Write up for the Twitminer

1) Software Used: -
The programming language used is Python (2.7.3) and using Python packages nltk,sklearn and numpy

2) Features: -The word occured in the tweets,dates,scores,numbers,hashtag, @symbol,length of the tweet has been used as features for the method.the words are first changed to their inflectional form by stemming (using Snowball Stemmer in nltk toolkit)
and changed to lower case and then used as features for the classifier .
compound words are broken iinto constituent words to be used as features for example RafaelNadal is broken into Rafael and Nadal and then these two words are used as features.

3) Similarity/Distance Measures (if any) :--No

3) Classifier (if you have used any standard classifier):-
Support Vector Classifier with linear kernel.

Please describe the algorithms in details:

1) Pre-processing step:- First of all one tweet is selected ,the tweet is parsed and a list of constituent word simultaneously the words are
added in the dictionary having all the words appearing in all training tweets ,the number of hashtag,dates ,numbers,years ,hashtag,@symbol
,length of the tweet is also saved .When all the tweets in the training data is processed ,the the word list of every tweet is checked if
for each word in the dictionary the word exists in the wordlist of the tweet then 1 is added to the feature list otherwise 0.All the other
features like number of dates,years,hashtags etcalso added in the list and list is added to x_train then the tag is read if it is Sports then
1 is added to the y_train .Similarly in the validation data the features are made ,only diffesrence is that the words in the validation data
don't change the dictionary ,dictionary consists of words only from training data.

2) Training Algorithm:Support Vector Classifier

Input Format:-The array of zeros and ones for presence and absence of words along with other features.
Tunable Parameters:-Softmargin Parameter C
Output Format:-Array of zeros and ones of the length equal to the numbers of tweets.One for Sports prediction and Zero for political prediction.
Algorithm:
About SVM: Given a set of training examples, each marked as belonging to one of two categories, a SVM training algorithm builds a model that assigns
new examples into one category or the other. A 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. New examples are then mapped into that same space and predicted to
belong to a category based on which side of the gap they fall on.
So in our method after features are made.
The model is trained on x_train and subsequent y_train using :svm.fit(x_train,y_train)
then the prediction is made on validation data using :ans=svc.predict(x_validate)

3) Validation and Parameter Tuning:We try to find the parameter that gives the best accuracy. For which we use gridsearch on the validation data .
As given Grid-search in The sklearn provides an object that, given data, computes the score during the fit of an estimator on
a parameter grid and chooses the parameters to maximize the cross-validation score. This object takes an estimator
during the construction and exposes an estimator API.

```
from sklearn.grid_search import GridSearchCV
gammas = np.logspace(1/30,10,6)
clf = GridSearchCV(estimator=svc, param_grid=dict(gamma=gammas),n_jobs=-1)
So best parameter is found using gridsearch and used for prediction in our case it is C=1.0
best parameter is given by :
clf.best_estimator_.gamma
```

4) Testing Algorithm:
    For testing some part of the training data is not used for training and reserved
    fortesting the accuracy of the algorithm .
    say last 1000 tweets are used for scoring then
    we will train on the remaining data and get the score on the last 1000 twwets by :
    np.score(x_train[-1000:],y_train[-1000:])
    which the precentage of entries which were correctly classified .

Explanation of results on validation data:

Why do you think your algorithm got the accuracy that it did on the validation data? Is scope for improvements?
The algorithm get its accuracy because it has been trained on the trained data using SVM with features made out of the words
and it is being used to make prediction on new test data .
Yes,there is much score for improvement and accuracy can be increased but we could not find better ways to improve effieciency .
