Institute of Technology of Cambodia

Department of Applied Mathematics and Statistics

Khmer Sentiment Analysis Using Machine Learning

Group Members

No.	Full Name	Student ID
1	Vong Pisey	e20210599
2	HENG Seaklong	e20210329
3	Ing Vitourotanak	e20210519
4	Haysavin Rongravidwin	e20211502
5	Chhorn Solita	e20210537
6	Long Ratanakvichea	e20210086

Lecturers

Dr. Khean Vesal *(Course Lecturer)*

Mr. Khean Vesal *(TP Lecturer)*

Academic Year 2025–2026

Contents

1 Introduction

In recent years, the rapid growth of digital communication has led to an enormous amount of textual data being generated on social media platforms, online forums, and review websites. Understanding public opinion and emotional tone within this data has become an important task in Natural Language Processing (NLP). Sentiment Analysis, also known as opinion mining, is the computational study of opinions, sentiments, and emotions expressed in text.

Most existing sentiment analysis systems focus on high-resource languages such as English, Chinese, and French. However, low-resource languages such as Khmer face significant challenges due to the lack of labeled datasets, linguistic tools, and pretrained language models. Khmer language has unique grammatical structures, lacks clear word boundaries, and exhibits complex contextual meanings, making sentiment classification particularly difficult.

This project focuses on developing and evaluating sentiment analysis models for Khmer text. Both traditional machine learning algorithms and deep learning models are explored and compared to understand their effectiveness under limited data conditions.

2 Problem Statement

Despite the increasing use of Khmer language on digital platforms, there is still no widely accepted or highly accurate sentiment analysis system tailored specifically for Khmer. Existing NLP models trained on other languages cannot be directly applied due to linguistic differences. Furthermore, deep learning models often require large datasets, which are not readily available for Khmer.

Therefore, the main problem addressed in this project is how to effectively classify sentiment in Khmer text using available data and models, and which modeling approach is more suitable under data-scarce conditions.

3 Objectives

The objectives of this research are summarized as follows:

- To study sentiment analysis techniques applicable to low-resource languages.
- To preprocess and represent Khmer text for machine learning models.
- To implement and compare traditional machine learning models including Logistic Regression, SVM, Random Forest, Naive Bayes, and XGBoost.
- To implement a Bidirectional LSTM model for sentiment classification.

- To evaluate all models using Accuracy, Precision, Recall, F1-score, and Cross-Validation scores.
- To analyze strengths and weaknesses of each model for Khmer sentiment analysis.

4 Data Collection and Preprocessing

1	text	target
2	ខ្ញុំស្រលាញ់ផលិតផលនេះណាស់	positive
3	មិនពេញចិត្តទេតើអ្វីកើតឡើង?	negative
4	ផលិតផលមធ្យមៗគ្មានអ្វីពិសេស	neutral
5	ពលរដ្ឋ៣ភូមិក្នុងស្រុកត្រាំកក់ទទួលបានផ្លូវបេតុងមួយខ្សែដើម្បីសម្រួលដល់ការធ្វើដំណើរ:://	positive
6	ផ្លូវបេតុងនេះមានប្រវែងជាង១៦00ម៉ែត្រ	neutral
7	ប្រជាពលរដ្ឋចំនួន៣ភូមិនឹងលែងជួបការលំបាកក្នុង?ការធ្វើដំណើរ	positive
8	សិស្សានុសិស្សដែលធ្វើដំណើរឆ្លងកាត់នឹងទទួលបានភាពងាយស្រួល	positive
9	អាជ្ញាធរស្រុកត្រាំកក់បានបើកការដ្ឋានសាងសង់ផ្លូវបេតុងមួយខ្សែ	positive
10	ផ្លូវបេតុងនេះមានប្រវែង១៦៣៧ម៉ែត្រ//	neutral
11	ពិធីបើកការដ្ឋានធ្វើឡើងក្រោមអធិបតីភាពលោកកេតម៉ៅ	neutral
12	លោកកេតម៉ៅជាអភិបាលនៃគណៈអភិបាលស្រុកត្រាំកក់	neutral
13	គម្រោងផ្លូវនេះជាគម្រោងមូលនិធិឃុំលាយបូរឆ្នាំ២០២៥	neutral
14	ផ្លូវមានទទឹង៦ម៉ែត្រនិងកម្រាស់១៥0ម៉ែត្រ	neutral
15	ផ្លូវនេះស្ថិតនៅក្នុងភូមិត្រពាំងព្រីង	neutral
16	ផ្លូវនេះក៏ស្ថិតនៅក្នុងភូមិអង្គគគីរ	neutral
17	ការសាងសង់ផ្លូវនេះឆ្លើយតបនឹងការទន្ទឹងរង់ចាំរបស់ប្រជាពលរដ្ឋ	positive
18	ផ្លូវបេតុងនេះជួយសម្រួលដល់ការធ្វើដំណើរប្រចាំថ្ងៃ	positive

Figure 1: Sample of Khmer Sentiment Dataset with Text and Target Labels

The dataset used in this project consists of Khmer-language text labeled with sentiment categories. The data was collected from publicly available online sources, including social media posts and user comments totality 1057 row. Each text instance was manually or semi-automatically labeled into sentiment classes.

4.1 Text Preprocessing

Before training the models, several preprocessing steps were applied:

- Removal of punctuation, numbers, and special characters.
- Normalization of Khmer Unicode text.
- Tokenization of Khmer sentences.
- Stop-word removal where applicable.

4.2 Feature Representation

For traditional machine learning models, text data was transformed into numerical vectors using techniques such as Bag-of-Words (BoW) and Term Frequency–Inverse Document Frequency (TF-IDF). For the deep learning model, sentences were converted into sequences of integer indices and padded to a fixed length.

5 Methodology

Let the dataset be formally defined as:

$$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$$

where x_i denotes the input Khmer text and $y_i \in \{1, 2, \ldots, C\}$ represents the sentiment label.

5.1 Logistic Regression

Logistic Regression is a linear classifier that estimates the probability of class membership using the softmax function:

$$P(y = c \mid x) = \frac{\exp(\mathbf{w}_c^\top \mathbf{x} + b_c)}{\sum_{k=1}^{C} \exp(\mathbf{w}_k^\top \mathbf{x} + b_k)}$$

The model parameters are learned by minimizing the cross-entropy loss:

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{C} y_{ic} \log P(y = c \mid x_i)$$

5.2 Support Vector Machine (SVM)

Support Vector Machine aims to find a hyperplane that maximizes the margin between classes. For multi-class classification, one-vs-rest strategy is applied.

The optimization problem is:

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N} \xi_i$$

subject to:

$$y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 - \xi_i$$

5.3 Random Forest

Random Forest is an ensemble learning method composed of multiple decision trees trained on bootstrapped samples. Each tree produces a class prediction, and the final output is obtained via majority voting:

$$\hat{y} = \arg\max_{c} \sum_{m=1}^{M} \mathbb{I}(T_m(x) = c)$$

5.4 Naive Bayes

Naive Bayes classifier is based on Bayes' theorem and assumes conditional independence among features:

$$P(y \mid x_1, \ldots, x_d) = \frac{P(y)\prod_{j=1}^{d} P(x_j \mid y)}{P(x_1, \ldots, x_d)}$$

The predicted label is:

$$\hat{y} = \arg\max_{y} P(y)\prod_{j=1}^{d} P(x_j \mid y)$$

5.5 XGBoost

XGBoost builds an ensemble of decision trees in a sequential manner. The prediction is given by:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i)$$

where each f_k represents a regression tree. The objective function is:

$$\mathcal{L} = \sum_{i=1}^{N} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

5.6 Bidirectional LSTM

Bidirectional LSTM captures long-range dependencies by processing sequences in both forward and backward directions:

$$\overrightarrow{h_t} = \text{LSTM}(x_t, \overrightarrow{h_{t-1}})$$

$$\overleftarrow{h_t} = \text{LSTM}(x_t, \overleftarrow{h_{t+1}})$$

The combined hidden state is:

$$h_t = [\overrightarrow{h_t}; \overleftarrow{h_t}]$$

6 Results and Discussion

Table 1: Performance Comparison of Models

Model	Accuracy	F1-Macro	Precision	Recall	Best CV Score
Logistic Regression	0.5166	0.5090	0.5055	0.5156	0.4616
SVM	**0.5403**	**0.5089**	**0.5260**	0.5012	0.4518
Random Forest	0.5213	0.4841	0.4852	0.4845	0.4712
Naive Bayes	0.5118	0.4371	0.5127	0.4352	0.4219
XGBoost	0.5071	0.4188	0.4756	0.4231	0.4269
Bidirectional LSTM	0.4834	0.2444	0.2951	0.3452	0.4675

The experimental results indicate that traditional machine learning models outperform the deep learning approach in this study. The SVM achieved the highest accuracy and precision, demonstrating its robustness in high-dimensional sparse feature spaces. Logistic Regression also performed consistently across all metrics, suggesting that linear classifiers are well-suited for TF-IDF representations.

Random Forest showed moderate performance but struggled to generalize due to sparse input features. Naive Bayes and XGBoost produced lower F1-scores, indicating difficulties in capturing sentiment-related patterns in Khmer text. The Bidirectional LSTM underperformed, likely due to insufficient training data and lack of pretrained embeddings.

7 Conclusion

This project presented a comprehensive study of Khmer sentiment analysis using multiple machine learning and deep learning models. Results show that under limited data conditions, traditional machine learning methods such as SVM and Logistic Regression provide superior performance compared to deep learning models. Future work should explore larger datasets, transfer learning, and transformer-based models to further improve sentiment classification accuracy.

8 References

1. Pang, B., & Lee, L. (2008). Opinion Mining and Sentiment Analysis.

2. Bishop, C. M. (2006). Pattern Recognition and Machine Learning.

3. Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory.

4. Chen, T., & Guestrin, C. (2016). XGBoost.