#AUTHOR – AMAN KOCHHAR

For our baseline we are using the code given to us by the Semeval team. This code produces an SVM model based on the tokenization of the data i.e. word frequency from the tweets. The model is then tested on the test data which is obtained by using the leave-one-out method.

To get our baseline we provide this code with the tweets data. All these files have been modified to remove a lot of "0" labeled tweets (tweets not selected in the top 10). This has been done to reduce the size of our data and decrease the time taken to train our model. However, each file still consists of the TOP 10 tweets and the winning tweets.

When running the default model on this data we get an accuracy of 49.5%

Corpus	Baseline accuracy
*Three hashtag files (Modified)	49.5%
*Three hashtag files (unmodified)	49.3%

- The three hashtag files used are – 420_celebs, Add_a_woman_improve_a_movie, Add_sports_ruin_a_song