CSL 772: Assignment 2
Write Up

Raghav Goyal
2010MT50612

Discussed with: Lovejeet Singh, Nimit Bindal
Data Processing - Initial processing is done with respect to bernoulli naive bayes to create data to even test classifiers. All steps are taken by observing processed tweets from previous operation.
Data Split: 80% Training set, 10% development set, 10% test set

Operation performed on data	**Classifier's precision**	**Possible explanation of why ?**	**Vocab**
Space based tokenization	Bernoulli NB: 0.767	-	
Discarded space based, TwitterNLP package tokenizer	Bernoulli NB:Increased (0.771)	more meaningful words	8 L
Removed "#" from hash tag Lower case all words	Bernoulli NB: Increased (0.773)	hash tags can mix up with other words enriching data	7 L
Words starting with digit	Bernoulli NB: Decreased	possible loss of words with content e.g 12Angry	7 L
URL removed hunnngrrry -> hungry isolated numbers removed tokens stripped based on: .,( )	Bernoulli NB: Increased (0.771)	Url doesn't contain sentiment, most of them were short url's, pics etc. Numbers and these symbols were unnecessary	6 L
Handles removed e.g. @userName	Bernoulli NB: Decreased	-	
Handles -> "_Handle_" regex to space the symbols: /:& tokens stripping symbols added: ;-*	Bernoulli NB: Back to where I started (0.773)	substituting handles reduced vocabulary rest of the operation were performed to set more words free	3 L
stop words, few selected ~50 words separated, attached with: :;. more stripping symbols added Lemmatized	Bernoulli NB: Decreased by 0.2 %	possible loss of sentiment in stop words operation retained because it reduced vocabulary size	2.5 L
Further more symbols added	Bernoulli NB: Increased (0.779)	based on data, symbols were added	2.3 L
	Multinomial NB: 0.784 Linear SVM: 0.782 Max-Ent: 0.791		
Negation added for one token next of " n't "	marginal increment		

- Initially **lemmatization** doesn't makes sense as tokenized data contains lots of punctuations and other symbols. Lemmatizer - Word Net
- Multinomial NB gave better accuracy than bernoulli, as it takes word frequency into account
- Final classifier was chosen to be **Maximum Entropy**, as it yielded better precision than others, probably because of its discriminative nature.
- Bi - grams could not be exploited as number of features increased significantly for any classifier to run