MES'S WADIA COLLEGE OF ENGINEERING, PUNE

Honors* in Artificial Intelligence and Machine Learning Fourth year of Engineering

410302: Machine learning Laboratory

NAME OF STUDENT:	**CLASS: BE**
SEMESTER/YEAR: VII	**ROLL NO:**
DATE OF PERFORMANCE:	**DATE OF SUBMISSION:**
EXAMINED BY: Dr. N. F. Shaikh	**EXPERIMENT NO: 05**

TITLE: Text classification for Sentimental analysis using KNN

AIM/PROBLEM STATEMENT: Perform Text classification for Sentimental analysis using KNN
Note: Use twitter data

OBJECTIVES:

- To understand text classification for Sentimental analysis
- To understand KNN

OUTCOMES:

- Processing a Sentimental analysis
- Design KNN

PRE-REQUISITES:

1. Theoretical knowledge of text pre-processing

2. Different pre-processing laws overview

THEORY:

Introduction

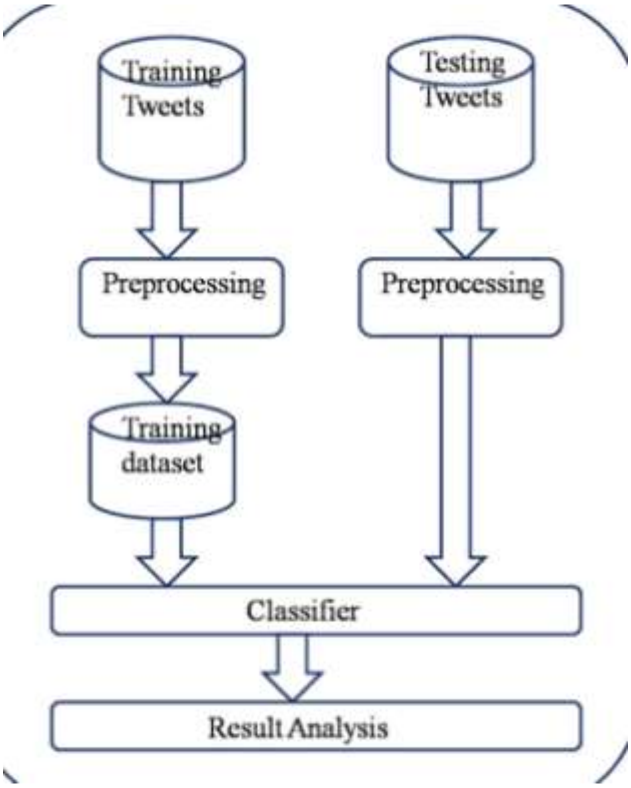
Micro blogging has turned into an extremely mainstream communication tool among web users. Many users share sentiments on various parts of life, every day on prevalent sites, for example, Twitter and Facebook. Prodded by this development, companies and media associations are progressively looking for approaches to mine these web-based social networking or information about what individuals consider for their organizations and items. Political gatherings might be intrigued to know whether individuals bolster their events or not. Associations that are social need to know individuals' assessment on verbal confrontations. The information can be obtained from micro blogging administrations, which users post sentiments on numerous parts of their life regularly. However, micro blogging information is unique in relation to normal content because it is extremely noisy in nature. A great deal of fascinating work is done keeping in mind the end goal to recognize feelings or sentiments from Twitter micro blogging information too. We intend an approach to automatically extract sentiment from a tweet. It is extremely supportive in nature that it enables input to be aggregated without any manual intercession. There are many researches in the area of sentiment classification. Mainly its greater part has concentrated on characterizing larger pieces of text, similar to reviews. Tweets (micro blogs) are dissimilar from reviews essentially due to their motivation: while tweet is easy going and partial to 140 characters of content. For the most part, tweets are not as insight fully created as reviews. Organizations can likewise utilize this to accumulate basic feedback about issues in recently launched items. In previous researches like Pang et al. [2] movie reviews are examined over many classifiers. This work of Pang et al. [2] is provided as a base in many research and many researchers have used the basic procedure in many areas. With the substantial scale of topics talked about on Twitter, it would be tremendously difficult to manually gather enough information to train a sentiment classifier for tweets. We run classifiers trained on emoticon data not in favor of a test set of tweets (which could conceivably have emoticons in them.

The steps taken are are as below:

A. Data

Data

For performing any classification, we need data set. Training and testing tweets are collected from Sentiment 140. We had 16, 00,000 training tweets and approx 448 testing tweets. The file was in excel format and converted in suitable format. We used MATLAB tool for our implementation. From the data, we took two main characteristics that were tweets itself and tweets labels (positive or negative)



B. Pre-processing

Tweets collected are not in usable format. Thus basic data pre-processing steps need to be taken in order to avoid noisy values. User names, any links, hash tags, emoticons, repeated letters, were removed. The Pre-processed data did not contain any common words which do not possess any sentiment like the, about, are etc. Stemming is also done for obtaining basic word for classification. For example word national was stemmed to nation

C. Working

After all pre-processing steps a corpus is made which has unique words alphabetically arranged. The model is trained first for polarity based classification and then for K-nn classifier. For K-nn classification, bag of word is maintained for each training and testing tweet. Then minimum Euclidian distance between training and testing bag of words is calculated. We set value of K as one-fourth of the data and choose the maximum occurred sentiment label from those one- fourth values.

Performing sentiment analysis on Twitter data usually involves four steps:

1. Gather Twitter data
2. Preprocess and prepare the data.
3. Train and test a Sentiment Analysis model
4. Visualize and verify the results

Question

1. What is Sentiment Analysis and its steps

2 How you can perform operation on training data?