# Twitter Sentiment Analysis

**DESCRIPTION:**

This project addresses the problem of sentiment analysis in twitter, that is classifying tweets according to the sentiment expressed in them (i.e positive, negative).Due to large amount of usage we hope to achieve a reflection of public sentiment by analysing the sentiments expressed in the tweets. Analysing the public sentiment is important for many applications such as firms trying to find out the response of their products in the market. The aim of this project is to develop a functional classifier for accurate and automatic sentiment classification of an unknown tweet stream.

**TECHNOLOGY USED:**

Jupyter notebook , Python, Rapid Miner

**MACHINE LEARNING:**

* Machine learning is an interdisciplinary research area which combines ideas from several branches of science namely, artificial intelligence, statistics, information theory, mathematics, etc.
* The prime focus of machine learning research is on the development of fast and efficient learning algorithms which can make predictions on data. When dealing with data analytics, machine learning is an approach used to create models for prediction. Machine learning tasks are mainly grouped into three categories- supervised, unsupervised and reinforcement learning. Supervised machine learning requires training with labeled data. Each labeled training data consists of input value and a desired target output value.
* The supervised learning algorithm analyzes the training data and makes an inferred function, which may be used for mapping new values. In unsupervised machine learning technique, hidden insights are drawn from unlabeled data sets, for example, cluster analysis.
* The third category, reinforcement learning allows a machine to learn its behavior from the feedback received through the interactions with an external environment. From a data processing point of view, both supervised and unsupervised learning techniques are preferred for data analysis and reinforcement techniques are preferred for decision making problems.

**SUPPORT VECTOR MACHINE:**

In [machinelearning](https://en.wikipedia.org/wiki/Machine_learning), support-vector-machines (SVMs,also support-vectornetworks**)** are [supervised learning](https://en.wikipedia.org/wiki/Supervised_learning) models with associated learning [algorithms](https://en.wikipedia.org/wiki/Algorithm) that analyze data used for [classification](https://en.wikipedia.org/wiki/Statistical_classification) and [regression analysis](https://en.wikipedia.org/wiki/Regression_analysis). Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-[probabilistic](https://en.wikipedia.org/wiki/Probabilistic_classification) [binary](https://en.wikipedia.org/wiki/Binary_classifier) [linear classifier](https://en.wikipedia.org/wiki/Linear_classifier) (although methods such as [Platt scaling](https://en.wikipedia.org/wiki/Platt_scaling) exist to use SVM in a probabilistic classification setting). An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible.The support-vector clustering algorithm, created by [Hava Siegelmann](https://en.wikipedia.org/wiki/Hava_Siegelmann) and [Vladimir Vapnik](https://en.wikipedia.org/wiki/Vladimir_Vapnik), applies the statistics of support vectors, developed in the support vector machines algorithm, to categorize unlabeled data, and is one of the most widely used clustering algorithms in industrial applications.

**NAIVE BAYES ALGORITHM:**

Naive Bayes is a simple technique for constructing classifiers: models that assign class labels to problem instances, represented as vectors of [feature](https://en.wikipedia.org/wiki/Feature_vector) values, where the class labels are drawn from some finite set. There is not a single [algorithm](https://en.wikipedia.org/wiki/Algorithm) for training such classifiers, but a family of algorithms based on a common principle: all Naive Bayes classifiers assume that the value of a particular feature is [independent](https://en.wikipedia.org/wiki/Independence_(probability_theory)) of the value of any other feature, given the class variable. For example, a fruit may be considered to be an apple if it is red, round, and about 10 cm in diameter. A Naive Bayes classifier considers each of these features to contribute independently to the probability that this fruit is an apple, regardless of any possible [correlations](https://en.wikipedia.org/wiki/Correlation_and_dependence) between the color, roundness, and diameter features.

For some types of probability models, naive Bayes classifiers can be trained very efficiently in a [supervised learning](https://en.wikipedia.org/wiki/Supervised_learning) setting. In many practical applications, parameter estimation for naive Bayes models uses the method of [maximum likelihood](https://en.wikipedia.org/wiki/Maximum_likelihood); in other words, one can work with the naive Bayes model without accepting [Bayesian probability](https://en.wikipedia.org/wiki/Bayesian_probability) or using any Bayesian methods.

Despite their naive design and apparently oversimplified assumptions, naive Bayes classifiers have worked quite well in many complex real-world situations. In 2004, an analysis of the Bayesian classification problem showed that there are sound theoretical reasons for the apparently implausible [efficacy](https://en.wikipedia.org/wiki/Efficacy) of naive Bayes classifiers. Still, a comprehensive comparison with other classification algorithms in 2006 showed that Bayes classification is outperformed by other approaches, such as [boosted trees](https://en.wikipedia.org/wiki/Boosted_trees) or [random forests](https://en.wikipedia.org/wiki/Random_forests).

**Existing system:**

The Existing work mainly focuses on classifying user into binary sentiment(i.e positive,negative) and they do not go further in mining user sentiment.The Existing approaches mainly leverage product category information or tag information to study the interpersonal influence.User reviews can provide us ideas in mining interpersonal interference and user preferences.

**PROPOSED SYSTEM**

* Proposed system will gives you the freedom to choose the data of any topic.
* It gives you the impact the results and statistics will have on the respective field.
* Proposed system allows retrieval of data based on the query entered by the user.
* Proposed system will provide accurate feature selection.

**PYTHON PACKAGES USED FOR TWITTER ANALYSIS**

* Tweepy: It is an open-sourced and enables Python to communicate with Twitter platform and use its API Mainly used for processing the datas.
* Text Blob: Processing textual data. It is used to process the twitter data in python packets.
* NLTK: Data understanding. It is used to understand the language used in data.These three tools are mandatory.

Hence by using these tools we bring the datas inside the python by tweepy, then we segregate the words by using text blob then by using NLTK we understand the meaning of the data.

**PROCESS:**

* Collecting Datasets from Twitter.
* Importing Datasets into Python.
* Packages used for python languages are tweepy, textblob and NLTK(Natural Language Tool Kit.
* For collected datasets we had performed data cleaning and analysed the total amount of tweets.
* We had used two algorithms to find the prediction and accuracy. First we had implemented Naïve Baye’s and then support vector algorithm.
* By using Rapid Miner we implemented these algorithms.
* In Naïve Baye’s first we retrieved our datasets, though by using cross validation we had done training and testing. All these datas are connected with Naïve.
* Through pie chart and graph we had shown the difference between Android and Apple
* Finally the accuracy for the tweets are analysed.
* The same steps had been followed for Support vector algorithm
* Finally the accuracy for both algorithms are equal with the difference in timings .Hence the prediction shows that the positive comments for iphone is more then Android.

**ACKNOWLEDGE:**

**Student Name:** Anitha Rajam.M, **Project Guide: Mrs.Yukthika**

D.Dheekshitha, **Signature:**

S.Atchaya **Date:**

**Class:** B.E IV-CSE-‘A’

**Signature:**

**Date:**