Tweets are collected automatically using Twitter API and they are manually annotated as positive or negative. A dataset is created by taking 600 positive tweets and 600 negative tweets. Table 1 shows how dataset is split into training set and test set.

### *B. Preprocessing of Tweets*

Keyword extraction is difficult in twitter due to misspellings and slang words. So to avoid this, a preprocessing step is performed before feature extraction. Preprocessing steps include removing url, avoiding misspellings and slang words. Misspellings are avoided by replacing repeated characters with 2 occurrences. Slang words contribute much to the emotion of a tweet. So they can't be simply removed. Therefore a slang word dictionary is maintained to replace slang words occurring in tweets with their associated meanings. Domain information contributes much to the formation of slang word dictionary.

### *C. Creation of Feature Vector*

Feature extraction is done in two steps. In the first step, twitter specific features are extracted. Hashtags and emoticons are the relevant twitter specific features. Emoticons can be positive or negative. So they are given different weights. Positive emoticons are given a weight of '1' and negative emoticons are given a weight of '-1'. There may be positive and negative hashtags. Therefore the count of positive hashtags and negative hashtags are added as two separate features in the feature vector.

Twitter specific features may not be present in all tweets. So a further feature extraction is to be done to obtain other features. After extracting twitter specific features, they are negative and '0' in its absence. Part of speech feature is given a value of '1' if it is relevant and '0' otherwise.

Thus feature vector is composed of 8 relevant features. The 8 features used are part of speech (pos) tag, special keyword, presence of negation, emoticon, number of positive keywords, number of negative keywords, number of positive hash tags and number of negative hash tags.

### *D. Sentiment Classification*

After creating a feature vector, classification is done using Naive Bayes, Support Vector Machine, Maximum Entropy and Ensemble classifiers and their performances are compared.

## IV. Classification Techniques

There are different types of classifiers that are generally used for text classification which can be also used for twitter sentiment classification.

### *A. Nave Bayes Classifier*

Nave Bayes Classifier makes use of all the features in the feature vector and analyzes them individually as they are equally independent of each other. The conditional probability for Naive Bayes can be defined as

$$P(X|y_j) = \Pi_{i=1}^{m} P(x_i|y_j) \quad (1)$$

'X' is the feature vector defined as X=$\{x_1,x_2,....x_m\}$ and $y_j$ is the class label. Here, in our work there are different independent features like emoticons, emotional keyword, count of positive and negative keywords, and count of positive and negative hash tags which are effectively utilized by Naive