Sentiment Analysis of Nepali Sentences Using Naive Bayes Classification

Ashok Chhetri (7926/072)
Abhishek Sapkota (7921/072)
Mahesh Acharya (7942/072)
Rabin Bhandari (7950/072)

Asian School of Management and Technology

Under the supervision of
Mr. Bikash Balami

Table of content

Introduction

- Sentiment Analysis is contextual mining of text
- It specifies the polarity(positive, negative, and neutral) of the sentences
- Sentiment analysis is difficult and challenging for Indo-Aryan language in which machine learning approaches are used in sentiment analysis by analyzing huge amount of data
- There are different algorithm are used in classification of sentiment like SVM and naïve bayes
- Among them we use naïve bays classification algorithm

Problem statement

Sentiment analysis is in research phase in Aryan Indo Language; Nepali Language. Due to lack of resources sentiment analysis is difficult and challenging for our language

In this project work , problem of nepali sentiment is addressed

The recognization task is carried out by supervised machine learning naïve bayes

.eg

मलाई यो मन परेन।[Negative]

आहा कती राम्रो घडी।[Positive]

Objectives

The main objective of the Sentiment Analysis System is:

. To classify the opinions of the customers for e-commerce based portal.

Literature review

There are lots of researches which have been done in the field of Sentiment Analysis,

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

In [1] "Named Entity Recognition for Nepali Text using Support Vector Machine," Author Surya Bahadur Bam. This system used supervised machine learning to classify Nepali phrases.

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.

Scope and limitations

Scopes

. It is used to find the opinion of the customers on the newly lunched products.

. It is used in the film industry to analysis the comments and reviews of the movies and songs.

. It is used in News Portals to find out the public opinions and reviews.

Limitations

. The accuracy of the reviews is low when there is the mixture of Nepali as well as English language in comments.

. Ambiguity can occur, and semantic analysis is difficult to address.

Methodology

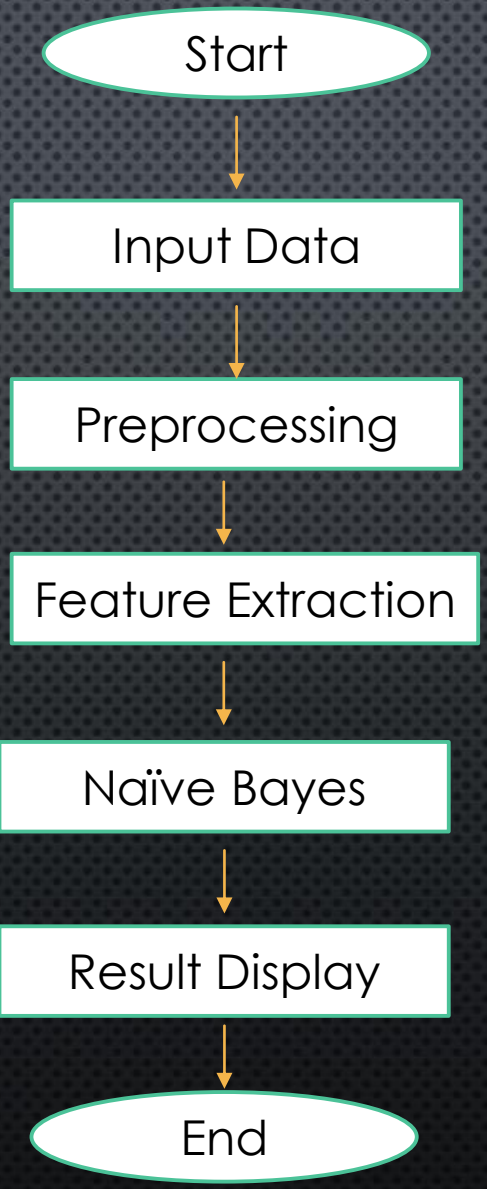
Data collection

We are manually collect data from different new and ecommerce sites of Nepal.

Preprocessing

Preprocessing is usually done manually by the humans in regard to the data sets. Here supervised machine

Proposed Implementation Model for Sentiment Analysis



Feature extraction

Features in the context of opinion mining are the words, terms or phrases that strongly express the opinion as positive or negative

TF-IDF

feature represents weight of the particular term present in the text document.

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

Where,

- tf = Term frequency,
- idf = Inverse document frequency,
- T_k = Term k in document D_i,
- n = Total number of document in collection of C,
- n_k= The number of document in C that contain T-k,
- idf_k= Inverse document frequency of T_k in document ,
- idf_k= $\log(\frac{nk}{N})$

Presence of Polar Word

Polar words are the words which represent the sentiment like good and bad

. Example: "यो सामान राम्रो छ।"

Count of Positive Word

We calculated the number of positive words in the sentence and added it as a feature

For example, "यो सामान राम्रो छ तेसैले मेरो छोराले धेरै मन परायो।"

has four positive words so it is a positive sentence.

Count of Negative word

We also calculated the number of negative words present in the sentence and added it as a feature.

E.g. "नेपालमा अन्लाईन सामानमा धेरै ठगि हुने हुँदा मान्छेहरु समान किन्न मन पराउनन्।"

has two negative words.

Naive Bayes

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[4].

$$P\left({}^{H}/_{X}\right) = \frac{P({}^{X}/_{H})P(H)}{P(X)}$$

Where,

$P({}^{X}/_{H})$ = posterior probability of H conditioned X

$P({}^{H}/_{X})$ = posterior probability of X conditioned H

$P(H)$ = prior probability of hypothesis H

$P(X)$ = prior probability of X

Classify Naive Bayes Text

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

Return VNB ,

Where,

$$VNB = \underset{v_j \in V}{\operatorname{argmax}} P(v_j) \prod_{i \in positions} P(a_i / v_j)$$

.eg

test = मेरो नाम राम्रो छ ।

P(test/+) = P(+)P(मेरो/+)P(नाम/+)P(राम्रो/+)P(छ/+)

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.

Recall

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

F-Measure

Harmonic mean of precision and recall. Mathematically,

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

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 propose scheme									
Preperation and abstract submission									
Preperation of model									
Implementation and debugging									
Analysis and simulation									
Result foermulation									
Final weite-up and submissiom									

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.

References

[1]Surya Bahadur Bam, "Named Entity Recognition for Nepali Text using Support VectoMachine," Tribhuvan University Department of Science and Technology, 2011.

[2]Peter D. Turney, "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]Ashok Panta, "Sentiment Analysis on Nepali Movie Reviews using Machine Learning," Tribhuwan University Department of Science and Technology, 2013.

[4]Lina L. Dhande and Dr. Prof. Girish K. Patnaik, —Analyzing Sentiment of Movie Review Data using Naive Bayes Neural Classifier‖, IJETTCS, Volume 3, Issue 4 July-August 2014, ISSN 2278-6856.

[5]Rajaraman, A., & Ullman, J. (2011). Data Mining. In Mining of Massive Datasets (pp. 1-17). Cambridge: Cambridge University Press.

THANK YOU!