Sentiment Analysis Of Nepali Sentences Using Naive Bayes Classification

Submitted by:
Mahesh Acharya (7942/072)
Ashok Chhetri (7926/072)
Abhishek Sapkota (7921/072)
Rabin Bhandari (7950/072)
Batch (2016 - 2020)

Submitted to:
Department of Computer Science and Information Technology
Tribhuvan University
Kirtipur, Nepal

Under the supervision of
Prof. Bikash Balami (IOST, TU)

.

1. Introduction

Sentiment Analysis is contextual mining of text in Negative, Positive or Neutral which identifies and extracts subjective information in source material, and helps a business to understand the social sentiment of their brand, product or service while monitoring online conversations. Proper identification and classification of SENTIMENT ANALYSIS are very crucial and pose a very big challenge to the NLP researchers. Sentiment analysis is difficult and challenging for Indo-Aryan language like Nepali due to lack of resources[1].

Now- a- days, Machine-Learning (ML) approaches are popularly used in sentiment analysis because these are easily trainable, adaptable to different domains and languages as well as their maintenance are also less expensive. On the other hand, rule-based approaches lack the ability of coping with the problems of robustness and portability. Each new source of text requires significant tweaking of rules to maintain optimal performance and the maintenance costs could be quite high. Some of the well-known machine learning approaches used in sentiment analysis are SVM and Naïve Bayes [2].

2. Problem Statement

Sentiment analysis is in research phase in Aryan Indo Language; Nepali Language. Since in the Nepali e-commerce the review is in Nepali or Roman language and the sentiment of the people is not exactly addressed by the e-commerce sites to better understand the opinions and reviews of the customers. Efficient analysis of the sentiment of the people reviews on the product can better help companies and sellers to make changes to their products and improvise according to the demands and desires of the people.

The sentiment analysis in Nepali is the problem which asks for the classification of each word of a document into predefined target of classes. In this research work, problem of Nepali Sentiments is addressed. The recognition task is carried out with supervised machine learning using Naive Bayes [9].

Given a set of classes, the sentences and reviews are categorized into Negative, Positive or Neutral Categories. For example,

मलाई यो घडी मन परेन । **[Negative]**

आहा कती राम्रो घडी । **[Positive]**

The main feature for the Sentiment Analysis is identifying customers opinion towards the products.

3. Objectives

The main objective of the Sentiment Analysis System is:

1. To classify/analyze the input opinions of the customers.

4. Literature Review

There are lots of researches which have been done in the field of Sentiment Analysis, but there is no any work in Nepali Sentiment Analysis.

The Naïve Bayes [6] method, is the classification technique that works in the conditional probability which means that the likelihood of occurrence of the next event given that one event has already occurred. The Naïve Bayes [6] is basically binary classification but it can be extended to multiclass classification.

In [2]"Semantic Orientation Applied to Unsupervised Classification of Reviews" Author Peter D. Turney. This system used unsupervised machine learning technique for the classification of the reviews.

In [3]"Semantic Analysis on Nepali Movie Reviews" Author Ashok Panta Naive Bayes Sentiment Analysis[3] This system uses Naive Bayes algorithm under the supervised machine learning to classify the Nepali movie reviews.

5. Scope and Limitation

The Scope of the Sentiment Analysis in Nepali reviews are as follows:

1. It is used to find the opinion of the customers on the newly lunched products.
2. It is used in the film industry to analysis the comments and reviews of the movies and songs.

3. It is used in News Portals to find out the public opinions and reviews.
4. It is used in Politics to know the political perceptions and views of the people.

The Limitation of the Sentiment Analysis In Nepali reviews are as follows:

1. The accuracy of the reviews is low when there is the mixture of Nepali as well as English language in comments.
2. Ambiguity can occur and semantic analysis is difficult to address.

6. Methodology

The research method for this module is discussed in this chapter. The different sections are:

6.1 Data Collection

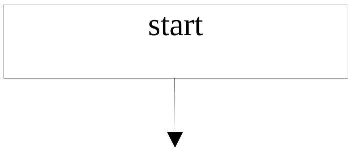
The sentences contain the Nepali reviews texts are collected from e-commerce, Nepali newspapers as well as from the blogs of different domains. Sentiment analysis sentences will be created manually.

6.2 Preprocessing

Preprocessing is usually done manually by the humans in regard to the data sets. Here supervised machine learning technique is applied with the already gathered data sets and the result is to be expected as per the domains. The main problem of preprocessing is time, if there are large data sets, processing time goes exponentially long. So to reduce time some techniques of segmentation is to be used.

a. Proposed Implementation Model for SA

As an implementation model, we will be implementing the Naive Bayes algorithm so as to extract the sentiment score from the given input set of data. The algorithm will be simulated for various size of sentences. The implementation model is given below:



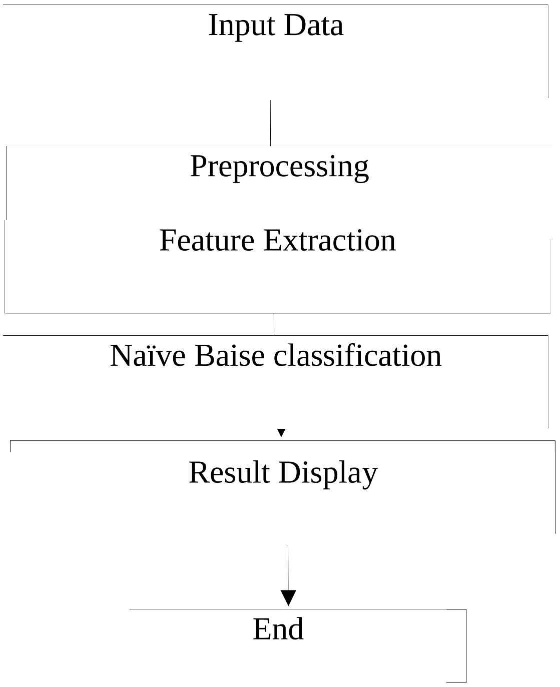


Figure 6.1 Implementation Model for Sentiment Analysis

6.4 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.

Main features extracted from the preprocessed document are described below.

6.4.1 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,

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

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.
$idf_k = \log(\frac{n_k}{N})$

6.4.2 Presence Of Polar Word

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: यो समान राम्रो छ ।

6.4.3 Count Of positive word

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, ”यो समान राम्रो छ तेसैले मेरो छोराले यो धेरै मन परायो । ″ has four positive words so it is a positive sentence.

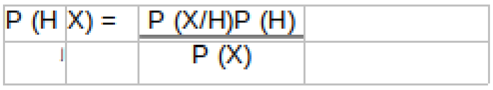
6.4.4 Count Of Negative word

We also calculated the number of negative words present in the sentence and added it as a feature. For example, ”नेपालमा अन्लाईन समानमा धेरै ठगी हुने हुँदा मान्छेहरु सामान किन्न मन पराउदैनन । ”

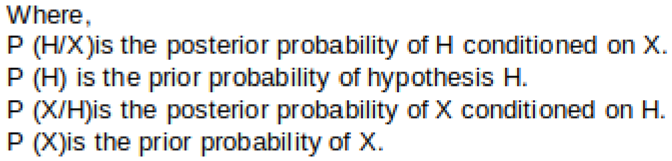
has two negative words.

6.5 Naive Bays

Naive Bayesian Classifier is a simple probabilistic classifier based on Bayes Theorem 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[3].



P (H	X) =	P (X/H)P (H)	
|		P (X)	



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.

6.5.1 Classify Naive Bayes Text

position = all word in Doc that contain tokens found in Vocabulary

Return V_{NB}, Where

$$V_{NB} = \underset{vj \varepsilon V}{\mathrm{argmax}} P(v_j) \prod_{i \varepsilon positions} P(ai/vj)$$

6.5.2 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.

6.5.3 Recall

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

6.5.4 F-Measure

Harmonic mean of precision and recall. Mathematically,

$$F = \frac{2(P*R)}{P+R}$$

7. Project Schedule

Year	2019								
Activity	1 Week	2 Week	3 Week	2 Week	3 Week	2 Week	2 Week	1 Week	2 Week
Collection of Literature	■								
Study of Literature		■							
Analysis of proposed Scheme			■						
Preparation And Abstract Submission				■					
Preparation Of Model				■	■				
Implementation and Debugging					■	■			
Analysis and Simulation						■	■		
Result Formulation								■	
Final Write-up& Submission									■

1. **8. EXPECTED OUTCOME**

At the end of this project, the proposed model is to be expected to classify general customer opinion towards the product in Nepali text.

9. References

[1] Surya Bahadur Bam, “Named Entity Recognition for Nepali Text using Support Vector Machine,” Tribhuwan University Department Of Science and Technology, 2011.

[2] Surya Bahadur Bam, “Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews,” Institute for Information Technology National Research Council of Canada Ottawa, Ontario, Canada, K1A 0R6, 2002.

[3] Surya Bahadur Bam, “Sentiment Analysis on Nepali Movie Reviews using Machine Learning,” Tribhuwan University Department Of Science and Technology, 2013.