## **Abstract**

With increase in internet usage, there are vast amounts of opinionated comments left by the Nepali users on Youtube and other social media platforms in their native language. Sentiment analysis of such comments can be implemented so that it can be beneficial for both research and business purposes. This paper presents a way to create a dataset for sentiment analysis, find sentence embeddings from pretrained language model BERT and implement various machine learning models to evaluate its performance. Random Forest classifier has been used for the classification of sentiment analysis with hyperparameter tuning and achieved 89% F1-score respectively.

Index Terms: Sentiment Analysis, Dataset, BERT, Random Forest Classifier, F1-Score, Hyperparameter Tuning

# **Introduction**

Sentiment Analysis is a classification technique that determines the sentiment of a given sentence. It falls under natural language processing which identifies and extracts the subjective information from the given textual documents and determines the contextual polarity of given sentences. With the recent increase of blogs, social media networks and video streaming websites; a lot of reviews, comments and ratings are expressed on these platforms. This rise of opinionated textual data on above platforms can be used to identify the sentiment of users on various topics and make a decision based on users sentiment. It can be used to make decisions making tasks like purchasing products, watching movies or even use services provided by organisations based on their reviews.

The penetration of smartphones to Nepalese users have certainly grown over the decades allowing users to interact with the social media platforms and leave their comments, opinions in Nepali language. Nepali language is written in Devnagari script and has its own grammatical structure, vocabulary and semantics. In this research project, we propose a sentiment analysis model trained for Nepali dataset.

In this paper, we discuss sentiment analysis on Nepali language datasets which were scraped from social media platforms and youtube comments. Furthermore we explore machine learning algorithms implemented to find sentiment of the given comments, data preprocessing techniques and data collection method for this project. For this paper we have used the BertForMaskedLM model trained on Nepali datasets to generate tokenizers and word embeddings.

In terms of accomplishing the work, our contributions to this papers are as follows:

* Manually collect a dataset; scraping from various platforms and labeling them.
* Develop a data scraping module for nepali language.
* Develop a Machine Learning based Sentiment Classifier based on collected labeled datasets.

## **Related Work**

There are few research works on Sentiment Analysis based on Nepali texts. (Birat Bade Shrestha, Bal Krishna Bal) have implemented sentiment analysis on Nepali news media texts. Their architecture consisted of Data Collection, Preprocessing, Named Entity Recognition, Anaphora Resolution and Sentiment Analysis. For the data collection step, news articles were scraped from four online news portals. Collected datasets were cleaned via article cleaning subcomponents and further lemmatized to split word into root forms and to remove unwanted suffixes under Preprocessing step. POS tagging, NER tagging and Lappin and Leass algorithm for anaphora resolution were performed on the collected datasets to increase the overall accuracy of the system. Word2Vec and FastText were used for feature extraction using Skip Gram parameters. Due to low datasets the sentiment analysis were trained on Support Vector Machine, Decision Tree and Random Forest where Support Vector Machine had better overall score. SVM had F1-score of 80.2 and 78.7 on Word2Vec and FastText respectively. Total of 3490 sentences were scraped for this research method out of which 2676 datasets were positive and 814 were negative.

(Oyesh) performed aspect based sentiment analysis on a dataset scraped from youtube comments. They used the BRAT annotation tool to annotate comment datasets at sentence level where Nouns phrases were assumed as an aspect term. Annotation was performed according to the annotation guidelines. Two different annotators annotated the sentence's polarity and their inter annotator agreement was found to be 0.703. Upon analysis of the corpus, the collected dataset had diverse use of language in social media platforms. The ratio of use of English words to Nepali words was around 0.0125 and found that Negative sentiment was directed towards politicians or persons but not toward any organizations. They have used BiLSTM+CRF model for aspect term extraction task and BiLSTM+CNN for sentiment polarity classification. Total of 3068 comments were collected from various youtube channels videos of which train, test and validation data were splitted into 80%, 10% and 10% respectively. They trained the dataset on 300-dimension skip-gram fasttext word embeddings using gensim to generate word embeddings. The overall F1-score of aspect term extraction and sentiment classification was 57.98 and 81.6 respectively.

(Ashok Kumar Pant) implemented sentiment analysis on Nepali Movie reviews. They collected 500 nepali movie review dataset from online sources. Preprocessing of data is performed to clean and extract features from the comments. TF-IDF vectorizer is used to extract word embeddings