DBSCAN or Density Based Spatial Clustering of Applications with Noise is an unsupervised clustering technique. It works on two parameters: *epsilon* and *min points*. *Epsilon* defines the maximum distance between two points for them to be a part of the same cluster. *Min points* defines the minimum number of points to form a dense region.

The main idea behind DBSCAN is to classify points into three groups: core point, border point and noise point.

A core point is one which has *min points* number of points within *epsilon* distance. A boundary point is a point which is a neighbor of a core point but does not have *min points* number of points within *epsilon* distance. A noise point is one which is neither a core point nor a boundary point.

Algorithm:

1. Choose a point in the space. Find the distance to all other points in that space. If there are at least *min points* number of points within at most *epsilon* distance of the point, mark this point as a core point and all its neighbors within *epsilon* distance as part of this cluster.
2. If this point is not a core point itself, but is in the vicinity of another core point, then mark this point as a boundary point of that cluster
3. If this point is neither a core point nor a boundary point, then mark it as a noise point
4. Repeat for all points in the space till each is not classified

It is a very straightforward algorithm that can be used to identify clusters of arbitrary shapes and sizes.

A multinomial naïve bayes classifier is a popular classifier for text data. It works on Bayes’ rule of probability which is:

Where represents the probability of A happening, given B has already taken place.

Multinomial naïve bayes classifier taken in the Tf-idf vector as the feature vector. Then, taking the vector value of each word, we calculate the probability of the word being in the cluster. Here, will represent the probability of a word occurring given we are given a cluster label. Based on the values of the probabilities obtained, each word is classified into a cluster with the maximum conditional probability.

Multinomial naïve bayes classifier was chosen for the purpose of verification because it works well with text data and tfidf vectors.