Tools that are used

Programming language: python3

nltk: library for tokenizing tweets

TurkishStemmer: library for stemming Turkish words

sklearn: library consisting of all classifiers that we used

Our Model

We used "Nearest Neighbors", "Linear SVM", "RBF SVM", "Decision Tree", "Random Forest", "Neural Net", "AdaBoost", "Naive Bayes", and "Logistic Regression" classifiers in the model.

First, we gave training data to each classifier and then split labeled data into two parts as training and test parts. (test data contains 3/10 of data) Then, We obtained accuracy of each of classifiers. After obtaining accuracy scores from classifiers, we assigned a coefficient that is linear to those accuracy scores to contribute to the model to classify tweets.

We filtered out classifiers that could not reach an accuracy value that is greater than 0.61 and so, we left with classifiers other than “Decision Tree”, ”Random Forest”, and “AdaBoost”.

Accuracy scores that we had for each classifier separately as belows:

*"Nearest Neighbors(kNN)" : 68.806,
*"Linear SVM" : 70.175,
*"RBF SVM" : 62.559,
*"Decision Tree" : 52.418,
*"Random Forest" : 43.047,
*"Neural Net" : 70.047,
*"AdaBoost" : 57.552,
*"Naive Bayes" : 67.180,
*"Logistic Regression" : 67.394

We used multiple classifiers as explained above because we wanted to increase individual accuracies of classifiers.

To increase accuracy, we used positive and negative words lists. Whenever a word in a positive tweet is in positive word list or a word in negative tweet is in negative word list, we added one more replica of the same tweet to data set. Stemming also had a positive effect on accuracy but we could not get the same effect when we do stop word removal.

One reason for us to not be able to enhance accuracy is shortness and both grammatically and semantically uncertainty of the tweets.

Some statistics

Below table shows accuracy scores when we applied the combination of stop word removal, stemming and duplicate data adding.

Stop word removal	***Stemming***	***Duplicate***	***Accuracy***
No	No	No	0.651
Yes	No	No	0.641
No	Yes	No	0.657
No	Yes	Yes	0.663
Yes	Yes	Yes	0.660