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

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 sentiment expressed on social media platforms, e-commerce sites or any other domain. With growing reach to the technology and availability of computing machines (as PC, laptops and smart phones) as well as the internet even in the remote areas, the usage of social media has become ubiquitous in Nepal. From the young to old, almost everyone has a presence at least in one of many social media platforms available. The usage of social media and way of expression on these platforms may have evolved from texts to 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 majority, prefers the usage of English as the mode expression. Even so, the actual number of people that use English language is very minimum, where most preferring to use only English alphabets to converse in and to 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 are 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 organization, as most of the working class people do not understand English, thus increasing the reachability. Also, the online news portals have to be attributed as they have long-since publishing news on Nepali script.

The world wide web and the internet have connected people from distant part of the world. Moreover, it has become a tool that allows us to connect even on 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.

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. 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 self-attention mechanisms to encode and decode sequential data. They can capture long-range dependencies and learn contextual representation so 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 [2] points out, there has been a paradigm shift on the field of NLP by the use of transformer based models (like BERT [3], GPT [4]). In particular, transformer-based models like BERT [3] and its variants have shown superior performance in various natural language processing tasks, including sentiment analysis.

1. 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 [5] [6] [7] [8] [9]. Some significant work has been done for building a only Nepali pre-trained BERT Language Models, such as NepBERT in [10] and NepBERTa in [11]. [6] 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 use of transformer based models for sentiment analysis or classification of Nepali texts.

2. Objective

The main objective 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 does the transformer based models perform in comparison 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.

3. 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.

4. 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 like Kaggle, GitHub, etc.

a. In [7], authors have created a dataset of Nepali tweets regarding covid-19, called NepCov19, for sentiment analysis which is publicly available in Kaggle.
b. In [6], authors have created a dataset of YouTube comments in Nepali for sentiment analysis and aspect term extraction, which is publicly available in GitHub.
c. Data augmentation through transliteration.
d. Data will be collected /scraped from Twitter, YouTube and Facebook if possible. We intend to use annotating tools and use clustering algorithms to annotate data, since manual annotation of data is time consuming.

Since manual annotation of data is time consuming, we intend to use annotating tools and use clustering algorithms to annotate data which will be collected / scraped from Twitter, YouTube and Facebook if possible.

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, unigram or 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, will employ 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 employ 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 a 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 (like made available by [10]) 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.

Libraries like Huggingface, Tensorflow, Trax, etc. will be used for model initialization and training.

V. Feature Extraction and Hybrid Approaches:

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 [5] and using domain specific and domain agnostic features as used by [7] [8] [9]. 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 [6] [7]. 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.

5. References

[1] "Attention Is All You Need".

[2] "On the Opportunities and Risks of Foundation Models".

[3] "BERT: Pre-training of Deep Bidirectional Transformers for".

[4] "Language Models are Unsupervised Multitask Learners (GPT)".

[5] "Detecting Sentiment in Nepali Texts: A Bootstrap Approach for Sentiment Analysis of texts in the Nepali Language".

[6] 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.

[7] 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.

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

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

[10] 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.

[11] 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.

[12] "Improving Nepali document classification by neural network".