Sentiment Analysis of Social Media Texts in Nepali Using Transformers

A Dissertation Proposal Report

Submitted by:

Regan Maharjan

Roll No. 17/075

Submitted to:

Central Department of Computer Science and Information Technology

Tribhuvan University

Kirtipur, Nepal

Under the Supervision of

Bikash Balami

Co-Supervison

Tej Bahadur Shahi

Table of Contents

Table of tables

1. Introduction

Sentiment Analysis is the task of identifying and extracting the polarity or emotion and subjective opinions in natural language texts. It plays a vital role in understanding public opinions and sentiments expressed on social media platforms, e-commerce sites, or any other domain. With the growing reach of the technology and availability of computing machines (such as PC, laptops, and smartphones) as well as the internet even in remote areas, social media usage has become ubiquitous in Nepal. From the young to old, almost everyone has a presence at least in one of the many social media platforms available. The usage of social media and way of expression on these platforms may have evolved from texts to the image to audio-video, however still, the basic form of expressing and interacting remains in-large the use of written form.

The young generation, in the majority, prefers the usage of English as the mode of expression. Even so, the actual number of people that use English language is very minimal, where most prefer to use only English alphabets to converse in and express in Nepali. But nonetheless, there are groups of people that express themselves on internet using Nepali (Devanagari Script). The number of these peoples are growing as there is the availability of typing in Nepali using, may it be Nepali character-labeled keyboards or English-to-Nepali Unicode converter. The growing number can also be attributed to the usage of Nepali by public figures (politicians, actors, and others) and as well as business organizations, as most of working class people do not understand English, thus increasing the reachability. Also, the online news portals must be attributed as they have long since published news in Nepali script.

The world wide web and the Internet have connected people from distant parts of the world. Moreover, it has become a tool that allows us to connect even in the most isolated period of time, just like the recent pandemic. It also can be taken into account that people are more comfortable sharing their emotions on social media rather than with a person. So, as the number of people who prefer Nepali on digital platforms increases, it is apparent that a proper analysis and sentiment classification of these posts/tweets/comments on digital/social media platforms is necessary.

[1]Transformers are a type of neural network architecture that has revolutionized the field of Natural Language Processing (NLP). Transformers are neural network architectures that rely on attention mechanisms [2] to encode and decode sequential data. They can capture long-range

dependencies and learn contextual representations of words and sentences, which had been previously a bottleneck as RNNs couldn't carry along long-range dependencies [1]. They were introduced in [1] and have since become the dominant architecture powering many state-of-the-art models. As the [3] points out, there has been a paradigm shift on the field of NLP by the use of transformer based models (like BERT [4], GPT [5]). In particular, transformer-based models like BERT [4] and its variants have shown superior performance in various natural language processing tasks, including sentiment analysis. In this thesis, we aim to address problem of sentiment analysis of social media texts in Nepali using state-of-the-art deep learning techniques, specifically transformer.

2. Problem Statement

Transformers are a state-of-the-art deep learning architecture that has achieved remarkable results in various NLP tasks, including sentiment analysis. However, most existing sentiment analysis techniques and tools are primarily developed for widely spoken languages and lack support for languages with limited resources, such as Nepali. The works which are done in Nepali sentiment analysis are mostly done using the RNNs, CNNs and traditional machine learning algorithms as can be found in [6] [7] [8] [9] [10]. Some significant work has been done for building a only Nepali pre-trained BERT Language Models, such as NepBERT in [11] and NepBERTa in [12]. Some works have been done for text classification of Nepali news texts using transformers, such as in [13] and [14]. [7] does make use BERT model for sentiment analysis, however, they fine-tuned a multi-lingual BERT model. Apparently, there hasn't been much study regarding the use of transformer-based models for sentiment analysis of Nepali texts.

3. Objective

The main objectives of this thesis are:

1. To compare and evaluate different transformer-based models on the available datasets for Sentiment Analysis in Nepali language (Devanagari Script).
2. To study how well the transformer-based models perform compared to other Neural Network Architectures on the premise of Nepali being a low-resource language.

3. To investigate methods to improve the performance of transformer based models on Nepali sentiment analysis using techniques such as data augmentation, cross-lingual transfer learning, domain adaptation, etc.

4. Literature Review

Sentiment analysis is a text classification problem where the target classes are the sentiments or emotions being conveyed in the given text. These sentiments are categorized as being Positive, Negative, or Neutral. Sentiment analysis is a well-studied problem in NLP and has been applied to various languages and domains. However, most of the existing works are focused on high-resource languages, such as English, and there is a lack of research and resources for low-resource languages, such as Nepali.

4.1. Related Works

Here we discuss works that have been done in the field of sentiment analysis of Nepali texts. We also take into account the work done in text classification. We do this because many works have been done in domain text classification of Nepali texts, not sentiment analysis but news classification. There are many news portals that publish news written in Nepali (Devanagari). Moreover, these news portals already have categorized the news articles. This provides abundant data. In addition to the data abundance, there are works that have been conducted for text classification (news classification) using transformers, some of which we will explore at the end of this section.

4.1.1 Sentiment Analysis

In [5], the authors claim to be the first to perform sentiment analysis on Nepali texts. They proceed on the task with two approaches, first, a resource-based approach, and second, an ML-based approach. They used the Naïve Bayes algorithm for the ML-based approach. In the case of the resource-based approach, they use SentiWordNet, a dictionary translated English-to-Nepali SentiWordNet. While the resource-based approach performed poorly they report 77.8% precision and 70.2% recall on the dataset of 20000 sentences. One thing to note here is, they did classification of two classes, subjective (positive or negative) and objective (neutral).

In [7], sentiment analysis was performed on data collection of YouTube comments. They study abusive sentiment analysis and sentiment analysis over different aspect terms within the sentence. They study four different aspects like profanity, violence, etc. and only use two class, positive and negative, for sentiment. In [7] for sentiment analysis task, they used BiLSTM, CNN, SVM and BERT models. They report 0.81 F1score by BiLSTM and CNN. BERT model performs slightly less with 0.799 F1 score. The maximum accuracy reported for BERT model is 81.5%. This should be noted that they used multi-lingual BERT, which is trained on 104 languages including Nepali, rather than using monolingual, only Nepali, pretrained BERT [4] [11] [12].

In [8], we find sentiment analysis on Nepali tweets regarding covid-19 using CNN model. They provide most extensive dataset in the domain of sentiment analysis of Nepali texts. They collected the data and classified the tweets in three polarities; positive, negative and neutral. The data is called NepCOV19Tweets. They propose three different approach to feature extraction for the representation of the data, namely fastText-based, domain-specific and domain-agnostic. A separate CNN model is trained using one of the feature representation. Then, a fusion layer is used to make combined decision based on the result from the three models. In [8] they make comparison of performance of the CNN model with other machine learning models like SVM, DT, RF, etc. When performance of individual CNN was evaluated, CNN trained using fastText representation performed best with 68.1 accuracy and 58.5 F1 score. Combined, the model achieved 68.7 accuracy and 56.4 F1 score.

[9] and [10] follows the work of sentiment analysis on NepCOV19Tweets dataset. In [9], we see that the authors focus on betterment of text representation and propose hybrid feature TF-IDF weighted fastText based method. They evaluate the performance of traditional ML algorithms for the task of sentiment analysis of covid-19 tweets by using TF-IDF, fastText, and hybrid representation. The performance is then compared with state-of-the-art presented [8]. [9] reports the highest achieved performance by the SVM+RBF model with 75.6 F1-score and 70.69% accuracy. We can see that F1-score and accuracy both increased with the use of hybrid text feature representation.

In [10], Multi-channel CNN is proposed with the hybrid approach for feature extraction from texts using fastText-based and domain-specific methods. To implement multi-channel CNN, four different CNNs with different kernel sizes (1, 2, 3, and 4 respectively). First, each CNN is fine-tuned. Each CNN is aggregated using a fusion layer, thus establishing a multi-channel. Then, the model is trained in an end-to-end fashion for the classification. They report the performance of MCNN with a 61.6 F1-score and 71.3 accuracy.

4.1.2 News Classification

News classification has become the de-facto Nepali text classification problem used for studies and for fair reasons. The availability of ready-to-use data in low-resource languages like Nepali where there is no major incentive for researchers and organizations to collect and accumulate large-scale data, for research purposes, is a blessing. In [15], we see a news classification task is done using SVM and three different feature extraction methods are used on the experiments. They used TF-IDF, word2vec and LSI based approach for feature extraction. LSI based approach achieved highest accuracy, 93.7%, followed by word2vec and TF-IDF with 86.3% and 85.4% accuracy, respectively. However, looking at the overall performance, and evaluating all the performance metrics (accuracy, precision, recall, and f1-score), the model performed better wword2vec-based feature representation than LSI and TF-IDF.

Similarly, in [16], they use SVM along with Naïve Bayes and Multi-layered Perceptron for the task of news classification. They built the dataset with twenty categories. A TF-IDF-based feature extraction method is used for feature vector representation. SVM with RBF kernel is shown to achieve the best performance with above 74% baseline across each performance metric (accuracy, precision, recall, and f1-score) with 74.65 % accuracy which is closely followed by linear SVM with 74.62% accuracy and 72.99% by MLP.

In [17] we see the use of RNN-based models, like LSTM, GRU, and adaptive GRU for Nepali news classification. As other works do, which are referenced so far, [17] also compares the performance of RNN-based models with other traditional ML approaches on the classification task. It makes use of TF-IDF and word2vec-based based feature extraction methods. The GRU model achieved the highest among RNN models with 77.44% accuracy, whereas, a simple perceptron achieved 78.56% accuracy. The author attributes the comparatively lower performance of RNN models (LSTM and GRU) is due to the limited amount of data available for training.

[11] and [12] focuses on training the monolingual BERT language model on a fairly large Nepali corpus. [13] and [14] builds upon the intuitions from [11] and [12] for Nepali news text classification. [11] uses the BERT tokenizer as it is, without taking into account the language

difference between English and Nepali leading to incoherent tokenized words, which can be seen in Table 1. [13] pre-trains DistillBERT and DeBERTa, two BERT-based models, and fine-tunes for the task of text classification. Pre-training is done through Masked Language Modelling (MLM). They then compare the performance of their pre-trained LMs with the pre-trained LMs from [11] and [12]. DeBERTa achieves 88.93% accuracy and DistillBERT achieves 88.31% accuracy on the downstream task. This is a significant performance improvement on the text classification task. We saw in [15] that SVM with LSI features achieves 93.7% accuracy, however, the f1-score is significantly lower than accuracy. From this, we know that the model fails to identify some of the categories. But since [13] doesn't provide those metrics, we can't make comparison of overall performance of two methods on news classification task.

	Before tokenization	After tokenization
Nepali (Devanagari):	फ्लु (फ + ◌् + ल + ◌ु)	फल (फ+ ल)
Translation:	Flu	Fruit

Table 1. Tokenization of Nepali texts by the BERT tokenizer.

[14] goes a step further on the use of transformer-models for news classification. While [13] used DistilBERT, DeBERTa and two separate pre-trained BERT models, [14] uses BERT, RoBERTa, DistilBERT, DeBERTa, mBERT, XLM-RoBERTa, and HindiRoBERTa. Along with those models, [14] also uses Bi-LSTM, MNB, RF, and SVM in their study. In their study, [14] shows that the transformer models performed better as the size of dataset was increased. The DistilBERT and DeBERTa achieved highest accuracy, 87.03% and 86.63% respectively. This finding corroborates with the finding of [13].

4.2 Transformer

[1] [2]. Explanation of transformer here.

4.2.1 Decoder Only Transformers - Generative Pre-trained Transformer (GPT)

[5] [18] Explanation of decoder-only transformers here.

4.2.2 Encoder Only Transformers - Bi-directional Encoder Representation Transformer (BERT)

[4] Explanation of Encoder-only transformers here. A brief explanation of all the BERT-based models, that we encountered in related works, is here.

5. Research Methodology

The research methodology for this thesis will involve the following steps:

I. **Data Collection:**

Research and collect already available representative dataset of Nepali social media text available in open-source data repository such as Kaggle, GitHub, etc.

a. In [8], authors have created a dataset of Nepali tweets regarding covid-19, called NepCov19, for sentiment analysis which is publicly available on Kaggle.

b. In [7], authors have created a dataset of YouTube comments in Nepali for sentiment analysis and aspect term extraction, which is publicly available on GitHub.

c. Data augmentation through transliteration and other approaches.

d. Data may be collected /scraped from Twitter, YouTube and Facebook if possible. We intend to use annotating tools and/or use clustering algorithms and/or Bhavanakos [6] to annotate data, since complete manual annotation of data is time consuming and is not within scope of this study.

II. **Data Preprocessing:**

The collected data will undergo preprocessing steps which include text normalization, tokenization and cleaning. Since the transformer models use tokenizing algorithms like wordpiece, sentencepiece, and byte-pair encodeing (BPE), we assume that traditional preprocessing steps like stemming and lemmatization are not required as they are inherently handled by underlining tokenizer algorithm based on the available corpus and expected vocabulary size. Data cleaning will be done to remove any foreign language character, including emoji's, and will be kept where assumed it will help in the task.

III. **Model Architecture:**

Initially, we will deploy encoder only transformer models, such as BERT or its variants, for sentiment analysis. These models have shown significant improvements in capturing contextual information and achieving state-of-the-art results in various NLP tasks. After that, we will use decoder only transformer models, such as GPT, and compare the efficiency of both models.

We also intend to test out sentiment analysis as a summarization or question answering problem. Bluntly put, as a sequence to sequence problem as shown below.

INPUT SEQUENCE: "कोभिड विरुद्धको खोपको अनुभव"

- Translation: Experience of covid vaccine

OUTPUT SEQUENCE: "यसले तटस्थ भावना व्यक्त गर्दछ"

- Translation: This conveys a neutral sentiment

Where { INPUT SEQUENCE } can be considered as a question to which the answer { OUTPUT SEQUENCE } can have one of three possible answers, which is related to the sentiment as shown above. So, we input a sequence (tweet or post or comment) as an input and we get an output sequence which is sentiment.

IV. Pre-training, Fine-tuning, and Training:

We will adapt and fine-tune pre-trained transformer models for sentiment analysis of Nepali social media texts. We will use existing Nepali language models (e.g. LM made available by [11]) and/or train our own language models on Nepali corpora to initialize the transformer models. We will also explore various techniques such as data augmentation, domain adaptation, cross-lingual transfer learning, etc. to improve the performance of the transformer models.

For model initialization and training, libraries such as Huggingface, Tensorflow, Keras, and Trax will be utilized.

V. Feature Extraction and Hybrid Approaches:

The representation learned by transformer models are self-sufficient. The learnable weights embedding and positional encoding used by transformers already robust. In addition to embedding, attention mechanisms used by transformers, both self-attention and cross-attention, are capable of capturing the long-range dependencies and learn contextual representations of words and sentences. Such that, the representations learned by transformer model like BERT are self-sufficient and doesn't need any other features extracted through traditional methods.

However, we will investigate hybrid feature extraction methods by combining transformer-based representations with other feature extraction techniques like Bhavanakos and

Sabdakos, a word2vec and doc2vec method respectively, as created by [6] and using domain specific and domain agnostic features as used by [8] [9] [10]. The goal is to boost the model's performance and efficiency by incorporating additional linguistic information specific to Nepali text.

VI. Evaluation and Performance Metrics:

Fine-tuned sentiment analysis models will be evaluated on the validation set separated from collected Nepali social media text dataset [7] [8]. The performance evaluation of sentiment analysis models will be done using standard evaluation metrics like accuracy, precision, recall, F1-score, and ROC_AUC score, etc.

We will compare the performance of our proposed transformer-based model with the existing sentiment analysis approaches for Nepali, including the referenced papers. This analysis will help assess the effectiveness and potential advantages of our approach.

In order to evaluate, calculate performance metrics, and visualize data, various libraries such as Scikit-Learn, Matplotlib, Seaborn, and Tensorflow will be utilized.

6. Time Schedule:

Some gantt chart here.

7. References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, N. A. Gomez, L. Kaiser and I. Polosukhin, "Attention Is All You Need," in *31st Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA, USA, 2017.

[2] D. Bahdanau, K. Cho and Y. Bengio, "NEURAL MACHINE TRANSLATION BY JOINTLY LEARNING TO ALIGN AND TRANSLATE," in *ICLR* , 2015.

[3] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. v. Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, E. Brynjolfsson, S. Buch, D. Card, R. Castellon, N. Chatterji and others.., "On the Opportunities and Risks of Foundation Models," 2022.

[4] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," Google AI Language, 2019.

[5] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei and I. Sutskever, "Language Models are Unsupervised Multitask Learners," OpenAI.

[6] C. P. Gupta and B. K. Bal, "Detecting Sentiment in Nepali Texts: A Bootstrap Approach for Sentiment Analysis of texts in the Nepali Language".

[7] O. M. Singh, S. Timalsina, B. K. Bal and A. Joshi, "Aspect Based Abusive Sentiment Detection in Nepali Social Media Texts," in *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 2020.

[8] C. Sitaula, A. Basnet, A. Mainali and T. B. Shahi, "Deep Learning-Based Methods for Sentiment Analysis on Nepali COVID-19-Related Tweets," *Computational Intelligence and Neuroscience,* 2021.

[9] T. B. Shahi, C. Sitaula and N. Paudel, "A Hybrid Feature Extraction Method for Nepali COVID-19-Related Tweets Classification".

[10] C. Sitaula and T. B. Shahi, "Multi-channel CNN to classify Nepali Covid-19 related tweets," 2022.

[11] S. Pudasaini, A. Tamang, S. Lamichhane, S. Adhikari, S. Adhikari, S. Thapa and J. Karki, "Pre-training of Masked Language Model in Nepali," in *36th Conference on Neural Information Processing Systems*, 2022.

[12] M. Gautam, S. Timalsina and B. Bhattarai, "NepBERTa: Nepali Language Model Trained in a Large Corpus," in *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the*, 2022.

[13] U. Maskey, M. Bhatta, S. R. Bhatta, S. Dhungel and B. K. Bal, "Nepali Encoder Transformers: An Analysis of Auto Encoding Transformer Language Models for Nepali Text Classification," in *Proceedings of SIGUL2022 @LREC2022*, 2022.

[14] S. R. T. a. C. Silpasuwanchai, "Comparative Evaluation of Transformer-Based Nepali Language Models," [Online]. Available: https://assets.researchsquare.com/files/rs-2289743/v1/aa3f3ba4a38a880db3d6c5dc.pdf?c=1670229384. [Accessed 19 06 2023].

[15] K. Kafle, D. Sharma, A. Subedi and A. K. Timalsina, "Improving Nepali document classification by neural network," in *Proceedings of IOE Graduate Conference*, 2016.

[16] T. B. Shahi and A. K. Pant, "Nepali News Classification using Naive Bayes, Support Vector Machines and Neural Networks," in *International Conference on Communication, Information & Computing Technology (ICCICT)*, Mumbai, 2018.

[17] O. M. Singh, "Nepali Multi-Class Text Classification," 2018. [Online]. Available: https://oya163.github.io/assets/resume/Nepali_Text_Classification.pdf. [Accessed 19 6 2023].

[18] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh and Dani, "Language Models are Unsupervised Multitask Learners," Open AI, 2020.