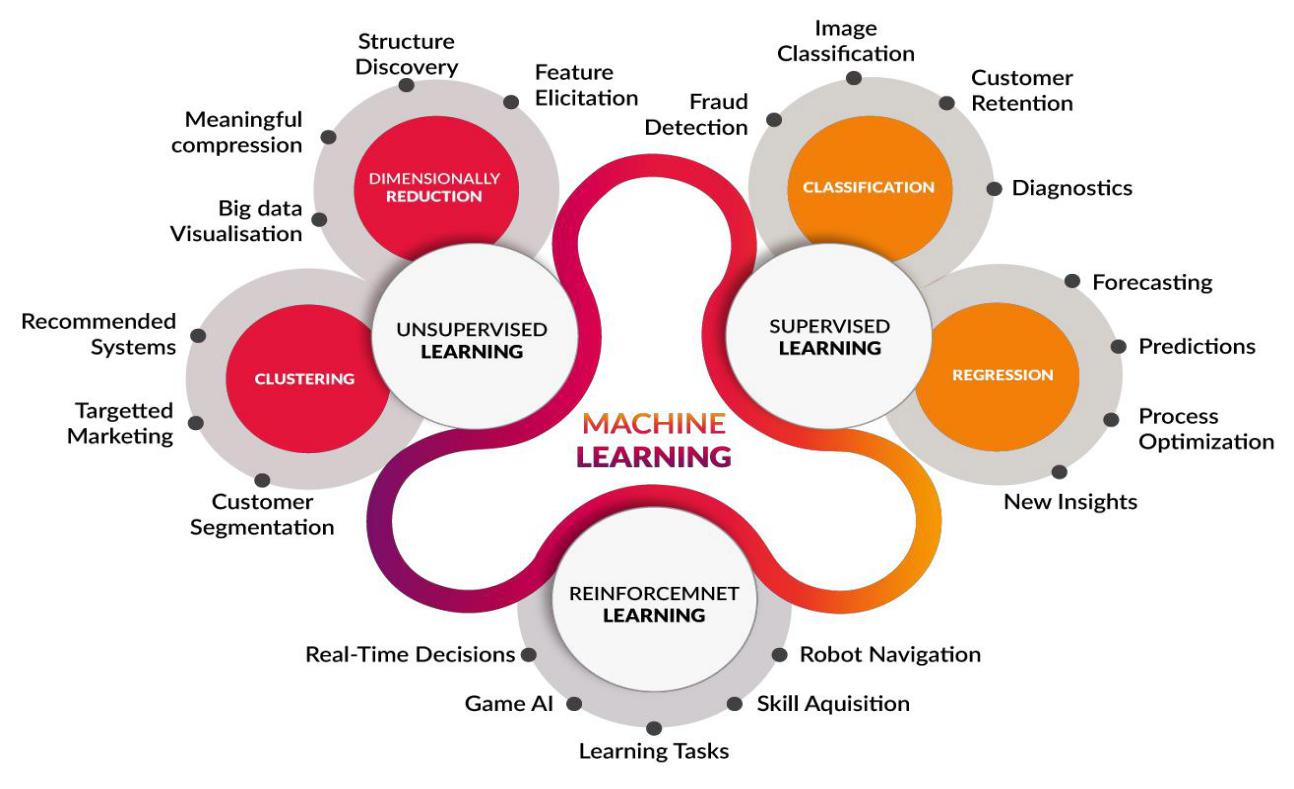
**CHAPTER 1**

**INTRODUCTION**

1. **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.

**1.1 PROBLEM DEFINITION**

* Twitter is an online micro-blogging and social-networking platform which allows users to write short status updates of maximum length 140 characters.
* This project addresses the problem of sentiment analysis in twitter; that is classifying tweets according to the sentiment expressed in them: positive, negative or neutral.
* Based on Sentiment Analysis we collect all the tweets from the public and will find which product is better than other.

**1.2 OBJECTIVE OF THE PROJECT**

The objective of this project is to develop a functional classifier for accurate and automatic sentiment classification of an unknown tweet streamout the response of their products in the market.

**MACHINE LEARNING ALGORITHMS**

The list of commonly used machine learning algorithms that can be applied to almost any data problem −

* Linear Regression
* Logistic Regression
* Decision Tree
* SVM
* Naive Bayes
* KNN
* K-Means
* Random Forest
* Dimensionality Reduction Algorithms

**SUPPORT VECTOR MACHINE**

In [machinelearning](https://en.wikipedia.org/wiki/Machine_learning" \o "Machine learning), support-vector-machines (SVMs, also support-vector networks**)** 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" \o "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).

**1.3 SIGNIFICANCE OF THE PROJECT**

* It is important to provide accurate values using Machine learning algorithms.
* Provide Feedback about a product.
* Comparison of machine learning algorithms to find the accuracy rate of Products.

**1.4 OUTLINE OF THE PROJECT**

This project addresses the problem of sentiment analysis in twitter,that is classifying tweets according to the sentiment expressed in them: positive, negative or neutral.

**CHAPTER 2**

**LITERATURE REVIEW**

Sentiment analysis of in the domain of micro-blogging is a relatively new research topic so there is still a lot of room for further research in this area. Decent amount of related prior work has been done on sentiment analysis of user reviews, documents, web blogs/articles and general phrase level sentiment analysis. These differ from twitter mainly because of the limit of 140 characters per tweet which forces the user to express opinion compressed in very short text. The best results reached in sentiment classification use supervised learning techniques such as Naive Bayes and Support Vector Machines, but the manual labelling required for the supervised approach is very expensive. Some work has been done on unsupervised approaches, and there is a lot of room of improvement. Various researchers testing new features and classification techniques often just compare their results to base-line performance. There is a need of proper and formal comparisons between these results arrived through different features and classification techniques in order to select the best features and most efficient classification techniques for particular applications.

The bag-of-words model is one of the most widely used feature model for almost all text classification tasks due to its simplicity coupled with good performance. The model represents the text to be classified as a bag or collection of individual words with no link or dependence of one word with the other, i.e. it completely disregards grammar and order of words within the text. This model is also very popular in sentiment analysis and has been used by various researchers. The simplest way to incorporate this model in our classifier is by using unigrams as features. Generally speaking n-grams is a contiguous sequence of “n” words in our text, which is completely independent of any other words or grams in the text. So unigrams is just a collection of individual words in the text to be classified, and we assume that the probability of occurrence of one word will not be affected by the presence or absence of any other word in the text. This is a very simplifying assumption but it has been shown to provide rather good performance. One simple way to use unigrams as features is to assign them with a certain prior polarity, and take the average of the overall polarity of the text, where the overall polarity of the text could simply be calculated by summing the prior polarities of individual unigrams.

Prior polarity of the word would be positive if the word is generally used as an indication of positivity, for example the word “sweet”, while it would be negative if the word is generally associated with negative connotations, for example “evil”. There can also be degrees of polarity in the model, which means how much indicative is that word for that particular class. A word like “awesome” would probably have strong subjective polarity along with positivity, while the word “decent” would although have positive prior polarity but probably with weak subjectivity.

Many of the researchers in this field have used already constructed publicly available lexicons of sentiment bearing words while many others have also explored building their own prior polarity lexicons .

The basic problem with the approach of prior polarity approach has been identified by Wilson et al. who distinguish between prior polarity and contextual polarity . They say that the prior polarity of a word may in fact be different from the way the word has been used in the particular context.

In this example all of the four underlined words “trust”, “well”, “reason” and “reasonable” have positive polarities when observed without context to the phrase, but here they are not being used to express a positive sentiment. This concludes that even though generally speaking a word like “trust” may be used in positive sentences, but this doesn’t rule out the chances of it appearing in non-positive sentences as well.

The task of twitter sentiment analysis can be most closely related to phrase level sentiment analysis. A seminal paper on phrase level sentiment analysis was presented in 2005 by Wilson et al. which identified a new approach to the problem by first classifying phrases according to subjectivity (polar) and objectivity (neutral) and then further classifying the subjective-classified phrases as either positive or negative. The paper noticed that many of the objective phrases used prior sentiment bearing words in them, which led to poor classification of especially objective phrases.

It claims that if we use a simple classifier which assumes that the contextual polarity of the word is merely equal to its prior polarity gives a result of about 48%. The novel classification process proposed by this paper along with the list of ingenious features which include information about contextual polarity resulted in significant improvement in performance (in terms of accuracy) of the classification process.

The paper reports positive results for their study that the more number of cases a word has of lengthening, the more chance there of that word being a strong indication of subjectivity. The most commonly used classification techniques are the Naive Bayes Classifier and State Vector Machines. Some researchers like Barbosa et al. publish better results for SVMs while others like Pak et al. support Naive Bayes and also report good results for Maximum Entropy classifier. Some of the earliest work in this field classified text only as positive or negative, assuming that all the data provided is subjectize. While this is a good assumption for something like movie reviews but when analyzing tweets and blogs there is a lot of objective text we have to consider, so incorporating neutral class into the classification process is now becoming a norm.

There has also been very recent research of classifying tweets according to the mood expressed in them, which goes one step further. Bollen et al. explores this area and develops a technique to classify tweets into six distinct moods: tension, depression, anger, vigrous, fatigue and confusion. They use an extended version of Profile of Mood States (POMS): a widely accepted psychometric instrument. They generate a word dictionary and assign them weights corresponding to each of the six mood states, and then they represented each tweet as a vector corresponding to these six dimensions. However not much detail has been provided into how they built their customized lexicon and what technique did they use for classification.

**Introduction:**

This project of analyzing sentiments of tweets comes under the domain of “Pattern Classification” and “Data Mining”. Both of these terms are very closely related and intertwined, and they can be formally defined as the process of discovering “useful” patterns in large set of data, either automatically (unsupervised) or semiautomatically (supervised). The project would heavily rely on techniques of “Natural Language Processing” in extracting significant patterns and features from the large data set of tweets and on “Machine Learning” techniques for accurately classifying individual unlabelled data samples (tweets) according to whichever pattern model best describes them. The features that can be used for modeling patterns and classification can be divided into two main groups: formal language based and informal blogging based. Language based features are those that deal with formal linguistics and include prior sentiment polarity of individual words and phrases, and parts of speech tagging of the sentence. Prior sentiment polarity means that some words and phrases have a natural innate tendency for expressing particular and specific sentiments in general. For example the word “excellent” has a strong positive connotation while the word “evil” possesses a strong negative connotation. So whenever a word with positive connotation is used in a sentence, chances are that the entire sentence would be expressing a positive sentiment. Parts of Speech tagging, on the other hand, is a syntactical approach to the problem. It means to automatically identify which part of speech each individual word of a sentence belongs to: noun, pronoun, adverb, adjective, verb, interjection, etc. Patterns can be extracted from analyzing the frequency distribution of these parts of speech (either individually or collectively with some other part of speech) in a particular class of labeled tweets. Twitter based features are more informal and relate with how people express themselves on online social platforms and compress their sentiments in the limited space of 140 characters offered by twitter. They include twitter hashtags, retweets, word capitalization, word lengthening , question marks, presence of URL in tweets, exclamation marks, internet emoticons and internet shorthand/slangs. Classification techniques can also be divided into a two categories: Supervised vs. unsupervised and non-adaptive vs. adaptive/reinforcement techniques. Supervised approach is when we have pre-labeled data samples available and we use them to train our classifier. Training the classifier means to use the pre-labeled to extract features that best model the patterns and differences between each of the individual classes, and then classifying an unlabeled data sample according to whichever pattern best describes it. For example if we come up with a highly simplified model that neutral tweets contain 0.3 exclamation marks per tweet on average while sentiment-bearing tweets contain 0.8, and if the tweet we have to classify does contain 1 exclamation mark then (ignoring all other possible features) the tweet would be classified as subjective, since 1 exclamation mark is closer to the model of 0.8 exclamation marks. Unsupervised classification is when we do not have any labeled data for training. In addition to this adaptive classification techniques deal with feedback from the environment. In our case feedback from the environment can be in form of a human telling the classifier whether it has done a good or poor job in classifying a particular tweet and the classifier needs to learn from this feedback.

**Algorithms can be used:**

* Support Vector Machine
* Naïve Bayes

**Software requirements:**

* Python
* Jupyter
* Anaconda
* Spreadsheet