Proceedings of 14th IOE Graduate Conference
Peer Reviewed
Year: 2023 Month: December Volume: 14
ISSN: 2350-8914 (Online), 2350-8906 (Print)

Transformer-Based Deep Learning Models for Sentiment Analysis in Romanized Nepali: A Comparative Investigation of BERT and RoBERTa

Abhash Pradhananga [a], Anand Kumar Sah [b]

[a, b] Department of Electronics and Computer Engineering, Pulchowk Campus, IOE, Tribhuwan University, Nepal
✉ [a] 078mscsk002.abhash@pcampus.edu.np, [b] anand.sah@pcampus.edu.np

Abstract

The complexity of sentiment analysis on Romanized Nepali text, a language that combines the Nepali and English/Latin alphabets that are used on Nepali e-commerce platforms, is explored in this study. Traditional sentiment analysis methods are inadequate for dealing with this complex and dynamic language. Our dataset, which includes 32,817 customer reviews from a well-known Nepali e-commerce site, showed a spectrum of emotions, with 22,880 positive, 8,382 negative, and the remaining neutral reviews. We hired three seasoned annotators with master's degrees who are fluent in Romanized Nepali to assure accuracy and reliability, yielding an outstanding Inter-Annotator Agreement (IAA) of 81.32%. The dataset underwent thorough text pre-processing, which included the creation of a stop word list specifically for Romanized Nepali, WordNet lemmatization, and Soundex stemming. We assessed the accuracy, precision, recall, and F1 score metrics of four classification models: Logistic Regression, Naive Bayes, Support Vector Machine, and BERT. BERT came out on top with an outstanding 79% accuracy, followed by Logistic Regression with 74%, Naive Bayes with 73%, and Support Vector Machine with 75%. These results highlight BERT's capability to analyze sentiment in Romanized Nepali text successfully, with positive ramifications for a range of applications in this linguistic setting.

Keywords

sentiment analysis, romanized nepali, nepali e-commerce, text pre-processing, stop words, wordnet lemmatization, soundex stemming, inter-annotator agreement (iaa), customer reviews, classification models, logistic regression, bert

1. Introduction

The proliferation of online platforms and the rapid growth of Nepal's e-commerce market have changed how people express their ideas and experiences about goods and services. Product reviews are now widely shared online thanks to the rise of user-generated content. Businesses should prioritize understanding the attitude expressed in these reviews since doing so helps them understand client preferences, make wise decisions, and improve their products and services. However, there are particular difficulties in performing sentiment analysis on Romanized Nepali product reviews.

Businesses typically use sentiment analysis[1] to track data from social media, analyse sentiment, assess brand reputation, and understand their customers. Deep learning advances in recent years have improved algorithms' capacity for text analysis. Deep exploration can benefit from the creative application of cutting-edge machine learning algorithms. Currently, emotion identification[2], aspect-based sentiment analysis[3], multilingual sentiment analysis[4], and fine-grained sentiment analysis[5] are some of the most popular types of sentiment analysis.

1.1 Sentiment Analysis

Sentiment analysis, a sub-field of natural language processing, has gained significant attention due to its ability to extract and analyze subjective information from text data. With the exponential growth of user-generated content on the internet, sentiment analysis has become crucial in understanding public opinion, customer feedback, and market trends. However, sentiment analysis in languages with unique characteristics, such as Romanized Nepali, poses distinct challenges. Romanized Nepali, which employs the Roman alphabet to represent the Nepali language, exhibits linguistic nuances that require specialized approaches for accurate sentiment analysis. This thesis focuses on investigating the effectiveness of transformer-based deep learning models, specifically BERT[6] and RoBERTa[7], for sentiment analysis in Romanized Nepali text.

The rise of e-commerce platforms and the increasing reliance on online reviews have amplified the need for accurate sentiment analysis in the context of product reviews. Understanding customer sentiment and preferences is crucial for businesses to make informed decisions, improve their products or services, and enhance customer satisfaction. Previous research on sentiment analysis in Romanized Nepali has primarily employed traditional machine learning algorithms, such as Support Vector Machine (SVM)[8], Logistic Regression (LR)[9], and Naive Bayes(NB)[10]. While these methods have achieved promising results, the advent of transformer-based models, such as BERT and RoBERTa, presents an opportunity to enhance sentiment analysis performance by capturing complex language patterns and contextual information.

1.1.1 Machine Learning Based Models

Support Vector Machine (SVM): Traditional machine learning models, such as Support Vector Machine, are utilised for categorization tasks like sentiment analysis. SVM operates by identifying the appropriate hyperplane for separating data points from various classes. Based on numerous text variables retrieved from the document, such as word frequencies or TF-IDF values, SVM learns to categorise text documents, such as product reviews, into positive, negative, or neutral sentiment categories. SVM is renowned for its proficiency with high-dimensional data and effectiveness in cases that can be separated linearly.

Logistic Regression(LR): Another conventional machine learning approach used frequently for binary classification applications like sentiment analysis is logistic regression. Logistic regression, despite its name, is utilised for classification as opposed to regression. The logistic function is used to model the likelihood that a text would fall into a specific sentiment class (such as positive or negative). When there is a fairly linear relationship between the features and the goal sentiment, logistic regression is relatively straightforward, understandable, and effective.

Naive Bayes(NB): Based on Bayes' theorem, Naive Bayes is a probabilistic machine learning algorithm. It is frequently employed in text categorization projects, such as sentiment analysis. Given the class label, naive Bayes implies that features (words or other textual tokens) are conditionally independent. Naive Bayes can be remarkably successful for text classification tasks despite this ”naive” assumption. It determines the likelihood that a document belongs to a particular sentiment class and then chooses the class with the highest likelihood.

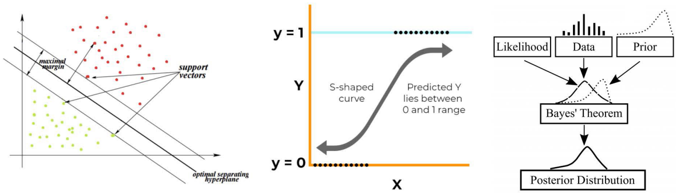


Figure 1: SVM, LR and NB Model

1.1.2 Transformer-Based Deep Learning Models

BERT (Bidirectional Encoder Representations from Transformers): BERT, a modern deep learning model, excels in understanding complex linguistic patterns and contextual links between words due to its pre-training on extensive text data. Customized for sentiment analysis, BERT performs exceptionally well, especially in tasks requiring deep contextual comprehension, such as sentiment analysis. Additionally, DistilBERT, a computationally efficient variant, retains much of BERT's performance while being smaller and faster, making it suitable for resource-constrained applications without sacrificing its ability to handle intricate language patterns and context.

RoBERTa (Robustly Optimized BERT approach): BERT's extension, RoBERTa, improves the pre-training procedure even further. It is intended to capture even more complex language links and patterns in text. Similar to BERT, RoBERTa is an effective deep learning model that works well for sentiment analysis applications. RoBERTa's capacity to comprehend the context allows it to deliver incredibly precise sentiment classifications. It has displayed outstanding performance on a variety of NLP tasks.

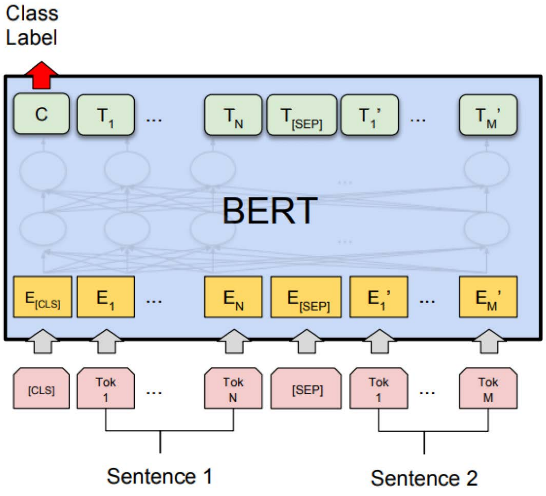


Figure 2: Bert Model Architecture

2. Literature Review

Nepali, spoken by millions of people worldwide, holds significant importance as the official language of Nepal. However, it is considered a Scarce Resource Language (SRL) with limited text processing capabilities and basic dictionaries. Conducting sentiment analysis in Nepali presents various challenges due to the lack of language-specific resources and tools. While some research has been carried out, most sentiment analysis studies have focused on languages like English and Chinese, which benefit from well-established resources and tools. Sentiment analysis, a field of natural language processing, focuses on extracting subjective information and sentiment polarities from text data. It plays a crucial role in understanding public opinion, market research, brand reputation monitoring, and customer experience analysis. With the rise of online platforms, sentiment analysis has become increasingly important, especially in analyzing unstructured data such as user reviews and social media posts. Traditional approaches have predominantly focused on document-level sentiment analysis, providing an overall sentiment assessment of a text. However, recent research has shown the significance of aspect-category sentiment analysis, which aims to predict sentiment polarities for different aspect categories within the same text, providing deeper insights.

Several transformer-based deep learning models have been proposed to enhance sentiment analysis performance. The studies by Narayanaswamy explore the application of RoBERTa (Robustly Optimized BERT Pre-training Approach) and BERT (Bidirectional Encoder Representations from Transformers) models for aspect-category sentiment analysis. Their model employs deep bidirectional Transformers to extract features from both the text and aspect tokens, utilizing the cross-attention mechanism to focus on the most relevant features for each aspect category. Experimental results demonstrate the superior performance of their proposed model in aspect-category sentiment analysis compared to other models.

Narayanaswamy [11] focuses on aspect-level sentiment analysis using BERT and RoBERTa models. Their research emphasizes the importance of aspect identification, which involves extracting attributes (aspects) from text data that people are commenting on. By leveraging the contextual embedding vector space offered by BERT and RoBERTa, they propose a framework for extracting aspects and apply it to these pre-trained models. Experimental results show that the aspect-based approach significantly improves the performance of sentiment analysis models, with the BERT model achieving the highest accuracy among all evaluated models.

Additionally, it's important to take note of a study that compares lexicon-based and BERT-based sentiment analysis in Italian (Catelli, Pelosi, and Esposito, citing Catelli2022) [12]. The usefulness of these technologies is examined in this study utilising a dataset specifically designed for the Italian language, which was carried out in one of the major e-commerce markets in the world. This study sheds light on how BERT and related language models perform in comparison to lexicon-based methods, especially when working with smaller datasets, which is typical for languages other than English or Chinese. BERT and related language models have demonstrated superiority in sentiment analysis.

To further investigate the performance of transformer-based models, Joshy and Sundar [13] analyze sentiment analysis using BERT, DistilBERT, and RoBERTa models. By applying these models to movie reviews and tweets datasets, they compare the performance using accuracy as the evaluation metric. Their findings indicate that the BERT model outperforms the other models in terms of sentiment analysis accuracy.

Considering the specific context of sentiment analysis in Romanized Nepali, there is a gap in the literature regarding the application of transformer-based deep learning models in this language. Therefore, this thesis aims to address this gap by conducting a comparative investigation of BERT and RoBERTa models for sentiment analysis in Romanized Nepali. By leveraging the power of these models in capturing contextual information and sentiment polarities, this study intends to enhance the accuracy and effectiveness of sentiment analysis in Romanized Nepali, contributing to a better understanding of public sentiment in this language.

3. Methodology

The methodology, depicted in Figure 3, involved several steps to extract meaningful features from text and classify sentiment. Initially, a diverse dataset from a top Nepali e-commerce platform captured various sentiments in Romanized Nepali. Data preprocessing includes a human-coded dictionary-based stemming approach and common techniques like stop word removal and lemmatization. TF-IDF analysis can be employed to identify significant features. For model selection, BERT and RoBERTa, known for their NLP prowess, can be chosen. Fine-tuning adapted these models to Romanized Nepali, considering linguistic nuances. The dataset can be split into training, validation, and test sets for experimentation, with evaluation metrics such as accuracy and F1-score. The iterative process can be used to identified the most effective transformer-based deep learning model for Romanized Nepali sentiment analysis.

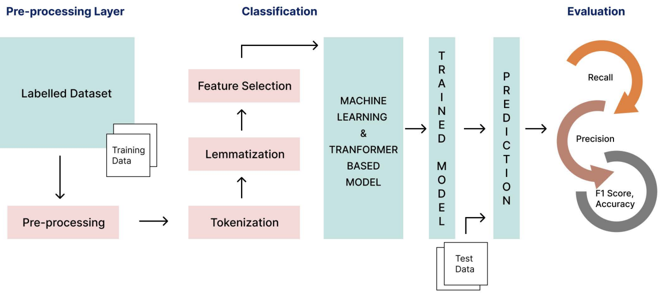


Figure 3: Abstract flow diagram of proposed framework

3.1 Data Collection

The dataset for this study is obtained from a well-known ecommerce site in Nepal that hosts a substantial collection of user-generated product reviews. A collaboration agreement with the platform allows access to this valuable dataset. The goal of the corpus selection procedure is to include reviews from a crosssection of product categories. Initially, various product domains, including electronics, fashion, home appliances, and more, are used to retrieve reviews. Only reviews in Romanized Nepali text are included, and any irrelevant or incomplete reviews are filtered out to maintain the integrity of the data. The resulting dataset includes a substantial number of reviews—more than 30,000 items in total—that are distributed somewhat equally among the product categories.

3.2 Annotation Guidelines

In this section, we describe the manual annotation process for corpus generation. This step involves preparing guidelines for annotation and manually annotating the entire dataset. The annotation rules are developed based on a literature review of existing works in the field of sentiment analysis. An example of user reviews classified as positive and negative can be seen in Table 1 below. The following criteria are used to annotate the sentences:

- A sentence was labeled as positive if it conveyed an overall positive sentiment or if it expressed both positive and neutral sentiments or if it contained expressions of agreement or approval.
- Sentences with words such as congratulations and admiration were also marked as positive.
- A sentence was labeled as negative if it conveyed an overall negative sentiment or if it contained more negative words than other sentiments.
- Sentences that showed any form of disagreement were also classified as negative.
- If a sentence contained terms such as ban, penalize, or assess, it was labeled negative.
- If a sentence included a negative word with a positive adjective, it was also classified as negative.

- A sentence was labeled as neutral if it did not convey an overall positive or negative sentiment and did not contain strong expressions of agreement, disagreement, approval, or disapproval.
- Sentences that contained factual information without emotional connotations were classified as neutral.
- When a sentence contained mixed sentiments, and the overall emotional tone was neither strongly positive nor negative, it was labeled as neutral.

Table 1: Dataset Sample

Positive Review	**Negative Review**	**Neutral Review**
dami lago value for money	Damaged in 5 minutes after installing it	Mild current fragrance
Thank you Daraz This watch is really awesome	Duplicate shampoo please kasaile nakinnu	Satisfactory product

3.2.1 Annotation Process

In this study, a meticulous annotation process is designed and executed to ensure the accuracy and reliability of sentiment annotations. The following section outlines the steps involved in the annotation process, as well as the methodology for calculating the Inter-Annotator Agreement (IAA)[14]. Comprehensive guidelines are developed to provide clear instructions for categorizing user reviews into negative, neutral, or positive sentiment classes, ensuring consistency. Three annotators with master's degrees and native proficiency in Romanized Nepali, along with extensive sentiment analysis experience, are engaged to perform annotations. They independently assess a random subset of 100 user reviews without collaboration, evaluating the initial agreement level. In cases of disparities between annotators X and Y, a third annotator, Z, resolves differences, following the guidelines to maintain consistency and address ambiguities in annotation.

The Inter-Annotator Agreement (IAA) is calculated to quantify the level of agreement between the annotations made by Annotator X and Annotator Y. The Cohens Kappa method, a widely recognized measure of inter-annotator agreement, is employed for this purpose. The formula for Cohens Kappa is as follows:

$$\kappa = \frac{P_o - P_e}{1 - P_e} \tag{1}$$

Where:

- P_o represents the observed agreement between annotators.
- P_e represents the expected agreement by chance.

3.3 Text Preprocessing

Text processing techniques play a crucial role in preprocessing the Romanized Nepali product reviews before conducting sentiment analysis.To improve the accuracy of the model, various elements are removed, including URLs, email addresses, phone numbers, numerical numbers, numerical digits, currency symbols, and punctuation marks. Also, specific stop words that are commonly used in the Romanized Nepali language are identified and eliminated. This step aims to reduce noise and focus on the most relevant content for sentiment analysis. Additionally, WordNet lemmatization is applied to reduce words to their base form, facilitating consistency in sentiment analysis even when dealing with different word variations. By employing WordNet lemmatization, variations of words are standardized, leading to a more robust and reliable sentiment analysis process.

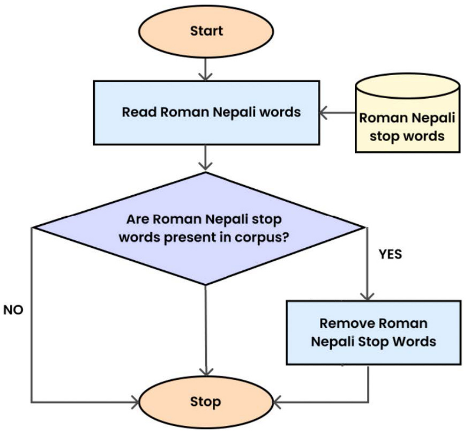

Start
Read Roman Nepali words
Roman Nepali stop words
Are Roman Nepali stop words present in corpus?
YES
NO
Remove Roman Nepali Stop Words
Stop

Figure 4: Flowchart of the Roman Nepali words removal module

3.3.1 Soundex Algorithm

The Soundex algorithm is a phonetic algorithm that was developed in the early 1900s to index surnames based on their pronunciation. The algorithm assigns a code to a word based on the way it sounds, rather than the way it is spelled. This code consists of a letter followed by three digits. The Soundex algorithm is utilized to address variations in pronunciation that may occur in the Romanized Nepali text. This algorithm aids in normalizing the data and enhancing the accuracy of sentiment analysis by accounting for phonetic similarities between words. This comprehensive text preprocessing ensures data cleanliness and enhances the effectiveness of the subsequent sentiment analysis of Romanized Nepali product reviews.

3.4 Feature Selection

Text vectorization is a crucial step in the process of preparing textual data for analysis. It involves converting text data into a numerical format that can be used for machine learning and statistical analysis. In this study, two methods, TfidfVectorizer and SelectKBest, were employed to perform text vectorization. TfidfVectorizer stands for Term Frequency-Inverse Document Frequency Vectorizer. It calculates the TF-IDF score for each word in the text data, which represents how important a word is within a document relative to the entire dataset. This method creates a matrix where each row corresponds to a document, and each column corresponds to a word, with the cell values indicating the TF-IDF score of each word in each document.

$$\text{Term-Frequency}\ (t) = \frac{Nt}{N}$$

An indicator of a term's importance to a document is the inversedocument frequency because phrase-frequency gives each term a nearly equal weight. Most often recurring keywords (stop words) have a strong likelihood of gaining greater weights. We use the Inverse Document Frequency (IDF) as follows to scale up rare terms (which reflect the genuine contribution) while weighing down frequently occurring terms:

$$\text{IDF}\ (t) = \log\left(\frac{Nd}{Nt}\right)$$

3.5 Model Selection and Fine Tuning

BERT and RoBERTa are selected as the transformer-based deep learning models for sentiment analysis. These models have demonstrated their effectiveness in various natural language processing tasks, including sentiment analysis, and are widely used in the field. Pre-trained versions of BERT and RoBERTa, trained on large-scale datasets such as multilingual or English corpora, are chosen to leverage their contextual representation capabilities. The selected BERT and RoBERTa models are fine-tuned on the Romanized Nepali sentiment analysis dataset. This process involves training the models on the labeled data to adapt them specifically for sentiment analysis in Romanized Nepali. If necessary, modifications are implemented to account for the linguistic characteristics of Romanized Nepali. This may include adjusting the tokenization process or incorporating language-specific resources like word embeddings or dictionaries.

3.6 Experiment Setup and Evaluation

For the experimental setup, we divide the dataset into training, validation, and test sets. The training set is used to train the adapted BERT and RoBERTa models, while the validation set is employed for hyperparameter tuning. The test set is then utilized to evaluate the models' performance. We select evaluation metrics such as accuracy, precision, recall, F1-score, or AUC-ROC based on the dataset and specific requirements. The training and evaluation process involves training the adapted models on the training set using the defined experimental setup. We monitor the models' performance by tracking relevant metrics such as loss values. Subsequently, we evaluate the models on the test set to assess their sentiment classification capabilities and compare their performance. This iterative process allows us to identify the most effective transformer-based deep learning model for sentiment analysis in Romanized Nepali.

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{Total Observations}}$$

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

$$\text{F1 Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

4. Result and Analysis

4.1 Human-Coded Dictionary-Based Stemming

Within the provided dataset, there was a comprehensive collection of 32,817 user reviews, encompassing a distribution of sentiments: 22,880 positive, 8,382 negative, and the remaining expressing neutrality. This section outlined the meticulous steps taken in the manual annotation process, which served as the foundation for creating the standardized corpora. Central to this process was the formulation of precise annotation guidelines, followed by the meticulous manual annotation of the entire corpus.

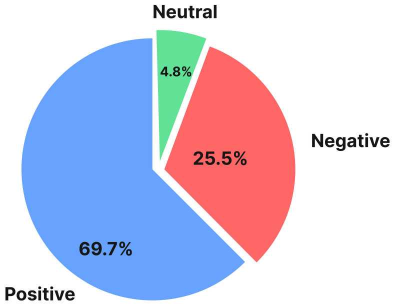


Figure 5: Proportion of Sentiment Categories in Dataset

To guarantee the accuracy and reliability of the annotation process, we engaged the expertise of three proficient annotators. These annotators, each holding a master's degree and possessing native proficiency in Romanized Nepali, undertook the crucial task of annotating every user review. The reviewers' deep familiarity with sentiment analysis further assured the quality of annotations. In order to validate the effectiveness of our guidelines, a subset of 100 reviews underwent evaluation by annotators X and Y. Their independent assessments meticulously categorized each review into one of three distinct sentiment classes: negative, neutral, or positive. Any disparities that arose between annotator X and Y were methodically addressed through the involvement of a third annotator, Z. These resolutions adhered to the aforementioned annotation guidelines. The application of the Cohens Kappa method yielded a noteworthy Inter-Annotator Agreement (IAA) of 81.32 percent, underlining the consistency and reliability of manual annotations. This IAA score, coupled with the moderate ratings, underscores the adept application and understanding of the annotation rules by the specialized annotators.

4.2 Text Preprocessing

The preprocessing of the Romanized Nepali text was necessary to make it usable for NLP tasks. Various elements, including URLs, email addresses, phone numbers, numerical numbers, numerical digits, currency symbols, and punctuation marks, were removed to enhance the accuracy of the model. Additionally, further text preprocessing steps were undertaken to improve the model's effectiveness for Roman Nepali text.

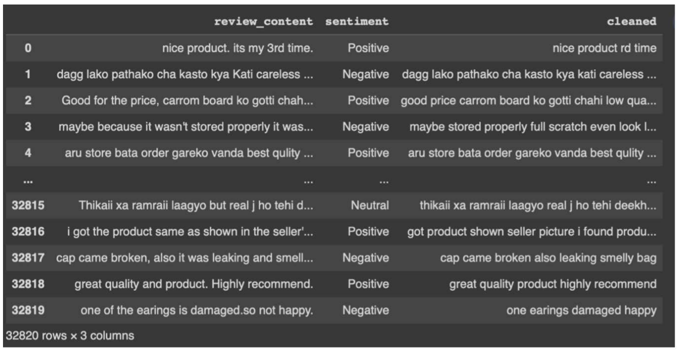
4.2.1 Stop Words

The elimination of commonly used words from the corpus was carried out to improve its quality. However, removing stop words automatically proved to be a challenging task due to the complex morphological structure of the Roman Nepali language and the limited resources available. The steps for removing Roman Nepali stop words were illustrated in the flowchart of Figure 3. A file was created, containing a collection of all commonly used Roman Nepali stop words, and then the corpus was cleared of those words.

Stop words, which are common words that typically do not contribute much meaning to text analysis, such as ”the,” “a,” “and,” “in,” etc., were targeted. In this case, the set of stop words included not only the English language stop words provided by the Natural Language Toolkit (nltk.corpus.stopwords) but also additional words specific to the Roman Nepali language (e.g., “vayo,” “bhayo,” “jhan,” “hai,” “yo,” etc.). These additional words were likely to be frequently used in the analyzed Nepali text data and may not have added significant value to the analysis. Therefore, they were included in the set of stop words to be removed during preprocessing.

4.2.2 Other Normalization

Several normalization steps were performed on the text data to simplify it and eliminate extraneous information that could interfere with the analysis process. The goal was to reduce variations and noise in the text data and provide a cleaner and more standardized representation suitable for sentiment analysis. The text data was transformed into a cleaner form through the application of normalization steps, including conversion to lowercase, removal of numbers, removal of stop words, and removal of punctuation. In addition to the normalization steps, the sentiment analysis model utilized a lemmatizer and stop word removal to enhance the accuracy of the analysis. Lemmatization, which reduced words to their base form, was used to eliminate variations and noise in the text data. Additionally, stop words, which are common words that do not carry any specific meaning in the text, were removed to simplify the text data and focus solely on the sentiment conveyed. As a result, the focus of the sentiment analysis model was solely on the sentiment conveyed in the text, leading to more accurate results. The improvement of the quality and accuracy of the analysis results was achieved through the normalization steps performed.



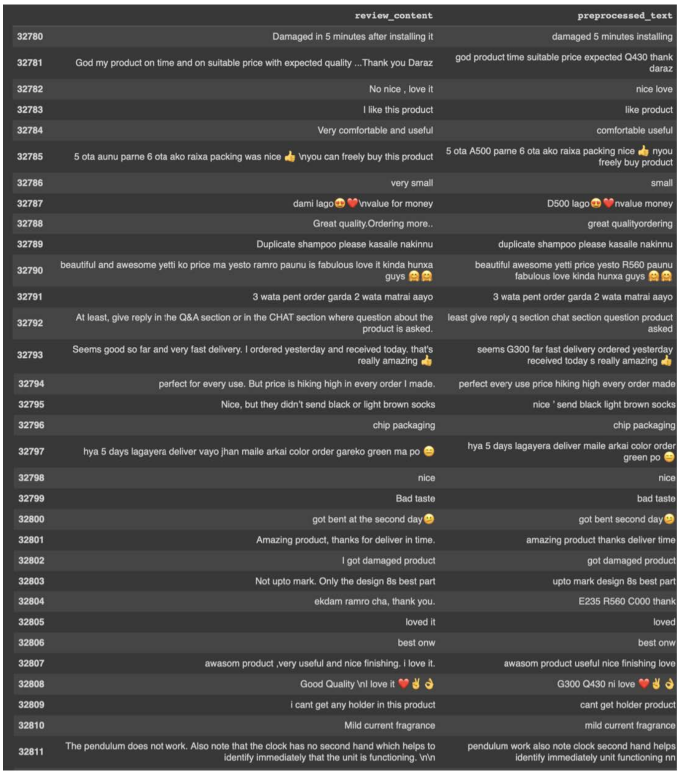
	review_content	sentiment	cleaned
0	nice product. its my 3rd time.	Positive	nice product rd time
1	dagg lako pathako cha kasto kya Kati careless ...	Negative	dagg lako pathako cha kasto kya kati careless ...
2	Good for the price, carrom board ko gotti chah...	Positive	good price carrom board ko gotti chahi low qua...
3	maybe because it wasn't stored properly it was...	Negative	maybe stored properly full scratch even look l...
4	aru store bata order gareko vanda best qulity ...	Positive	aru store bata order gareko vanda best qulity ...
...	...	...	...
32815	Thikaii xa ramraii laagyo but real j ho tehi d...	Neutral	thikaii xa ramraii laagyo real j ho tehi deekh...
32816	i got the product same as shown in the seller'...	Positive	got product shown seller picture i found produ...
32817	cap came broken, also it was leaking and smell...	Negative	cap came broken also leaking smelly bag
32818	great quality and product. Highly recommend.	Positive	great quality product highly recommend
32819	one of the earings is damaged.so not happy.	Negative	one earings damaged happy

32820 rows × 3 columns

Figure 6: Normalized Sentences

4.2.3 Soundex Algorithm

The Soundex technique was crucial to the text preparation in this thesis. Based on how each word is spoken, the Soundex algorithm generates a four-character string that uniquely encodes the sound of each word. With this strategy, it is ensured that words with comparable pronunciations—even if they have different spellings—are grouped together. For instance, the Soundex code for “damme” and “dami” is both “D500,” whereas the Soundex code for “quality” and “qulity” is both “Q430.” The noise in the data was greatly decreased during the normalisation process using the Soundex method, and terms with similar meanings but different spellings were successfully grouped.



	review_content	preprocessed_text
32780	Damaged in 5 minutes after installing it	damaged 5 minutes installing
32781	God my product on time and on suitable price with expected quality ...Thank you Daraz	god product time suitable price expected Q430 thank daraz
32782	No nice , love it	nice love
32783	I like this product	like product
32784	Very comfortable and useful	comfortable useful
32785	5 ota aunu parne 6 ota ako raixa packing was nice 👍 \nyou can freely buy this product	5 ota A500 parne 6 ota ako raixa packing nice 👍 nyou freely buy product
32786	very small	small
32787	dami lago😍❤️\nvalue for money	D500 lago😍❤️nvalue money
32788	Great quality.Ordering more..	great qualityordering
32789	Duplicate shampoo please kasaile nakinnu	duplicate shampoo please kasaile nakinnu
32790	beautiful and awesome yetti ko price ma yesto ramro paunu is fabulous love it kinda hunxa guys 🤗🤗	beautiful awesome yetti price yesto R560 paunu fabulous love kinda hunxa guys 🤗🤗
32791	3 wata pent order garda 2 wata matrai aayo	3 wata pent order garda 2 wata matrai aayo
32792	At least, give reply in the Q&A section or in the CHAT section where question about the product is asked.	least give reply q section chat section question product asked
32793	Seems good so far and very fast delivery. I ordered yesterday and received today. that's really amazing 👍	seems G300 far fast delivery ordered yesterday received today s really amazing 👍
32794	perfect for every use. But price is hiking high in every order I made.	perfect every use price hiking high every order made
32795	Nice, but they didn't send black or light brown socks	nice ' send black light brown socks
32796	chip packaging	chip packaging
32797	hya 5 days lagayera deliver vayo jhan maile arkai color order gareko green ma po 😑	hya 5 days lagayera deliver maile arkai color order green po 😑
32798	nice	nice
32799	Bad taste	bad taste
32800	got bent at the second day😕	got bent second day😕
32801	Amazing product, thanks for deliver in time.	amazing product thanks deliver time
32802	I got damaged product	got damaged product
32803	Not upto mark. Only the design 8s best part	upto mark design 8s best part
32804	ekdam ramro cha, thank you.	E235 R560 C000 thank
32805	loved it	loved
32806	best onw	best onw
32807	awasom product ,very useful and nice finishing. i love it.	awasom product useful nice finishing love
32808	Good Quality \nI love it ❤️✌️👌	G300 Q430 ni love ❤️✌️👌
32809	i cant get any holder in this product	cant get holder product
32810	Mild current fragrance	mild current fragrance
32811	The pendulum does not work. Also note that the clock has no second hand which helps to identify immediately that the unit is functioning. \n\n	pendulum work also note clock second hand helps identify immediately unit functioning nn

Figure 7: Application of Soundex Algorithm

4.3 Performance Evaluation

After assessing the effectiveness of various algorithms, such as Logistic Regression, Naive Bayes, Support Vector Machine, and BERT, using a dataset containing 32,817 observations and employing a 5-fold cross-validation approach, the evaluation incorporated metrics like accuracy, precision, recall, and F1 score. The dataset was initially divided into a training set (80 percent) and a test set (20 percent) for the initial model assessment. In this section, the analysis presents an overview of the outcomes from comparing these algorithms, with a focus on their performance across multiple folds to ensure robustness and reliability. Logistic Regression (LR) model: Our dataset was utilized to assess the Logistic Regression (LR) model, and the results were presented in Table 6. The LR model achieved an accuracy of 74%, highlighting its ability to classify situations accurately. With a precision score of 0.76, it frequently made accurate predictions for positive emotions. The F1-Score of 0.75 indicated a balanced trade-off between recall and precision. The model effectively captured a substantial portion of positive sentiment events, as evident from the recall score of 0.75. Despite the LR model's overall good performance, further optimization could enhance precision.

Naive Bayes (NB) model: Achieving an accuracy of 73%, the NB model demonstrated reasonable classification capabilities. The precision score of 0.72 indicated its ability to provide dependable predictions for positive sentiment. With an F1-Score of 0.73, it struck a balance between precision and recall. The recall score of 0.73 highlighted the model's effectiveness in capturing a substantial portion of positive sentiment instances. While the NB model delivered a competitive performance, there was room for improvement through fine-tuning to enhance precision.

Support Vector Machine (SVM) model: The SVM model displayed strong classification capabilities by achieving an accuracy of 75%. Its accuracy in foretelling favourable thoughts was demonstrated by its precision score of 0.74. It demonstrated a well-balanced trade-off between recall and precision with a remarkable F1-Score of 0.77. The model was successful in accurately identifying occurrences of positive emotion, as evidenced by the recall score of 0.75. The SVM model showed promise for a variety of applications while excelling at sentiment analysis.

BERT model: BERT demonstrated its sophisticated capabilities in sentiment classification by achieving a noteworthy accuracy of 79%. The F1-Score of 0.75 indicated an efficient balance between recall and accuracy, while the precision score of 0.77 indicated reliable predictions of positive mood. The BERT model performed exceptionally well in capturing instances of positive sentiment, with a recall score of 0.76. Overall, BERT fared better than the other models in terms of accuracy and showed great promise for tasks involving sentiment analysis.

Table 2: LR model's performance for the five-fold cross-validation

Fold	Accuracy	Precision	F1 Score	Recall
Fold 1	0.80	0.78	0.77	0.78
Fold 2	0.75	0.75	0.76	0.74
Fold 3	0.72	0.77	0.73	0.76
Fold 4	0.73	0.79	0.74	0.73
Fold 5	0.70	0.74	0.70	0.77
Average	0.74	0.76	0.75	0.75

Table 3: NB model's performance for the five-fold cross-validation

Fold	Accuracy	Precision	F1 Score	Recall
Fold 1	0.70	0.70	0.75	0.70
Fold 2	0.71	0.72	0.73	0.71
Fold 3	0.72	0.73	0.72	0.73
Fold 4	0.74	0.71	0.74	0.75
Fold 5	0.77	0.75	0.71	0.76
Average	0.73	0.72	0.73	0.73

Table 4: SVM model's performance for the five-fold cross-validation

Model	Accuracy	Precision	f1-Score	Recall
Fold 1	0.74	0.72	0.78	0.74
Fold 2	0.75	0.73	0.77	0.75
Fold 3	0.76	0.75	0.76	0.76
Fold 4	0.74	0.75	0.76	0.74
Fold 5	0.76	0.76	0.78	0.76
Average	0.75	0.74	0.77	0.75

Table 5: BERT model's performance for the five-fold cross-validation

Fold	Accuracy	Precision	F1 Score	Recall
Fold 1	0.80	0.78	0.85	0.83
Fold 2	0.79	0.77	0.70	0.81
Fold 3	0.79	0.76	0.71	0.76
Fold 4	0.78	0.76	0.71	0.71
Fold 5	0.80	0.78	0.70	0.71
Average	0.79	0.77	0.75	0.76

Table 6: Overall Average result of Performance Metrics for Models

Fold	Accuracy	Precision	F1 Score	Recall
LR	0.74	0.76	0.75	0.75
NB	0.73	0.72	0.73	0.73
SVM	0.75	0.74	0.77	0.75
BERT	0.79	0.77	0.75	0.76

LR, NB, and SVM put up strong performances, displaying admirable aptitudes for sentiment classification accuracy. However, LR and NB suggested that fine-tuning could be able to improve precision. However, SVM demonstrated strong classification abilities with a good trade-off between recall and precision, making it adaptable for a variety of applications. Notably, BERT distinguished itself as a top performer with great recall and accuracy, making it a superb option for precision-focused sentiment analysis tasks.

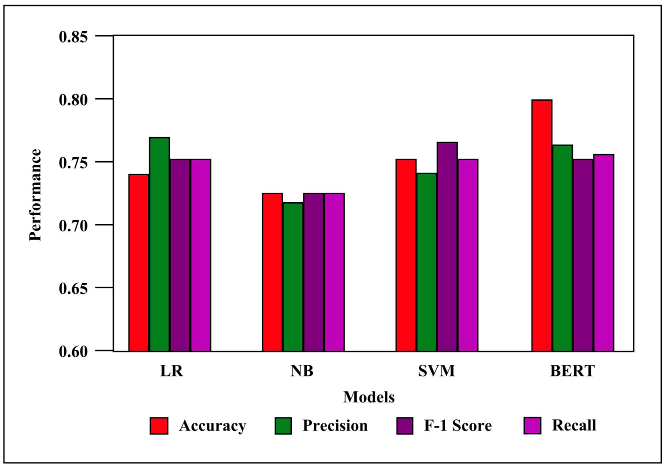


Figure 8: Performance comparison of all models

5. Discussion and Conclusion

In the modern era, sentiment analysis has become a crucial tool for businesses to monitor customer feedback from various sources, including social media, customer reviews, and surveys. With advancements in deep learning, sentiment analysis algorithms have become more powerful, enabling businesses to conduct in-depth analysis and understand customer sentiment. However, one of the major challenges in sentiment analysis is the language used in the text data. For instance, around 45 million people worldwide speak Nepali, which is the mother-tongue of Nepal, and most of them are not proficient in English. This leads to the use of Romanized Nepali script by people in Nepal who are more comfortable with Romanized Nepali words or sentences.

Romanized Nepali is a unique combination of English/Latin alphabet and Nepali language, making it incredibly dynamic and irregular. People in Nepal use Romanized Nepali text to share their comments and views on e-commerce platforms. However, due to the irregularity and dynamic nature of Romanized Nepali, it is hard to detect the sentiment using typical sentiment analysis techniques. This study aims to address this challenge by conducting sentiment analysis on Romanized Nepali text data.

The pre-processing step of stemming is crucial in sentiment analysis, as it reduces words to their root form and simplifies the text while also improving accuracy in sentiment classification. However, for Romanized Nepali, there is currently no available stemmer. Therefore, a human-coded dictionary-based stemming approach was suggested. A study was conducted to examine the effectiveness of sentiment analysis on Romanized Nepali using a range of machine learning techniques. The primary goal of this research is to simplify the text data and reduce variability, resulting in more accurate sentiment classification and a better understanding of customer sentiment. The findings of this study could potentially assist businesses in Nepal in improving their corporate strategies and meeting the needs and demands of their customers.

Acknowledgments

The authors express their gratitude to the Department of Electronics and Computer Engineering, Pulchowk Campus, for offering the opportunity to work on this research. The department provided both direct and indirect, yet invaluable, support and assistance throughout the conception of this research.

References

[1] Walaa Medhat, Ahmed Hassan, and Hoda Korashy. Sentiment analysis algorithms and applications: A survey. Ain Shams Engineering Journal, 5(4):1093–1113, 2014.

[2] Ema Ku˘sen, Giuseppe Cascavilla, Kathrin Figl, Mauro Conti, and Mark Strembeck. Identifying emotions in social media: Comparison of word-emotion lexicons. pages 132– 137, 2017.

[3] Hai Ha Do, PWC Prasad, Angelika Maag, and Abeer Alsadoon. Deep learning for aspect-based sentiment analysis: A comparative review. Expert Systems with Applications, 118:272–299, 2019.

[4] Kerstin Denecke. Using sentiwordnet for multilingual sentiment analysis. pages 507–512, 2008.

[5] Feilong Tang, Luoyi Fu, Bin Yao, andWenchao Xu. Aspect based fine-grained sentiment analysis for online reviews. Information Sciences, 488:190–204, 2019.

[6] M. V. Koroteev. BERT: A review of applications in natural language processing and understanding. CoRR, abs/2103.11943, 2021.

[7] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. 2019.

[8] Shan Suthaharan. Support vector machine. pages 207–235, 2016.

[9] Anjuman Prabhat and Vikas Khullar. Sentiment classification on big data using na¨ıve bayes and logistic regression. pages 1–5, 2017.

[10] L. Jiang, Z. Cai, and D. Wang. Improving naive bayes for classification. International Journal of Computers and Applications, 32(3):328–332, 2010.

[11] Gagan Reddy Narayanaswamy. Exploiting bert and roberta to improve performance for aspect based sentiment analysis. 2021.

[12] Rosario Catelli, Serena Pelosi, and Massimo Esposito. Lexicon-based vs. bert-based sentiment analysis: A comparative study in italian. Electronics, 11(3), 2022.

[13] Archa Joshy and Sumod Sundar. Analyzing the performance of sentiment analysis using bert, distilbert, and roberta. pages 1–6, 2022.

[14] Victoria Bobicev and Marina Sokolova. Inter-annotator agreement in sentiment analysis: Machine learning perspective. pages 97–102, September 2017.