SENTIMENT ANALYSIS

Team Members:

Arun Teja Dutta(999991302)

Sukanya Neela(999991566)

**DATA MINING:**

Introduction: Data mining is a technique of extraction hidden patterns from large data bases.

**Sentiment Analysis**

**Project outline:**

The outline of the project is sentiment identification system using Machine learning technique to classify the polarity of the movie review as positive, negative.

**Problem:**

Problem is to find the polarities (sentiments) based on the reviews of the users, whether they are positive or negative (good or bad).

## Data:

Data is obtained from tweets. The data in tweets is raw and contains unwanted numeric and special characters, they are removed in the code before performing the analysis.



#### NAÏVE BAYES TECHNIQUE

Bayes theorem provides a way of calculating the posterior probability, *P*(*c|x*), from *P*(*c*), *P*(*x*), and *P*(*x|c*). Naive Bayes classifier assume that the effect of the value of a predictor (*x*) on a given class (*c*) is independent of the values of other predictors. This assumption is called as class conditional independence.

#### Naïve Bayes rule

P(c|x)=P(x1|c) P(x2|c) ….. P(xn|c) P(c)

* + - * *P*(*c|x*) is the posterior probability of *class* (*target*) given *predictor*(*attribute*).
      * *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*.

#### Naïve Bayes Conditional Independence Assumption

Assume that the probability of observing the conjunction of attributes is equal to the product of the individual probability P(Xi|Cj).

W: Words C: Class

Tweet

W2

W3

W4

W1

**Naïve Bayes Conditional Independence**

Features detect term presence and are independent of each other given the class P(W1,…….W4|C) =P(W1|C)\*P(W2|C)\*…….P(X4|C)

This model is appropriate for binary variables. Zero probabilities cannot conditioned away, no matter the other evidence

Here Tweet representing the sentence and w1, w2, w3, w4 are representing words. Probability of the tweet is equal to the product of the probabilities of the words. Probability of each word in the each sentence is found using the posterior probability.

Thus each sentence probability will be found accordingly with the multiplication of each word probability.

## SYSTEM IMPLEMENTATION

**Architecture**

Reviews, Tweets

Positive Sentence

Negative Sentence

Sentiment (Naive Bayes) classifier

Positive Polarity

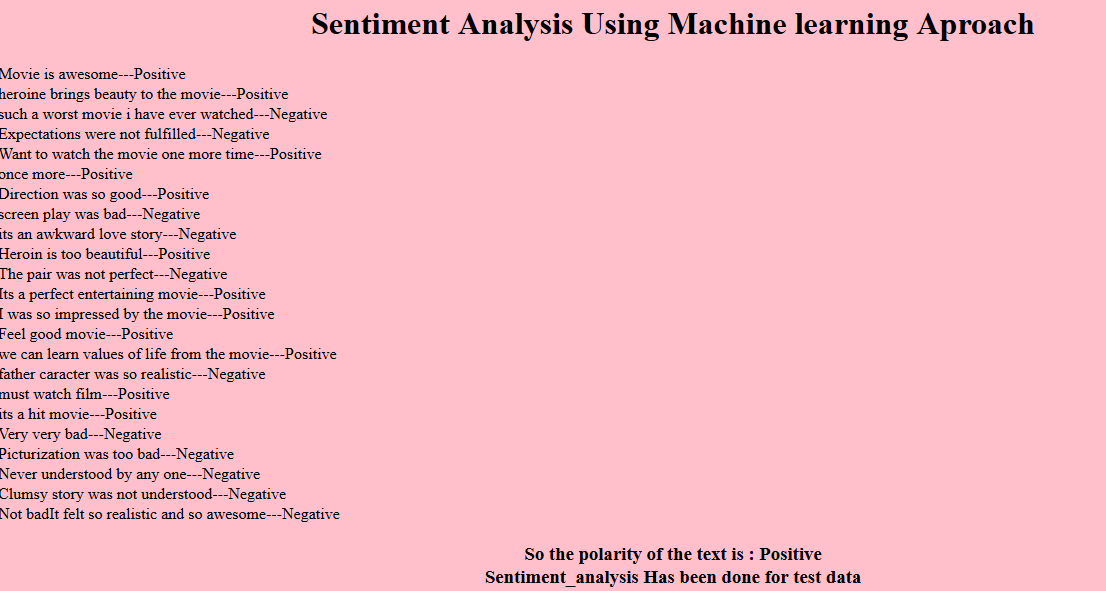
Negative

Architecture of the project

e Polarity

## Testing & Results





### SOFTWARES REQUIREMENTS:

Python (v2.7.4), notepad

### PROJECT MANAGEMENT

| **Team member** | **Roles and skills** |
| --- | --- |
| Arun Teja Dutta | Design, Coding, Rework |
| Sukanya Neela | Documentation, Testing, Review |

### DELIVERABLES AND CHECKPOINTS

|  |  |  |  |
| --- | --- | --- | --- |
| **Checkpoint Date** | **Expected Deliverable** | **Responsible team member(s)** | **Checkpoint results** |
| 3/8/2016 | Documentation | Arun Teja Dutta, Sukanya Neela | Project proposal |
| 3/17/2016 | Documentation | Arun Teja Dutta, Sukanya Neela | Github up to date |
| 4/25/2016 | Source code for the project | Arun Teja Dutta, Sukanya Neela | Implementation |
| 4/27/2016 | Test Plan | Arun Teja Dutta, Sukanya Neela | Testing |
| 5/1/2016 | Final Submission | Arun Teja Dutta, Sukanya Neela | Final Document and results |

#### References:

**[1] Chenhao Tan, Lillian Lee, Jie Tang, Long Jiang, Ming Zhou, Ping Li Cornell University "User-Level Sentiment Analysis Incorporating Social Networks"**

**[2] Bo Pang and Lillian Lee, Shivakumar Vaithyanathan “Thumbs up? Sentiment Classification using Machine Learning Techniques”**

**[3] McGrawHill\_-\_Machine\_Learning\_ by Tom\_Mitchell – Bayesian Learning**

**[4] Data\_Mining\_Concepts&Techniques\_Edi2 by Jiawei han and Micheline Kamber [5]**[**http://pymol.sourceforge.net/newman/user/S0120install.html**](http://pymol.sourceforge.net/newman/user/S0120install.html)