# Sentiment Analysis on Nepali Movie Reviews using Machine Learning

Ashok Kumar Pant
Central Department of Computer Science & IT, TU,
Kirtipur, Kathmandu, Nepal
Email: ashokpant87@gmail.com

Abhimanu Yadav
College of Applied Business
Naksal, Kathmandu, Nepal
Email: abhimanupro@gmail.com

***Abstract*—This research article presents machine learning methods for detecting the sentiment expressed by movie reviews. The semantic orientation of a review can be positive or negative. Analysis of opinion for particular product, news or document could be beneficial to many companies, institutions and individuals for marketing, advertising, question answering, product selection and so on. We have created a Nepali movie review dataset with total $500$ samples having $250$ samples per each positive and negative class of sentiment from various online sources. Sentiment analysis system implements various natural language processing techniques for document preprocessing and feature extraction. Naive Bayes based machine learning technique is used for the classification of the sentiment. Empirical results shows, classification accuracies are, $79.23\%$ of precision, $78.57\%$ of recall and $78.90\%$ of F-score.**

Index Terms: Sentiment Analysis, Nepali Movie Review, Natural Language Processing, Machine Learning, Naive Bayes.

## I. INTRODUCTION

Sentiment analysis is a natural language processing task that deals with the identification and extraction of subjective information or the opinion from the given text documents. It determines the attitude or the contextual polarity of the document. Sentiment analysis carries the basic task of classification of the expressed opinion in a document into "positive", "negative", or "neutral" class. Beyond polarity, sentiment classification can be used with the emotional states such as "happy", "sad", and "angry."

Recently, sentiment analysis has taken great interests as the rise of social media such as blogs and social networks. With the proliferation of reviews, ratings, recommendations and other forms of online expression, online opinion has turned into a kind of virtual currency for businesses looking to market their products, identify new opportunities and manage their reputations. It can also be used to make decisions to purchase or to use services by individuals. Ad market can use sentiment analysis to place ads on praised sites. And,sentiment analysis can also be used for opinion retrievals [1].

Movie review sentiment analysis can be done to identify user attitude and opinion toward particular movie. In this research, we have considered two class of global subjective polarity (positive and negative) of movie review texts.

Sentiment classification of reviews has been the focus of recent research. It has been attempted in different domains such as movie reviews, product reviews, and customer feedback reviews [2]–[7]. Much of the research untill now has focused on training Machine Learning algorithms such as Support Vector Machines [8] to classify reviews. Research has also been done on positive/negative term-counting methods and automatically determining if a term is positive or negative [9]. Some of other machine learning techniques used in sentiment analysis include statistical approaches [10], Fuzzy Logic [11], Probabilistic models [12], [13],and Neural networks [14]–[16]. Different techniques and applications of sentiment analysis can be found in [1]. Movie review sentiment analysis techniques are described in [17]–[20]

## II. RESEARCH METHODOLOGY

There exist three approaches towards sentiment analysis; machine learning based methods, lexicon based methods and linguistic analysis. Machine learning methods are based on training an algorithm, mostly classification on a set of selected features for a specific mission and then test on another set whether it is able to detect the right features and give the right classification. A lexicon based method depends on a predefined list or corpus of words with a certain polarity. An algorithm is then searching for those words, counting them or estimating their weight and measuring the overall polarity of the text. Lastly the linguistic approach uses the syntactic characteristics of the words or phrases, the negation, and the structure of the text to determine the text orientation. This approach is usually combined with a lexicon based method.

This research paper deals with word level feature extraction method for machine learning based sentiment analysis.

### A. System Model

Figure 1 shows the top level sentiment classification system for Nepali movie reviews. It is divided into four sub-systems, data acquisition, preprocessing, feature extraction, and classification.

## III. PREPROCESSING

Pre-processing the data is the process of cleaning and preparing the text for feature extraction and classification. In this stage, noise and uninformative text are removed from the input text document [21]. Keeping those words makes the dimensionality of the problem high and hence the classification

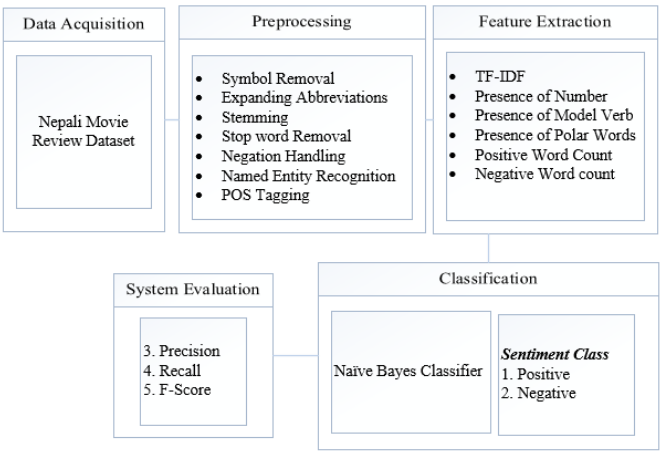


Fig. 1. System model of sentiment analysis.

more difficult since each word in the text is treated as one dimension. Here is the hypothesis of having the data properly pre-processed: to reduce the noise in the text should help improve the performance of the classifier and speed up the classification process, thus aiding in real time sentiment analysis. The main preprocessing techniques used are given below.

1) White space and special symbol removal.
2) Expanding abbreviations.
3) Stemming.
4) Stop word removal
5) Negation handling.
6) POS tagging.
7) Named entity recognition.

### A. Subjective/Objective classification

Pre-processed clean texts of movie review text document are classified into subjective (opinionated) and objective sentences. Objective sentences don't play any role in calculating the sentiment of the text since they do not consist of sentiment-bearing words or phrases. For example the following sentence in the review is not opinionated thus will not play a role in determining the sentiment: "The story was written by Dharabasi based on a story he found written in leaf booklet in a bag (Jhola) left at his home by an elderly man who had come from Manipur, India."

On the other hand, "It's intelligent, thought provoking, emotional, and damn well entertaining" definitely is opinionated/subjective and will be taken to the next level to determine whether the sentiment involved is positive or negative. So with this classification module we pruned all the not important sentences for future processing.

## IV. FEATURE EXTRACTION

Features in the context of opinion mining are the words, terms or phrases that strongly express the opinion as positive or negative. This means that they have a higher impact on the orientation of the text than other words in the same text. There are several methods that are used in feature selection, where some are syntactic, based on the syntactic position of the word such as objectives, and some are univariate, based on each feature's relation to a specific category, and some are multivariate using genetic algorithms and decision trees based on features subsets [21]. Main features extracted from the preprocessed document are described below.

### A. TF-IDF

TF-IDF feature represents weight of the particular term present in the text document. It reflects how important a word is to a document in a collection or corpus and every term are represented as a vector. Mathematically, TF-IDF weight can be calculated as,

$$W_{ik} = \frac{tf_{ik} log(\frac{N}{nk})}{\sum_{k=1}^{t}(tf_{ik})^2[\log(\frac{N}{nk})]^2} \quad (1)$$

Where,
$tf$ = Term frequency.
$idf$ = Inverse document frequency.
$T_k$ = Term $k$ in document $D_i$.
$tf_{ik}$ =frequency of term $T_k$ in document $D_i$.
$idf_k$ =Inverse document frequency of $T_k$ in document $C$.
$N$ = Total number of document in the collection $C$.
$n_k$ = The number of document in $C$ that contain $T_k$.
$idf_k = \log(\frac{n_k}{N})$

### B. Presence of Number

We extracted a binary feature based on the presence or absence of the number in the sentence. Sentences that contain numbers generally are objective sentences.

### C. Presence of Modal Verb

Presence of modal verbs like have to, must, can, should, wish, want, need play a major role in the classification of the subjective-objective sentences.

### D. Presence of Polar Words

Polar words are the words which represent the sentiment like good and bad. A binary feature is extracted on the presence or absence of the polar word. Sentences which contain polar words generally are subjective sentences. Example: Loot, the best movie of 2013!

### E. Positive Words Count

We calculated the number of positive words in the sentence and added it as a feature. This is a very important feature because if there are more positive words then the sentence tends to be a positive sentence. For example, "It's intelligent, thought provoking, emotional, and damn well entertaining" has four positive words so it is a positive sentence.

### F. Negative Words Count

We also calculated the number of negative words present in the sentence and added it as a feature. For example, "The story is elongated unnecessarily and it more boring than binding" has two negative words.

## V. Naive Bayesian Classifier

Naive Bayesian Classifier is a simple probabilistic classifier based on Bayes Theorem (Eq. 2) with strong independence assumptions of feature space. Depending on the precise nature of the probability model, Naive Bayes classifier can be trained very efficiently in a supervised learning setting.

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)} \tag{2}$$

Where,
$P(H|X)$is the posterior probability of $H$ conditioned on $X$.
$P(H)$ is the prior probability of hypothesis $H$.
$P(X|H)$is the posterior probability of $X$ conditioned on $H$.
$P(X)$is the prior probability of $X$.

For the given hypothesis (or text document) $H$, Naive Bayes classifier assign it to class $X^* = \arg\max_x P(H|X)$. Since $P(X)$ plays no role in selecting $X^*$. To estimate the term $P(X|H)$, Naive Bayes decomposes it by assuming the features $F_i$'s (Section IV) are conditionally independent given $X$'s class:

$$P(X|H) = \frac{P(X)\prod_{i=1}^{n} P(F_i|X)}{P(H)} \tag{3}$$

The training method performs the relative frequency estimation of $P(X)$ and $P(F_i|X)$.

## VI. Evaluation Methods

The correctness of a classification can be evaluated by computing the number of correctly recognized class examples (true positives), the number of correctly recognized examples that do not belong to the class (true negatives), and examples that either were incorrectly assigned to the class (false positives) or that were not recognized as class examples (false negatives).

### A. Precision

Precision (also called positive predictive value) is the number of correctly classified positive examples divided by the number of examples labeled by the system as positive.

### B. Recall

Recall(also called sensitivity)is the number of correctly classified positive examples divided by the number of positive examples in the test dataset.

### C. F-Score

F-Score is the combination of the precision and recall.

$$Fscore = \frac{(\beta^2+1)Precision * Recall}{\beta^2 * Precision + Recall} \tag{4}$$

## VII. Training and Testing Dataset

Nepali movie review dataset contains total 500 text documents which are collected from [22]–[24] and other online sources. Documents are manually annotated to positive or negative sentiments. Dataset is than divided into two classes, "positive" and "negative" for training and testing with 250 samples each class.

## VIII. Experiments and Results

Traning and Testing dataset statistics is given in Table I. Classification accuracies are given in Table II.

TABLE I
Nepali Movie Review Dataset Statistics.

| | Training | Testing | Total |
|---|---|---|---|
| **Positive** | 200 | 50 | **250** |
| **Negative** | 200 | 50 | **250** |
| **Total** | **400** | **100** | **500** |

TABLE II
Sentiment Classification Accuracies.

| Precision (%) | Recall (%) | F-score (%) |
|---|---|---|
| 79.23 | 78.57 | 78.9 |

## IX. Conclusion and Feature Work

Sentiment analysis is a challenging field of data mining as it involves natural language processing, text analysis and computational linguistics. It has a wide variety of applications that could benefit from its results, such as review analytics, news analytics, marketing, question answering, knowledge bases and so on. Knowing important insights from opinions expressed for certain products, news, or documents by users on the internet is lively for many companies and institutions, whether it is in terms of product feedback, public mood, or investors opinions.

In this research paper, Sentiment analysis is done for Nepali movie reviews found online. Various natural language processing techniques are used for document preprocessing and for semantic feature extraction. Naive Bayes based classifier is used as a sentiment classifier.

We have created a Nepali movie review dataset with total 500 samples having 250 samples per each positive and negative class of sentiment. Empirical results shows, classification accuracies are obtained as, $79.23\%$ of precision, $78.57\%$ of recall and $78.90\%$ of F-score.

Sentiment analysis system can be further enhanced by adding more training and testing datasets. System can also be tested by adding more features and selecting good features. Other machine learning techniques like SVM, Neural networks, Fuzzy Logic can also be applied to the problem of sentiment analysis of movie reviews.

## References


[1] R. Feldman, "Techniques and applications for sentiment analysis," *Commun. ACM*, vol. 56, no. 4, pp. 82–89, Apr. 2013.

[2] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? sentiment classification using machine learning techniques," *CoRR*, vol. cs.CL/0205070, 2002.

[3] P. D. Turney and M. L. Littman, "Measuring praise and criticism: Inference of semantic orientation from association," *CoRR*, vol. cs.CL/0309034, 2003.

[4] B. Pang and L. Lee, "A sentimental education: Sentiment analysis using subjectivity," in *Proceedings of ACL*, 2004, pp. 271–278.

[5] P. Beineke, T. Hastie, and S. Vaithyanathan, "The sentimental factor: Improving review classification via human-provided information," in *ACL*, D. Scott, W. Daelemans, and M. A. Walker, Eds. ACL, 2004, pp. 263–270.
[6] M. Gamon, "Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis," in *COLING*, 2004.
[7] M. Zitnik, "Using sentiment analysis to improve business operations," *ACM Crossroads*, vol. 18, no. 4, pp. 42–43, 2012.
[8] T. Mullen and N. Collier, "Sentiment analysis using support vector machines with diverse information sources," in *EMNLP*. ACL, 2004, pp. 412–418.
[9] P. D. Turney and M. L. Littman, "Unsupervised learning of semantic orientation from a hundred-billion-word corpus," Tech. Rep., 2002.
[10] L. Dong, F. Wei, S. Liu, M. Zhou, and K. Xu, "A statistical parsing framework for sentiment classification," Jan. 24 2014.
[11] M. A. Haque, "Sentiment analysis by using fuzzy logic," Mar. 13 2014.
[12] W. Fan, S. Sun, and G. Song, "Probability adjustment naive bayes algorithm based on nondomain-specific sentiment and evaluation word for domain-transfer sentiment analysis," in *FSKD*. IEEE, 2011, pp. 1043–1046.
[13] V. Narayanan, I. Arora, and A. Bhatia, "Fast and accurate sentiment classification using an enhanced naive bayes model," pp. 194–201, Sep. 16 2013, comment: 8 pages, 2 figures.
[14] N. R. Hassan and S. Q. Wong, "The relationship between market sentiment and equity premium: an artificial neural network analysis," 2008.
[15] A. Sharma and S. Dey, "Using self-organizing maps for sentiment analysis," Sep. 16 2013, comment: 13 Pages.
[16] B. Jayanag, K. Vineela, and S. Vasavi, "A study on feature subsumption for sentiment classification in social networks using natural language processing," no. 18, 2012.
[17] P. Chaovalit and L. Zhou, "Movie review mining: a comparison between supervised and unsupervised classification approaches," in *Proceedings of HICSS-05, the 38th Hawaii International Conference on System Sciences*, 2005, p. 112c.
[18] H. Ghorbel and D. Jacot, "Sentiment analysis of french movie reviews," in *Advances in Distributed Agent-Based Retrieval Tools*, ser. Studies in Computational Intelligence, V. Pallotta, A. Soro, and E. Vargiu, Eds. Springer, 2011, vol. 361, pp. 97–108.
[19] P. Kalaivani and D. K. L. Shunmuganathan, "Sentiment classification of movie reviews by supervised machine learning approaches," 2013.
[20] S. M. Basheer and S. Farook, "Movie review classification and feature based summarization of movie reviews," p. 167, 2013.
[21] E. Haddi, X. Liu, and Y. Shi, "The role of text pre-processing in sentiment analysis," in *ITQM*, ser. Procedia Computer Science, Y. Shi, Y. Xi, P. Wolcott, Y. Tian, J. Li, D. Berg, Z. Chen, E. Herrera-Viedma, G. Kou, H. Lee, Y. Peng, and L. Yu, Eds., vol. 17. Elsevier, 2013, pp. 26–32.
[22] [Online]. Available: http://xnepali.net/movies/category/movie-review/
[23] [Online]. Available: http://www.merocinema.com
[24] [Online]. Available: http://www.youtube.com/