother does not play”)
The Line 1) refers to positive sentence and Line 2) refers to negative sentence.

b. There is absence of well-annotated corpus which makes the Sentiment Analysis task in the Nepali Language a challenging one.

4. METHODOLOGY

In this section the methodology involved in sentiment detection has been presented in the form of a flow chart (See Fig. 2).

Part A- Sentiment analysis for sentences:

Using the data from Kaggle [11] the sentences were encoded using skip-gram and MLP classifier. The resulting accuracy was noted. The errors arising out if it were recorded and classified into the following 4 types.
Type 1: Negative words are not available but still negative.
Type 2: Positive words are not available but still positive.
Type 3: Negative word is available but still positive.
Type 4: Interrogative sentences whose meaning is very challenging for the machine to understand.

Part B- Sentiment analysis for Verb Chunks:

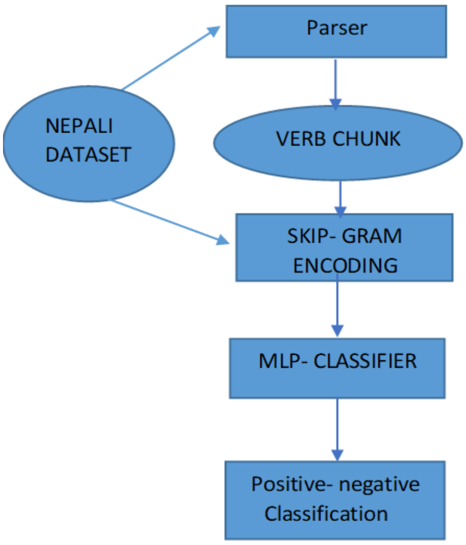

Fig. 2. Flow Diagram of the Methodology

As shown in fig. 2, the words are represented as vectors using the Skip-gram model [4], and the verb chunks and sentences are identified as positive or negative using an MLP classifier.

5. Results and Discussion

Experiment I (For Sentences)

Experiment I resulted in an accuracy of 68%. As mentioned in previous section Part- A, Methodology. However, there were a lot of errors along with it. The errors were analysed and were categorised into four types. Out of which Type 3 was, given more emphasis for the Experiment II.

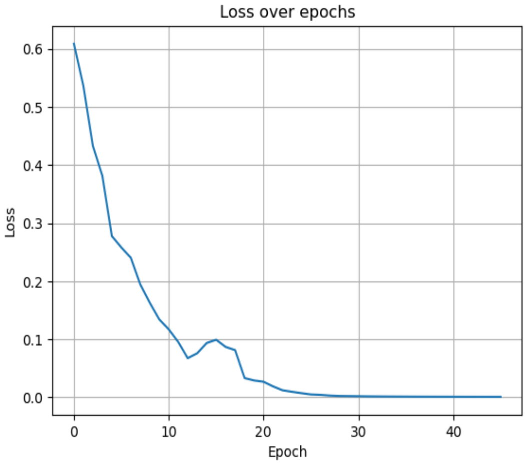

Fig: 3 Loss over epoch,Train Accuracy: 1.0 Test Accuracy: 0.6486

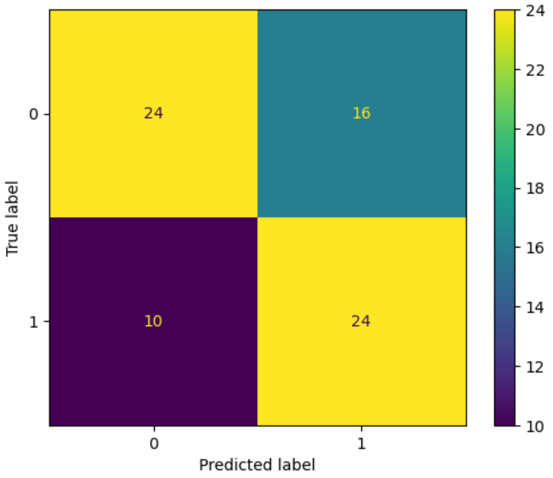

Fig 4. Confusion matrix for sentence F1 score= 0.6486

However, some errors were encountered during the Experiment I, which were broadly classified into four types.

In Type 1 error: भाइ गीत गाईरहेको छ ("Brother is singing song") is detected as negative sentence but negative words are not available.

In Type 2 error: घिन लाग्दो दानब ("Forlorn Demon") is detected as positive sentence but positive words are not available.

In Type 3 error: राम बजार गएन ("Ram did not go to the market") is detected as positive sentence but negative word गएन ("did not go") is available.

Another example of Type 3 error: रामलाई बजार जान लगाइएन ("Ram wasn't made to go to the market") was detected earlier as positive sentence even if negative word (लगाइएन) ("wasn't made to go") is available.

In Type 4 error: the classification of positive and negative sentences depends on the meaning it forms, which is very challenging for the machine to understand. खाना खाने भए यस्तै गर्थे? ("Would have done this while eating the food") is wrongly detected as negative sentence and आखिर देशविकासको बाधक को रहेछ त? ("Afterall who is the obstacle to the country's development") is wrongly detected as positive sentence.

The Experiment I, the percentage of total errors occurred during the testing for sentiment analysis of sentences are Type 1 is 30.76%, Type 2 is 7.69%, Type 3 is 53.84% and Type 4 is 8.69%. Experiment I resulted in an overall accuracy of 68%.

Experiment II (for verbs)

The Experiment II was performed using the verb chunks to improve Type 3 errors encountered in Experiment I, and an improvement of 90% was achieved. Making the final accuracy of the experiment as 85%.

Fig. 5 shows that the loss over epochs is least after 30 epochs. Hence it shows that 48 epochs used in this study were enough to train the data. In our study, Training data accuracy was 98.98%

Fig. 6 shows that the accuracy and correctness of results in 878 verb chunks. The experiment showed that out of 878 verbs tested, True negative (0-0, green box) and True positive (1-1, yellow box) were 384+221 = 605. So, Test accuracy using MLP was 605/878 = 0.6890
The Experiment II was performed using the verb chunks to improve Type3 errors and an improvement of 90% was achieved. Making the final accuracy of the experiment as 85%.

So, Chunker based sentiment analysis using Verb Chunks is proven to be better than sentiment analysis using sentences.

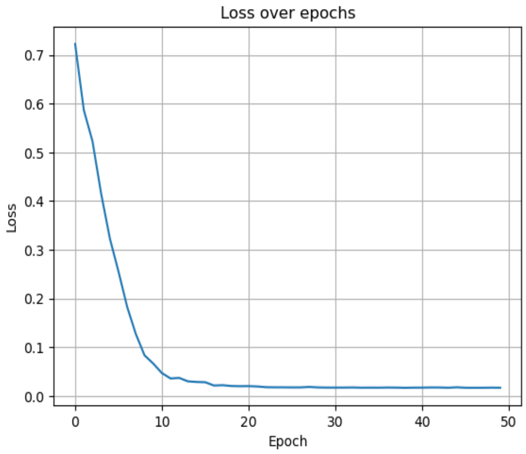

Fig 5. Loss over epoch, Train Accuracy: 0.9898 Test Accuracy using MLP: 0.6890 Test Accuracy using RBF: 0.6697.

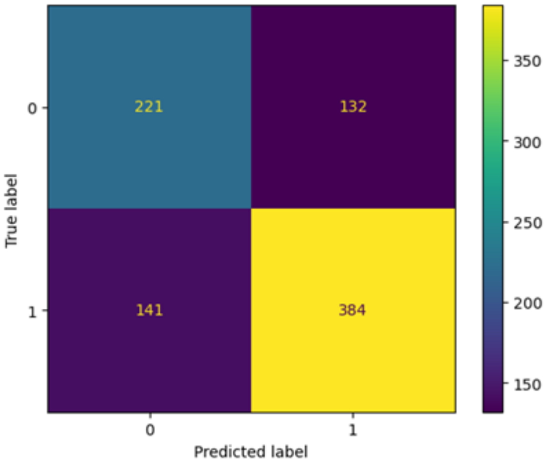

Fig 6. Confusion matrix for verbs F1 score= 0.6779

6. Conclusion

The article represents the sentiment analysis for Nepali Language using Skip-gram model. The identification of the tags of positive-negative is carried out in Nepali dataset available in Kaggle. The Experiment I, carried out for sentiment analysis of sentences, resulted in an overall accuracy of 68% and F1 score of 0.6486 (Fig- 4). The percentage of total errors occurred during the testing for sentiment analysis of sentences are Type 1 is 30.76%, Type 2 is 7.69%, Type 3 is 53.84% and Type 4 is 8.69%.

In Experiment II, Type 3 errors have been improved and an improvement of 90% was achieved. The highest F1 score of 0.6779 (Fig-6) is achieved for the positive –negative classification using skip-gram encoding technique employed on verb chunks with an accuracy of 85%. It has also been found that Multilayer Perceptron (MLP) classifier has performed better than Radial Basis Function (RBF) model for the Skip-gram model.

So, we can conclude that Chunker based sentiment analysis using Verb Chunks, has been proven to be better than sentiment analysis using sentences.

SUMMARY OF THE TWO EXPERIMENTS

	Experiment -I	**Experiment -II**
Principle	Sentences encoded using skip-gram and MLP classifier.	verb chunks encoded using skip-gram and MLP classifier
Overall Accuracy	68%.	85%
MLP Test accuracy	0.6486	0.6890
F1 Score	0.6486	0.6779
Type 3 error	53.84%	Improved 90%

ACKNOWLEDGEMENTS

We would like to acknowledge and express our sincere gratitude to the "Department of Science and Technology, Government of India", for sponsoring the project entitled "Study and develop a natural language parser for Nepali language "reference no. SR/CSRI/- 28/2015(G)" under the "Cognitive Science Research Initiative (CSRI)" to carry out this work. We also acknowledge the "TMA Pai University (Sikkim Manipal University)" research grant for supporting this work.

REFERENCES

[1] K. &. K. D. S. Shrivastava, "A Sentiment Analysis System for the Hindi Language by Integrating Gated Recurrent Unit with Genetic Algorithm.," The International Arab Journal of Information Technology., no. 17. 954-964. 10.34028/I ajit/17/6/14., (2020).

[2] S. Ghosh, "Multitasking of sentiment detection and emotion recoignition in code-mixed Hinglish data.," Knowledge-Based Systems., no. 260.110182.10.1016/j.knosys.2022.110182, (2022).

[3] B. K.Bal., Structure of Nepali Grammar (1st.ed.)., ,Nepal. : Madan PuraskarPustakalaya, 2004.

[4] T. Mikolov, K. Chen, G. Corrado and a. J. Dean., "Efficient estimation of word representations in vector space.," ICLR Workshop, 2013.

[5] T. Mikolov, I. Sutskever, K. Chen, G. Corrado and J. Dean., "Distributed representations of words and phrases and their compositionality.," NIPS ,, 2013.

[6] K. Kiprono Elijah Koech, "https://towardsdatascience.com/cross-entropy-loss-function-f38c4ec8643e," 20 Oct 2020. [Online]. Available: https://towardsdatascience.com/cross-entropy-loss-function-f38c4ec8643e.

[7] A. Pradhan, A. Yajnik and a. Prajapati., "A Conceptual Graph Approach to the Parsing of Projective Sentences .," International Journal of Mathematics and Computer Science,, no. 15(1) 199–221, (2020).

[8] A. Pradhan and A. Yajnik, "Parts-of-speech tagging of Nepali texts with Bidirectional LSTM, Conditional Random Fields and HMM," Multimedia Tools and Applications, 2023.

[9] A. Pradhan and A. Yajnik, "Probabilistic and Neural Network Based POS Tagging of Ambiguous Nepali text:," A comparative Study. ISEEIE, Association for Computing Machinery, Seoul,Republic of Korea., no. https://doi.org/10.1145/3459104.3459146., (2021).

[10] A. MacKinlay., The effects of Part –Of-Speech Tagsets on Tagger Performance (Bachelor's thesis)Master's thesis, .University of Melbourne ,Australia., 2005.

[11] https://www.kaggle.com/datasets/aayamoza/nepali-sentiment-analysis

[12] Piryani, Rajesh & Piryani, Bhawna & Singh, Vivek & Pinto, David. (2020). Sentiment analysis in Nepali: Exploring machine learning and lexicon-based approaches. Journal of Intelligent and Fuzzy Systems. 1-12. 10.3233/JIFS-179884.

AUTHORS

Dr Archit Yajnik, is an Additional Professor in the Department of Mathematics at Sikkim Manipal Institute of Technology. He is a member of DPRC Committee and Coordinator of Workshops and Seminar Committee. His area of interest is Wavelet Analysis and Natural Language Processing and has published several papers in Indexed Journals in NLP.

Ms. Sabu Lama Tamang is a PhD. Research Scholar at Sikkim Manipal Institute of Technology under the guidance of Dr. Archit Yajnik. She has completed her M.Sc. Mathematics from the prestigious NIT Durgapur. Her area of interest is NLP.