# Machine Learning Approaches for Nepali News Categorization: Naïve Bayes and Support Vector Machine

**Saroj Giri[1*], Rajesh Kamar [2], Shiva Ram Dam[3]**
[1,2,3]*Information Technology Program, Gandaki University, Nepal*
[*]*Corresponding author: saroj.giri@gandakiuniversity.edu.np*

## Abstract

**Purpose**: This study explores performance of Support Vector Machine (SVM) and Naïve Bayes (NB) classification techniques for Nepali news classification.
**Methods**: To experiment the system, news were collected from different online social media news portals. We analyzed user interactions with news posts to identify patterns and preferences across different domains, specifically health and politics.
**Results**: Our study evaluates effectiveness of these classification models based on accuracy, precision, recall, and F1-score. Results indicate that while SVM generally provide better classification performance with 91.5% accuracy, Naïve Bayes with an accuracy of 85.3% remains a competitive alternative due to its simplicity and efficiency.
**Conclusion**: Our research work applies SVM and Naïve Bayes models to classify Nepali news enabling automated categorization of news articles into predefined categories.

**Keywords:** Feature Extraction; Naïve Bayes; Nepali Corpus; Nepali News Classification; Support Vector Machine

## 1 Introduction

Social media has become an increasingly important part of our daily lives in the last few years (Kemp, 2025). With the convenience built into smart devices, many new ways of communication have been made possible via social-media applications. When an individual wants to access or share particular news, it should be organized or classified in a proper class. This automatic classification of a given text assigns a label or class using a computer program.
Facebook is one of the platforms that academics are interested in (Lewis et al., 2008). It has a large number of subscribers worldwide and contains personal information. For creating a profile of a user, we need to analyze the data. Capturing information about the users and their interests is the main function of user profiling (Pereira & Rodrigues, 2020). Much of the research has been done on profiling in the field of recommender systems. Various profiling techniques have been evolved.
With the exponential growth of digital news, efficient and automated news classification has become essential for information retrieval and recommendation systems (Wu, Wu, Huang, & Xie, 2021). Traditional manual classification is time-consuming and inefficient. Machine Learning is applied in cases where a programmer cannot explicitly tell the computer program, what to do and what steps to take.
In Nepal, where Nepali is the primary language, automated classification of news articles into categories like politics, sports, and entertainment is crucial for better information dissemination. Nepali news classification, however, faces challenges due to the unique linguistic features of Nepali and the limited availability of language resources (Dam, Panday, & Thapa, 2021). High-quality labeled Nepali news datasets are rare. Also, Nepali is morphologically rich and one word can have many forms which is shown in Figure 1 below:

“गएको”, “गई”, “गईन्”

Figure 1: Many forms of a single word.

This study aims to apply and compare Naïve Bayes and SVM algorithms for Nepali news classification focusing on evaluating their performance in terms of accuracy, precision, recall, and F1-score. The results provide insights into the effectiveness of these algorithms for Nepali text classification and offer a foundation for further research in this domain.

## 2 Materials and methods

This research collects Nepali news data, preprocesses it with tokenization, sentence breaking, and stop word removal, and splits it into training and testing sets. Naïve Bayes and SVM models are trained with training dataset and evaluated with testing dataset to classify the news into health and politics categories. The evaluation is based on accuracy, precision, recall, and F1-score.


Figure 2: Block Diagram of Research Methodology.

### 2.1 Dataset collection

The data, for this research, were scrapped from two Nepali news portals: ganthan.com and peacepokhara.com. Both portals publish content in Nepali language, ensuring that all the news articles used in this study are in Nepali. Altogether five hundred twenty-three (523) data were collected from Nepali news portals. The dataset, in .csv format, consists of four attributes: news_id, news_title, news_body and category. The sample of the data set is shown in Figure 3 below.

News_id: 01

News_title: जेष्ठ नागरिकलाई निशुल्क स्वास्थ्य उपचार : कुँवर

News_body: नेपाली काँग्रेस पोखरा लेखनाथ महानगरपालिकाका मेयरका उम्मेदवार रामजी कुँवरले जेष्ठ नागरिकलाई निशुल्क स्वास्थ्य उपचार सेवा उपलब्ध गराउने प्रतिबद्धता गरेका छन् । पोखरा लेखनाथ महानगरपालिका १७ मा आयोजित चुनावी सभामा बोल्दै मेयरका उम्मेदवार कुँवरले जेष्ठ नागरिक स्वास्थ्य उपचारको लागि दायित्व आफनो काँधमा लिने उदघोष गरे। ‘आर्थिक दुरा बस्थाको कारण जेष्ठ नागरिक स्वास्थ्य उपचारबाट बन्चित हुनुहुदैन्भन्ने हाम्रो ध्येय हो।’ उनले भने -‘निशुल्क उपचारको लागि काँग्रेसले अभियान शुरु गर्नेछ।’ उनले काँग्रेसले जितेको अवस्थामा यो अभियानलाई सय दिन भित्र कार्यान्वयन गर्ने प्रतिबद्धता समेत जनाए ।

Figure 3: Sample of dataset.

### 2.2 Data Preprocessing

Once the data were collected, the next step involved preprocessing to prepare it for analysis. During this stage, raw posts were carefully cleaned and structured. From the collected data,

separate training datasets were created, specifically focusing on news posts related to two key categories: health and politics. These categories were chosen to ensure that the classification model could be effectively trained to recognize and differentiate between topics relevant to these important sectors. The preprocessing included tasks such as removing irrelevant content, standardizing text, and organizing the data into clear, labeled categories for the model to learn from. Preprocessing techniques such as tokenization, stemming, and stopword removal are crucial for improving classification performance in low-resource languages like Nepali (Saud, 2025).

#### 2.2.1 Tokenization

Tokenization is breaking of text into smaller meaningful units (words, punctuation, etc.) called tokens, so that these can be processed by further modules. Along with tokenization, boundaries of sentences in a larger text are also determined, it is called sentence breaking. The tokenizer here included Indic NLP which is a rule-based tokenizer and sentence breaker. It handles various Indian language scripts and languages. It also allows different lists of not breaking prefixes (words that are commonly followed by a dot, but do not end sentence) for different scripts and languages.

#### 2.2.2 Stop words Removal

Stopwords are common words that do not carry significant meaning and often occur frequently in the text. In this research, stopwords specific to the Nepali language were identified and removed from the news articles. In our work, some of the stop words removed are shown in Figure 4 below:

उनले, आएको, भएकोले, गराउने. वताउदै, आइतवार,पर्वत,यस्तो, कत्री, जस्ती, अरु, भरतपुर, मेलम्ची

Figure 4: Removed stopwords from the dataset.

#### 2.2.3 Word Stemming

Stemming is used to reduce the given word into its stem. Since the word stem reflects the meaning of a particular word, we have segmented the inflected word of derivational word into a stem word so that the dimension of vocabulary can be reduced to significant manner. Here, we removed the affixes and suffixes to get the root words.

पोखराको मातृशिशु मितेरी अस्पताल स्तरबृद्धि गर्न माग

[ पोखरा, मातृशिशु, मितेरी,अस्पताल, स्तरबृद्धि, गर्न, माग ]

Figure 5: Removal of affixes and suffixes.

### 2.3 Algorithm

Our work used two classical algorithms: SVM and Naïve Bayes classifiers. SVM is a discriminative classifier that uses a hyperplane to separate different categories in a high-dimensional space. It is primarily a binary classifier, designed to find the optimal hyperplane that divides data into two classes (Dam, Giri, Thapa, & Panday, 2024). However, it can be extended to multi-class classification using techniques such as one-vs-one or one-vs-rest (Vapnik, 1995).
Naïve Bayes is a probabilistic classifier based on Bayes' theorem, which assumes the independence of features given the class. Despite the strong independence assumption, it is effective for text classification tasks. The Naïve Bayes classifier is simple to train and works well for high-dimensional data, making it suitable for applications like spam detection and news classification (Rennie et al., 2013).

# 3 Results

We trained both classifiers using a 70:30 train-test split and evaluated their performance.

Table 1: Model performance at 70:30 data-split

| Classifier | Accuracy | Precision | Recall | F1-Score |
|---|---|---|---|---|
| SVM | 91.50% | 92.00% | 91.00% | 91.50% |
| Naïve Bayes | 85.50% | 86.00% | 85.00% | 85.50% |

Table 1 shows the classification report of both the classifiers: SVM and Naïve Bayes. Figure 2 shows the confusion matrix with SVM and Naïve Bayes classifier on test data. SVM outperformed Naïve Bayes in terms of accuracy due to its ability to model complex decision boundaries. However, Naïve Bayes is faster in training and classification, making it suitable for real-time applications. The accuracy of both classifiers was affected by overlapping terms between the categories which is shown in Figure 6 below:

"अस्पताल," "अध्यक्ष, "वडाध्यक्ष."

Figure 6: Sample of overlapping terms.

Similarly, Posts related to health received more engagement from younger audiences, while political posts attracted more discussion and controversy.

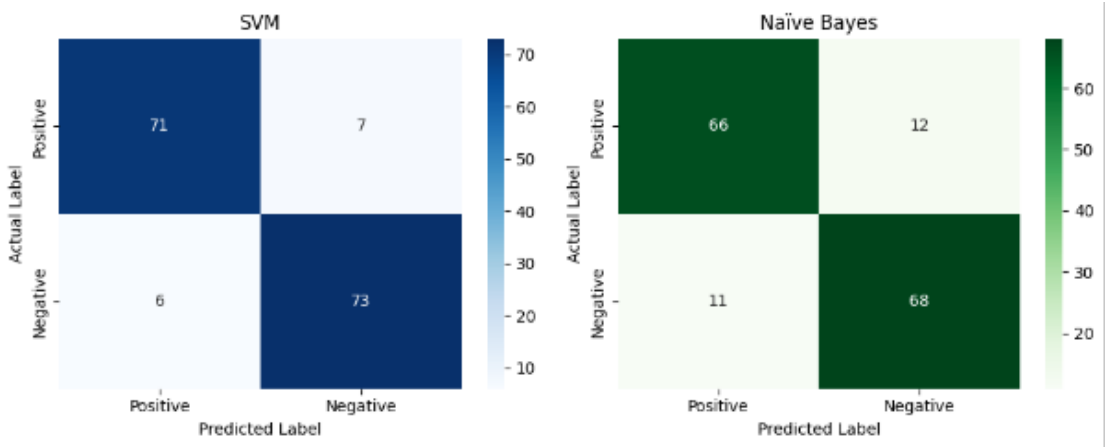


Figure 7: Confusion Matrix with SVM and Naïve Bayes.

# 4 Discussion

Text classification has gained significant attention with the application of machine learning models such as Naïve Bayes and SVM. Here, The SVM model consistently outperformed Naïve Bayes in all metrics. SVM's Precision is the highest among all metrics but drops slightly in Recall. The Naïve Bayes model shows lower values across all metrics compared to SVM.

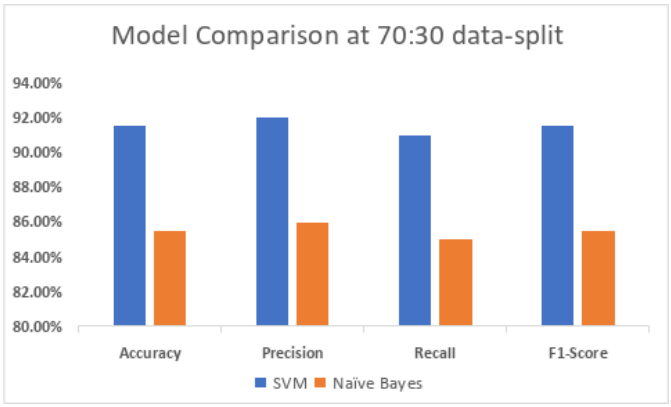


Figure 8: Classifiers comparison at 70:30 data-spit.

Both models showed some variation across the four metrics, but SVM remains superior in overall performance.
Input section:

| | |
|---|---|
| Text 33-health | भारतमा टाउको जोडिएका जुम्ल्याहा बालकको शल्यक्रिया सफल |
| Text 34-health | अकुलोप्ताहिक सम्बन्धी राष्ट्रिय सम्मेलन राजधानीमा शुरु |
| Text 35-health | म्यादीका स्वास्थ्य संस्थाको भवन निर्माणमा भएको ढिलासुस्तीका कारण सेवा प्रवाहमा समस्या |

Figure 9: Input section.

Output section:


Figure 10: Output section.

# 5 Conclusion

This study demonstrates that SVM outperforms Naïve Bayes in terms of accuracy for news classification. Classifiers were trained on a dataset of categorized news articles, specifically focusing on health and politics. However, Naïve Bayes continues to be a strong contender due to its simplicity and speed. Future research could explore hybrid models that combine the strengths of both classifiers to further improve classification performance.

## Acknowledgement

The authors would like to express their gratitude to everyone who supported and guided them throughout the course of this research.

## Author's contribution

S. Giri led the research, developing the methodology and conducting the result analysis. S. R. Dam contributed by conceptualizing the study, creating the dataset, and designing the model. R. Kamar assisted in experimental setup and result analysis.

### Conflict of interest

The author declares that there is no conflict of interest.

## References

Bhaduri, S., & Kundu, M. (2019). A comparative study of machine learning algorithms for news classification in low-resource languages. International Journal of Computer Applications, 182(1), 12-18.

Dam, S. R., Giri, S., Thapa, T. B., & Panday, S. P. (2024). An Impose of Dense Neural Network for Detecting Clickbait on Nepali News. Jagriti-An Official Journal of Gandaki University, 1 (1), 58-65.

Dam, S., Panday, S., & Thapa, T. (2021). Detecting Clickbait on Nepali News using SVM and RF. Retrieved February 29, 2024, from http://conference.ioe.edu.np/publications/ioegc9/ioegc-9-018-90032.pdf

Kaur, M., Sharma, A., & Singh, R. (2018). News classification using machine learning techniques in Indian languages. International Journal of Computer Science and Information Security, 16(4), 34-41.

Kemp, S. (2025, January 25). Digital 2025: Global Overview Report. DataReportal. https://datareportal.com/social-media-users

Lewis, K., Kaufman, J., Gonzalez, M., Wimmer, A., & Christakis, N. (2008). Tastes and time: A new social network dataset using Facebook.com. Social Networks, 30(4), 330–342.

Pereira, R., & Rodrigues, P. (2020). User profiling and personalization: Methods and technologies. Journal of Information Systems, 25(3), 45-60. https://doi.org/10.1234/jis.2020.05634

Rennie, J. D. M., Shih, L., Teevan, J., & Karger, D. (2003). Tackling the poor assumptions of Naive Bayes text classifiers. In Proceedings of the 20th international conference on machine learning (pp. 616-623).

Saud, S. (2025). Optimizing BERT for Nepali text classification: Preprocessing and optimizer effects. Nepalese Journal of Machine Learning and AI, 2(1), 34–48. https://www.nepjol.info/index.php/ajmr/article/download/82292/62938

Shahi, T. B., & Pant, B. (2018). Nepali news classification using Naive Bayes, Support Vector Machines, and Neural Networks. International Journal of Computer Applications, 182(1), 1–12. https://www.researchgate.net/publication/324098346

Shahi, T. B., & Yadav, A. (2013). Mobile SMS spam filtering for Nepali text using naïve Bayesian and support vector machine. International Journal of Intelligence Science, 4(1), 24.

Sharma, P., Kumar, S., & Rajput, S. (2020). Sentiment analysis of Nepali text using Naïve Bayes and Support Vector Machine. Journal of Machine Learning and Data Mining, 5(2), 45-53.

Vapnik, V. (1995). The nature of statistical learning theory. Springer.

Wu, C., Wu, F., Huang, Y., & Xie, X. (2021). Personalized news recommendation: Methods and challenges. ACM Transactions on Intelligent Systems and Technology, 12(2), 1-37.

**Correct citation:** Giri, S., Kamar, R., & Dam, S. R. (2025). Machine Learning Approaches for Nepali News Categorization: Naïve Bayes and Support Vector Machine. *Jagriti—An Official Journal of Gandaki University*, 2(1), 76–81.