# PURBANCHAL UNIVERSITY

# DEPARTMENT OF COMPUTER ENGINEERING

# KHWOPA ENGINEERING COLLEGE
# LIBALI-2, BHAKTAPUR

**A PROJECT PROPOSAL**

**ON**

# SENTIMENT ANALYSIS

Project work submitted in partial fulfillment of requirements for the award of the degree of Bachelor of Engineering in Computer Engineering (Seventh Semester)

**SUBMITTED BY**

1. Arun Prajapati (740305)
2. Neetu Phaiju (740324)
3. Rabin Phaiju (740329)
4. Rodip Duwal (740334)

**SUBMITTED TO:**

DEPARTMENT OF COMPUTER ENGINEERING

KHWOPA ENGINEERING COLLEGE

2nd July 2021

# ABSTRACT

Sentiment analysis is the computational study of people's opinions, sentiments, attitudes, and emotions expressed in written language. It is one of the most active research areas in natural language processing and text mining in recent years. It has a wide range of applications because opinions are key influencers of our behavior. In decision making, the opinions of others have a significant effect. The approaches of text sentiment analysis typically work at a particular level like phrase, sentence or document level. This system aims at analyzing a solution for the sentiment classification. In this project, we are going to compare Naïve Bayes and Support Vector Machine on sentiment analysis. We will use twitter posts mainly focused on political statements of our country to find polarity of the posts.

# Table of Contents

# List of Figures

# List of Abbreviation

1. SA: Sentiment Analysis
2. NLP: Natural Language Processing
3. SVM: Support Vector Machine

# CHAPTER 1

## INTRODUCTION

### 1.1 Background

Sentiment analysis (or opinion mining) is a natural language processing technique used to determine whether data is positive, negative or neutral. Sentiment analysis is often performed on textual data to help businesses monitor brand and product sentiment in customer feedback, and understand customer needs.[4]

There are many research works that focus on sentiment classification, with different goals and different supervised and unsupervised methods. There are usually based on machine learning models (e.g., support vector machines, maximum entropy, time series, Bayes, Hoeffding trees, artificial neural networks, etc.). Some of them use lexical resources and different features (e.g., unigrams, bigrams, etc.). The accuracy of such methods is incrementally better, being their main goal, the extraction of the sentiment based on the subjectivity and linguistic features of the words used in an unstructured text. Thus, it is quite common the division into two well differentiated types of methods for SA:

- • Lexicon based methods - These techniques are based on dictionaries of words annotated with their semantic polarity [2].

- • Machine learning based methods - These methods are also divided in groups: supervised and unsupervised techniques [3], together with semi-supervised learning, a hybrid between them.

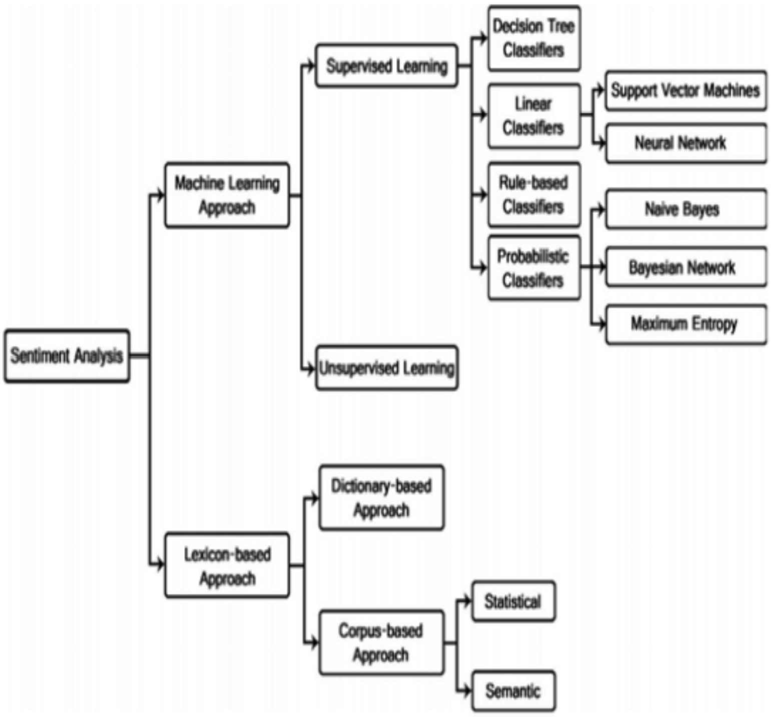


Fig 1.1, Sentiment Classification

In this paper proposes a model to analyses in comments, posts from twitter, Facebook and/or other social media. One of the main tasks to do is the preparation and the process of the corpus as well as the selection of the features that finally will support the process of classification of the comments in positive, neutral, and negative.

### 1.2 Motivation

Due to the exponential growth of the social network, the sentiment analysis has been applied to analyze the user's opinions. The majority of the data that is constantly generated in the social networks could contain valuable information like perceptions and tendencies from the users to the objects, personalities, or services.

Due to its popularity nowadays, SA has been expanded to other fields like education, medicine, politics, and others. But reading all reviews is both time and money consuming therefore instead of spending times in reading and figuring out the positivity and negativity of text we can use automated techniques for sentiment analysis.

### 1.3 Statement of Problems

A major benefit of social media is that we can see the good and bad things people say about the particular brand or personality.

The bigger your company gets difficult it becomes to keep a handle on how everyone feels about your brand. For large companies or a public figure with thousands of daily mentions on social media, news sites and blogs, it's extremely difficult to do this manually.

To combat this problem, sentimental analysis is necessary. It can be used to evaluate the people's sentiment about particular brand or personality.

### 1.4 Objectives

- To obtain writer's feelings/emotions/sentiments expressed in positive or negative comments from plain text.

### 1.5 Scope & Limitation

Sentiment analysis is the area which deals with judgments, responses as well as feelings, which is generated from texts, being extensively used in fields like data mining, web mining, and social media analytics because sentiments are the most essential characteristics to judge the human behavior [5]. Hence this project can be a significant measure for predicting polarity of political statements of our country based on twitter's posts.

In context of existing techniques, there are inadequate accuracy, incapability to deal with complex sentences and inability to perform well in different domains.

# CHAPTER 2

## LITERATURE REVIEW

As we know the success of the company or product depends on their customers. So, the customer likes your product it's your success if not then you certainly need to improvise it by making some changes in it. But how will you know whether your product is successful or not. For that you need to analyze your customers and one of the attributes of analyzing your customer is to analyze the sentiment of them and this is where the sentiment analysis comes into picture.

Sentiment analysis is the process of computationally identifying and categorizing opinions from piece of text, and determine whether the writer's attitude towards a particular topic or the product, is positive, negative, or neutral. Finding the sentiments of a person (also called the polarization of a text) is a classical NLP problem. This problem brings a relatively new perspective by considering twitter tweets and Facebook posts, which are effectively a dialect of the English language. This problem has been considered already by a few authors. Since it is a necessary step but is not central to our questions, we plan to reuse some filters already available. If these detectors do not work as well as expected, we plan to use a robust binary classifier such as SVM. Furthermore, instead of using directly the words as features, we can use a Kernelized SVM that takes into account the spelling mistakes. Posts will be filtered by hashtags. For sentiment analysis, we will first use naive Bayes classifiers on a bag-of-words model, potentially with n-grams and stop words (uh. uh…. uh… a, the, am, is) are removed. If training is required, that could be done on Twitter tweets and Facebook posts or recorded sound and text.

We will compare most popular algorithms such as Naïve Bayes (Kononenko, 1993), Support Vector Machine (Cortes and Vapnik,1995) on sentiment analysis. We will use twitter posts mainly focused on our country's political statements to find polarity of the posts.

In 2011, Han-Xiao Shi, Xiao-Jun Li created a sentiment analysis model for hotel reviews based on supervised learning approach using unigram feature with two types of information (frequency and TF-IDF) to realize polarity classification of documents [1].

In 2019, Saad and Yang have aimed for giving a complete tweet sentiment analysis on the basis of ordinal regression with machine learning algorithms. The suggested model included pre-processing tweets as first step and with the feature extraction model, an effective feature was generated [6].

# CHAPTER 3

## PROJECT MANAGEMENT

In order to design our system, first we collect the related information. Then we will plan for the success of the concerning concept. The main contribution of this work will be as following:

- Compare Naïve Bayes, Support Vector Machine based on accuracy, precision, Recall and F1 score.
- Evaluate the impact of such political statements on final Polarity review.

### 3.1 Team Members

For this project, we have a group of four members:

1. Arun Prajapati (740305)
2. Neetu Phaiju (740324)
3. Rabin Phaiju (740329)
4. Rodip Duwal (740334)

### 3.2 Feasibility Study

The aim of feasibility study is to understand thoroughly all aspects of a concept, or plan. During the study, problems in the system are determined. So, it's always good to have a contingency plan that test to make sure it's a viable alternative in case the first plan fails. This study will be reliable for forecasting the detail of the system.

### 3.3 Work break down structure

All the team members will work on different modules. During the course of work, team members will be in touch so that no problems will arise in future. After the completion of individual work, whole work will be combined to develop a proposed system.

| S. N | Week Job Description | 1st Week | 2nd Week | 3rd Week | 4th Week | 5th Week | 6th Week | 7th Week | 8th Week |
|---|---|---|---|---|---|---|---|---|---|
| 1. | Problem Identification | ■ | | | | | | | |
| 2. | Analysis | | ■ | | | | | | |
| 3. | Design | | | ■ | ■ | ■ | ■ | ■ | |
| 4. | Coding | | | | | ■ | ■ | ■ | ■ |
| 5. | Testing and debugging | | | | | | ■ | ■ | ■ |
| 6. | Documentati on | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ |

Fig 3.1, Workbreak down structure of our project

# CHAPTER 4

## METHODOLOGY

### 4.1 Background

We will compare most popular algorithms such as Naïve Bayes (Kononenko, 1993), Support Vector Machine (Cortes and Vapnik,1995) on sentiment analysis. We will use twitter posts mainly focused on our country's political statements to find polarity of the posts.

The step involved to develop our system are described as follows:

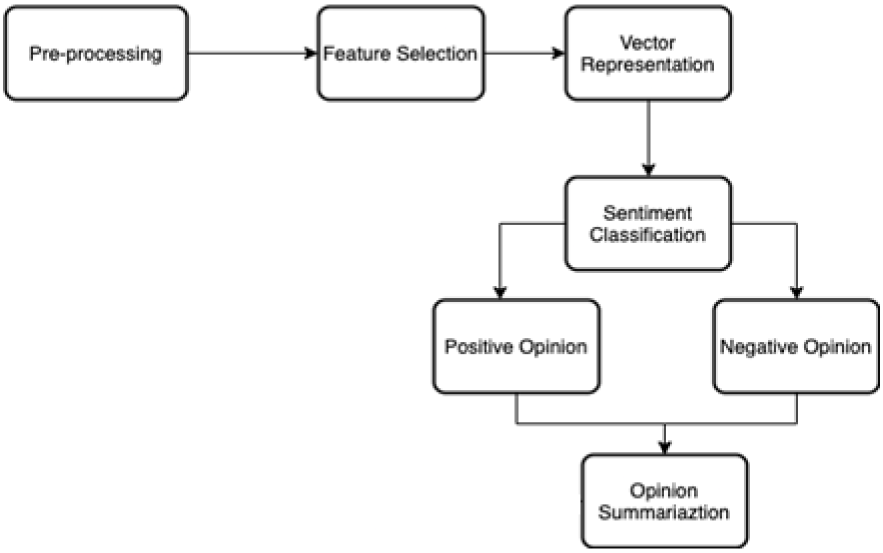


Fig 4.1, The proposed methodology for sentiment analysis.

a. **Preprocessing**: In Preprocessing the unanalyzed data is handled for feature extraction. It is further divided into below step:

- Tokenization: White spaces, symbols and special characters are removed and a sentence is divided into words.
- Stop Word Removal: Articles are removed.
- Stemming: Token or words are reduced for root forms.

b. **Feature extraction:** Feature extraction handles the following task:

- Feature Type: In this step features are identified like the term frequencies, term cooccurrences, Opinion word, OS information, Negation Syntactic Dependencies.
- Selection of Feature: Good features are selected for classification using the following ways like Information gain, Document frequency, Odd ratio and Mutual Information.
- Feature Weighting Mechanism: The features are ranked by computing the weight using term presence, term frequency and Inverse document frequencies.
- Reduction of Feature: To optimize the classifiers performance the vector size is reduced.

c. **Sentiment Analysis:** Polarity of text is classified by Sentiment Analysis. This process is done in 3 different levels.

- Document Level: The entire document is taken and is labeled as either positive or negative.
- Sentence Level: The entire document is parsed into sentence and the polarity is classified as positive, neutral or negative.
- Word or Phrase Level: Product attributes or components are analyzed.

d. **Sentiment Classification:** Sentiment classification uses two approaches to classify the nature of documents/sentence. They are Machine Learning Approaches and Lexicon Based Approaches.

- The Machine Learning belongs to supervised learning and classification of text in particular. So, it is called as ―Supervised Learning.
- Lexicon Based approach consist of Dictionary based and corpus based.

e. **Visualization**: An application software and different plots to show the frequency of words in the customer tweets and the sentiment scores.

### 4.2 Naïve Bayes

It is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.

Bayes theorem provides a way of calculating posterior probability P(c|x) from P(c), P(x) and P(x|c). Look at the equation below:

$$\underbrace{P(c\mid x)}_{\text{Posterior Probability}} = \frac{\overbrace{P(x\mid c)}^{\text{Likelihood}}\,\overbrace{P(c)}^{\text{Class Prior Probability}}}{\underbrace{P(x)}_{\text{Predictor Prior Probability}}}$$

$$P(c\mid \mathrm{X}) = P(x_1\mid c)\times P(x_2\mid c)\times\cdots\times P(x_n\mid c)\times P(c)$$

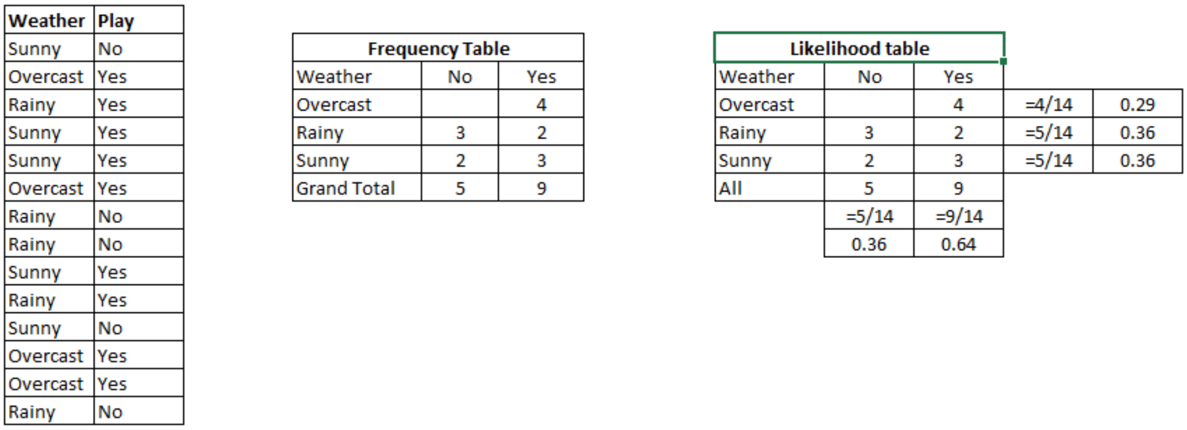
Above,

- P(c|x) is the posterior probability of class (c, target) given predictor (x, attributes).
- P(c) is the prior probability of class.
- P(x|c) is the likelihood which is the probability of predictor given class.
- P(x) is the prior probability of predictor.

Let's understand it using an example. Below I have a training data set of weather and corresponding target variable 'Play' (suggesting possibilities of playing). Now, we need to classify whether players will play or not based on weather condition. Let's follow the below steps to perform it.

Step 1: Convert the data set into a frequency table

Step 2: Create Likelihood table by finding the probabilities like Overcast probability = 0.29 and probability of playing is 0.64.



| Weather | Play |
|---|---|
| Sunny | No |
| Overcast | Yes |
| Rainy | Yes |
| Sunny | Yes |
| Sunny | Yes |
| Overcast | Yes |
| Rainy | No |
| Rainy | No |
| Sunny | Yes |
| Rainy | Yes |
| Sunny | No |
| Overcast | Yes |
| Overcast | Yes |
| Rainy | No |

| Frequency Table | | |
|---|---|---|
| Weather | No | Yes |
| Overcast | | 4 |
| Rainy | 3 | 2 |
| Sunny | 2 | 3 |
| Grand Total | 5 | 9 |

| Likelihood table | | | | |
|---|---|---|---|---|
| Weather | No | Yes | | |
| Overcast | | 4 | =4/14 | 0.29 |
| Rainy | 3 | 2 | =5/14 | 0.36 |
| Sunny | 2 | 3 | =5/14 | 0.36 |
| All | 5 | 9 | | |
| | =5/14 | =9/14 | | |
| | 0.36 | 0.64 | | |

Fig 4.2. Likelihood tables

Step 3: Now, use Naive Bayesian equation to calculate the posterior probability for each class. The class with the highest posterior probability is the outcome of prediction.

**Problem:** Players will play if weather is sunny. Is this statement is correct?

We can solve it using above discussed method of posterior probability.

P(Yes | Sunny) = P( Sunny | Yes) * P(Yes) / P (Sunny)

Here we have P (Sunny |Yes) = 3/9 = 0.33, P(Sunny) = 5/14 = 0.36, P( Yes)= 9/14 = 0.64

Now, P (Yes | Sunny) = 0.33 * 0.64 / 0.36 = 0.60, which has higher probability.

Naive Bayes uses a similar method to predict the probability of different class based on various attributes. This algorithm is mostly used in text classification and with problems having multiple classes.

### 4.3 Support Vector Machine

Support Vector Machine" (SVM) is a supervised machine learning algorithm which can be used for both classification or regression challenges. However, it is mostly used in classification problems. In the SVM algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiates the two classes very well (look at the below snapshot)

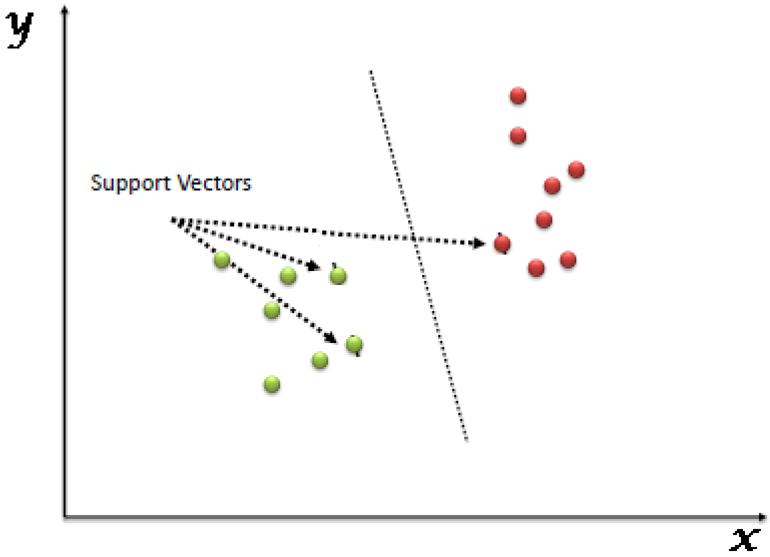


Fig 4.3, SVM classifier

## 4.4 Use case diagram

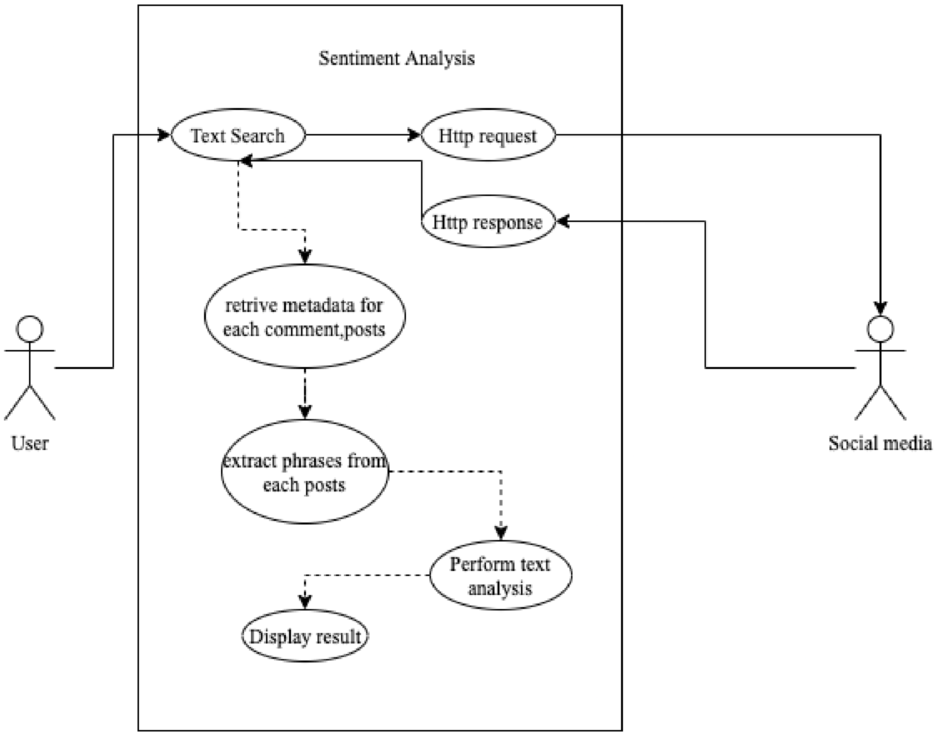


Fig 4.4, Use case diagram for sentiment analysis

## 4.5 Tools and Platform

1. Python
2. Google Colab
3. Jupyter Notebook
4. VS code

# CHAPTER 5

## EXPECTED OUTCOMES

The project will comprise two parts: a classifier of text snippets into topic modeling and sentiment analysis. This will evaluate the sentiment of a person and analyze it. It will show positive and negative sentiment of a person. This system will be executed successfully and will meet its objective. This will be a reliable approach for predicting people’s opinions.

# REFERENCES

### Paper in Conference

[1] Han-Xiao Shi, Xiao-Jun Li. A sentiment analysis model for hotel reviews based on supervised learning. (10-13 July 2011)

[2] M. Taboada, J. Brooke, M. Tofiloski, K. Voll and M. Stede, Lexicon-based methods for sentiment analysis, Computational linguistics 37(2) (2011)

[3] D. Vilares, C. G´omez-Rodrıguez and M.A. Alonso, Universal, unsupervised (rule-based), uncovered sentiment analysis, Knowledge-Based Systems 118 (2017)

### Website

[4] Sentiment Analysis: A definitive Guide, Retrieved from: monkeylearn.com/sentiment-analysis/

[5] Sentiment Analysis: ScienceDirect, Retrieved from: sciencedirect.com/topics/engineering/sentiment-analysis

[6] S. E. Saad and J. Yang, "Twitter Sentiment Analysis Based on Ordinal Regression," IEEE Access, vol. 7, pp. 163677-163685, (2019).