Web Mining Final Project Emotion Analysis from Tweets.

Tegjyot Singh Sethi
T0seth01@louisville.edu

Ranjith Kumar Nandella
Nandella.ranjithkumar@gmail.com

Introduction

- The goal of this project is to Analyze Tweets to classify them as either having a Positive or a Negative emotional undertone.
- Classification done solely based on the text in the tweets.
- Trying to find a relationship between the use of certain words and the mood of the user.
- Major challenges: Tweets tend to be messy and Short
- Most of the documents will end up becoming highly sparse vectors and hence might not lead to any useful information.

Tools Used

- Pattern, Python, Weka, Matlab

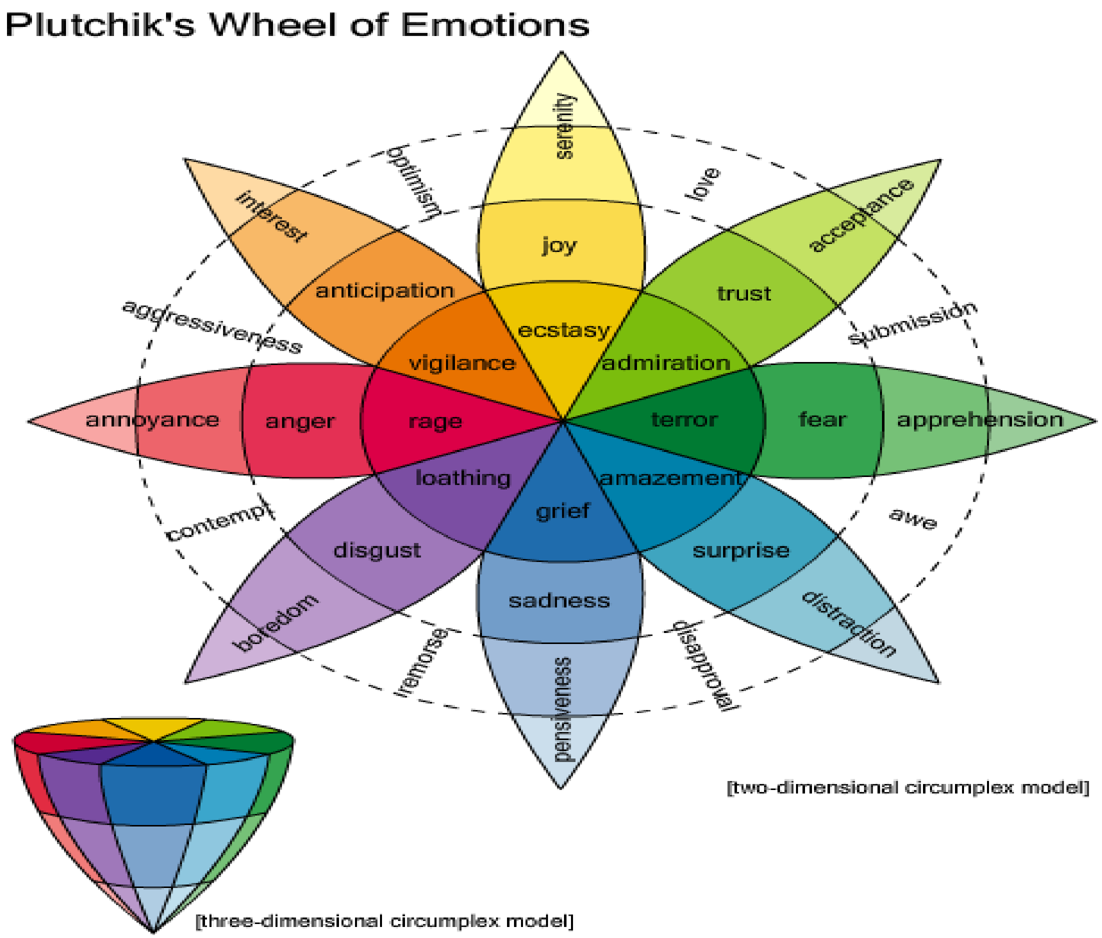
Pattern

Pattern is a web mining module for the Python programming language.

It bundles tools for data retrieval (Google + Twitter + Wikipedia API, web spider, HTML DOM parser), text analysis (rule-based shallow parser, WordNet interface, syntactical + semantical n-gram search algorithm, tf-idf + cosine similarity + LSA metrics), clustering and classification (*k*-means, *k*-NN, SVM), and data visualization (graph networks).

The module is bundled with 30+ example scripts and 350+ unit tests.

Motivation


Plutchik's Wheel of Emotions
serenity
joy
ecstasy
love
acceptance
trust
admiration
submission
apprehension
fear
terror
awe
distraction
surprise
amazement
disapproval
pensiveness
sadness
grief
remorse
boredom
disgust
loathing
contempt
annoyance
anger
rage
aggressiveness
interest
anticipation
vigilance
optimism
[two-dimensional circumplex model]
[three-dimensional circumplex model]

Data Used

- Twitter Data Stream queried for a week. Search based on any of the below mentioned hashtags.

- *Positive Emotions:*

  #joy, #happy, # bliss , #ecstasy, #merry

- *Negative Emotions:*

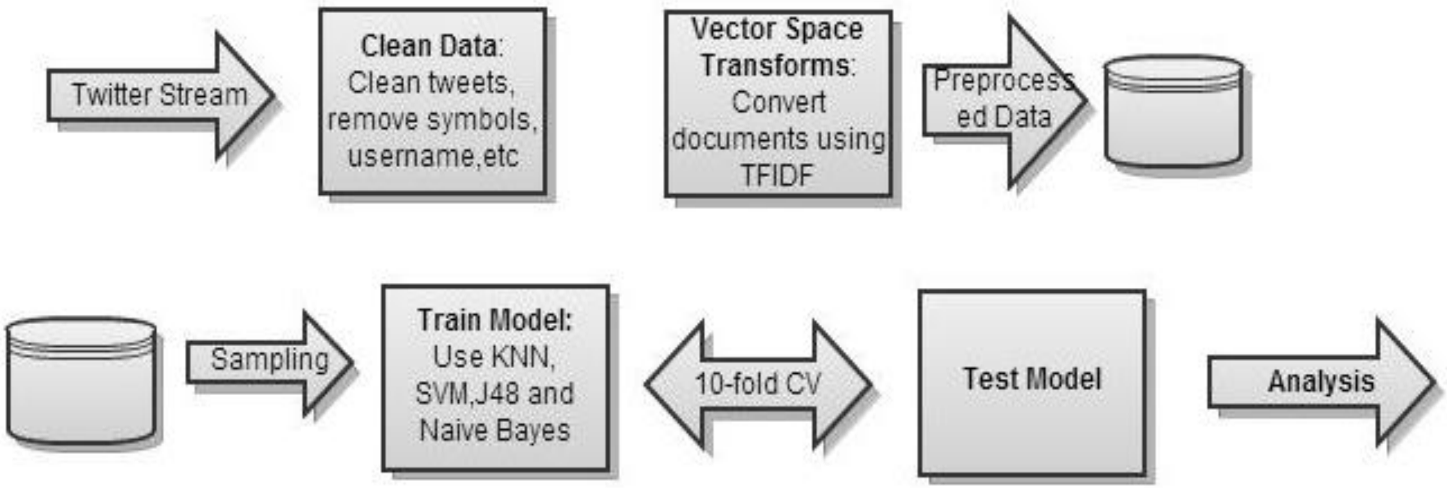
  #sad, #gloomy, # depressed, #mourn, #despair.

- The resulting tweets are sorted and duplicates are removed. The resulting files now have the following size:

  Positive Tweets: 22088; Negative Tweets: 16175

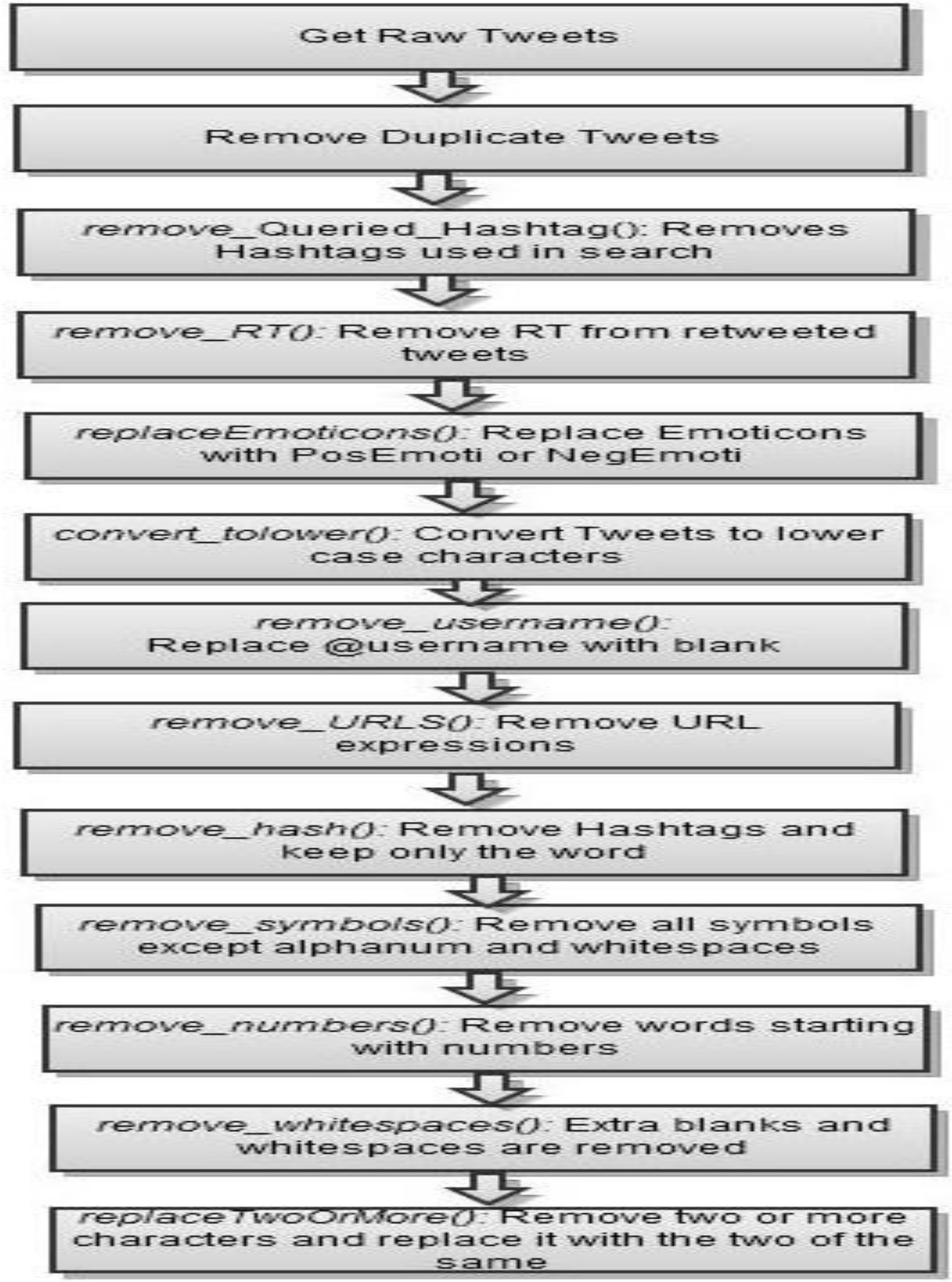
Architecture of the System

Objective: Analyzing Tweets to classify them as either having a Positive or a Negative mood.



Preprocessing (Data Cleaning)

Descriptive Text	Symbols
PosEmoti	:), :-),:o), :],:3,:c),:D, C:, ;), ;},:8
NegEmoti	:'(,;(,',D:, :{, :<, :-D, ', v.v, DX,D=,D;,D8,:C,:c , :-(, :(,
Heart	'<3'
BrokenHeart	'</3'



Preprocessing (Data Cleaning)

Sample Tweet

Raw:

- How to Avoid the and #Discouragement of Long Term #JobLoss. http://t.co/1RuLoLPg62 #Depression #Networking #HiddenJobMarket
- How to deal with #pessimism and even in the midst of hardship, with @carter_phipps: http://t.co/RrRfpmCwhA
- @hunterr_hancock @hannahkshumate #coldshoulder #ignore #sadness #depression #bacon #lubricant #yellowpages #brush #randomhashtags
- I advised my teenage cousin to checkout the #GWU podcast from @RealJudgeJules. His reply, "I'm an indie rock kinda guy".
- I can't find my Star Wars T-Shirt... @sonofsammie ! #despondency

After Preprocessing:

- how to avoid the and discouragement of long term jobloss depression networking hiddenjobmarket
- how to deal with pessimism and even in the midst of hardship with
- coldshoulder ignore sadness depression bacon lubricant yellowpages brush randomhashtags
- i advised my teenage cousin to checkout the gwu podcast from his reply im an indie rock kinda guy
- i cant find my star wars t shirt despondency

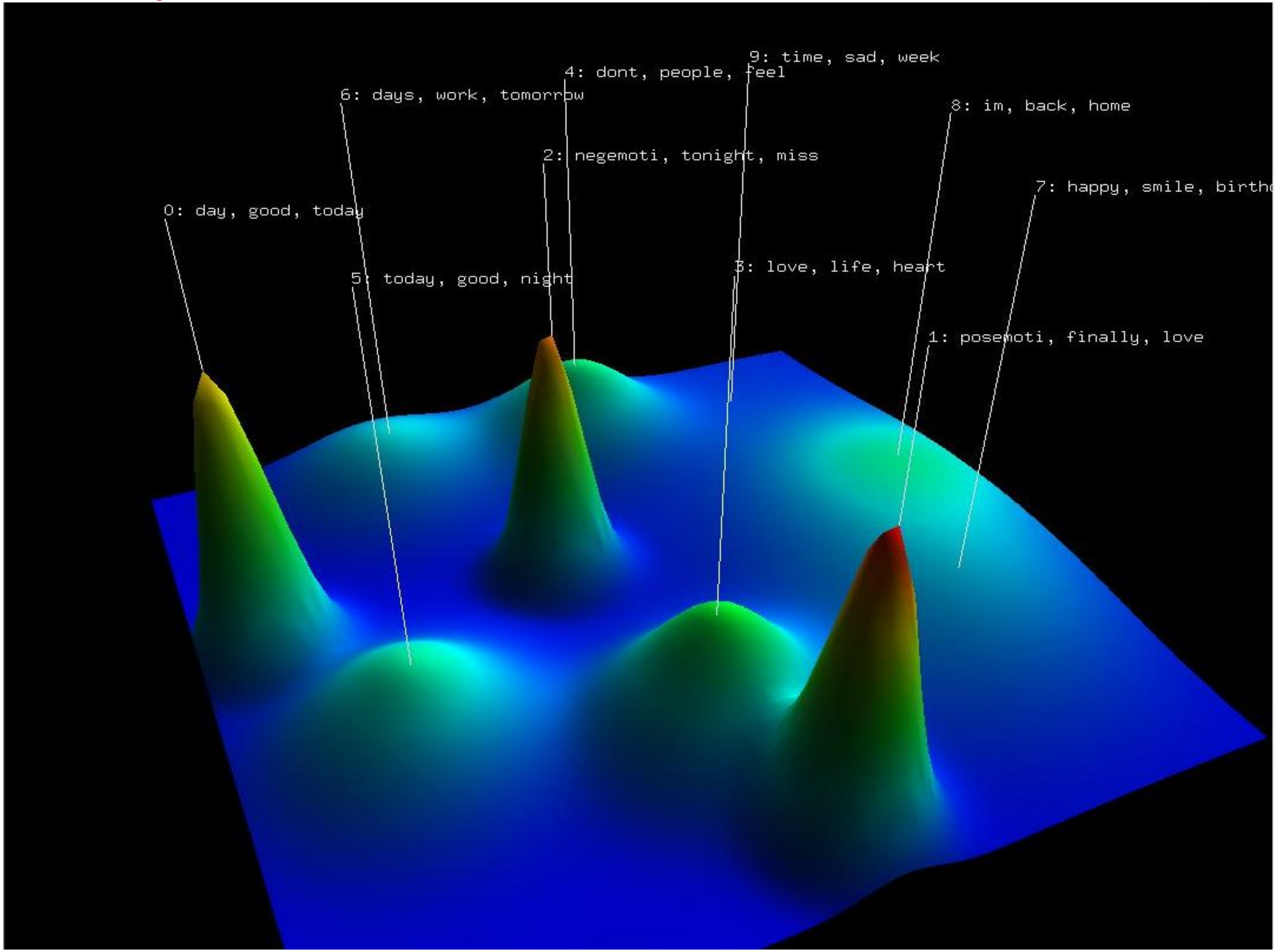
Preprocessing (Vector Space Transform)

- *TF-IDF transform*
- *Stemming* : PORTER Stemmer
- *Stop Words* Removal
- *Specify Wordcount:* 300,500,1000
- Resampling to Balance Class

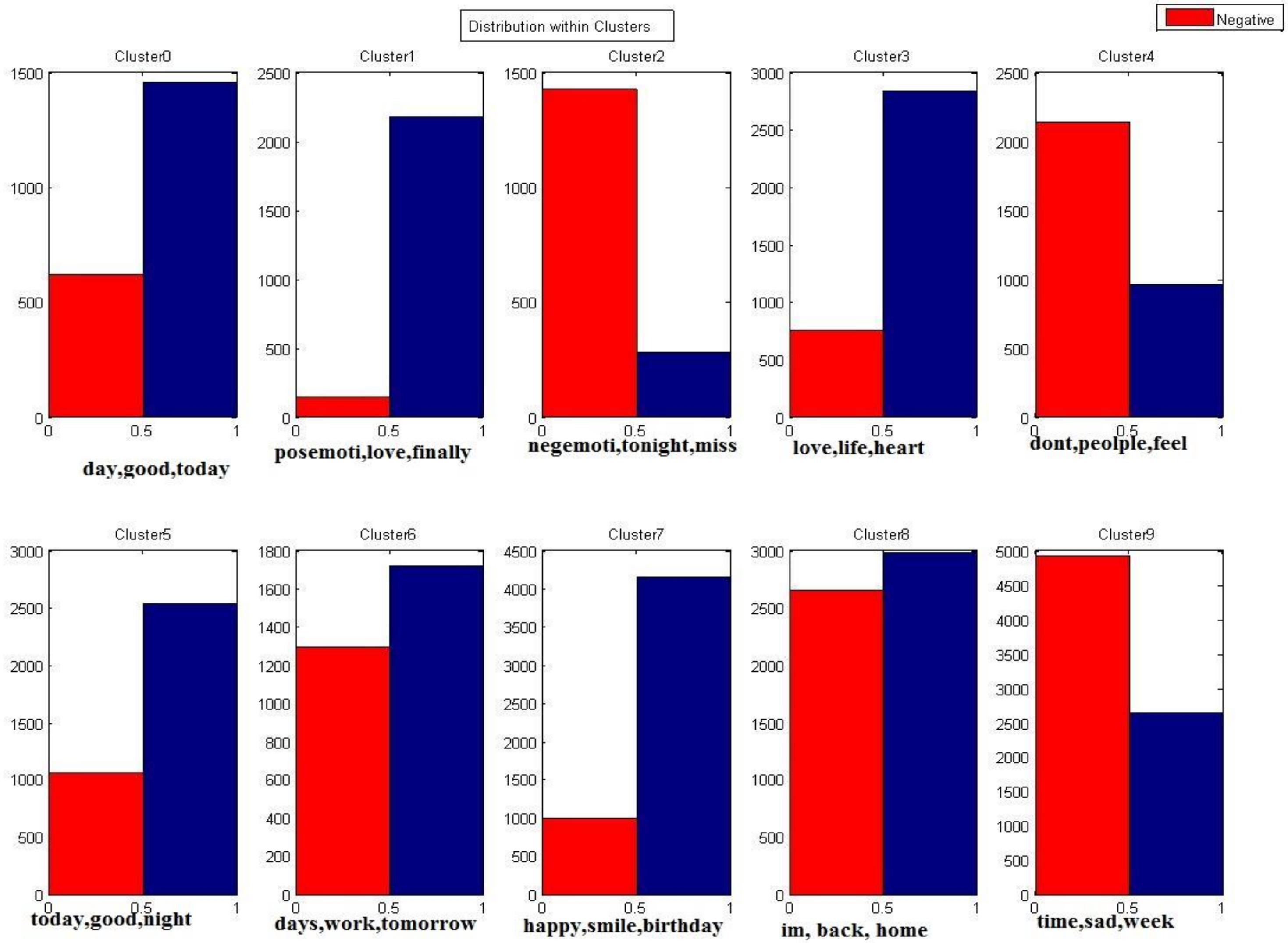
Data After Preprocessing

Negative Tweets: 15156
Positive Tweets: 15042
Instances: 30198
Attributes: 730

Data Exploration With CLUTO



Data Exploration With CLUTO



Data Exploration With CLUTO

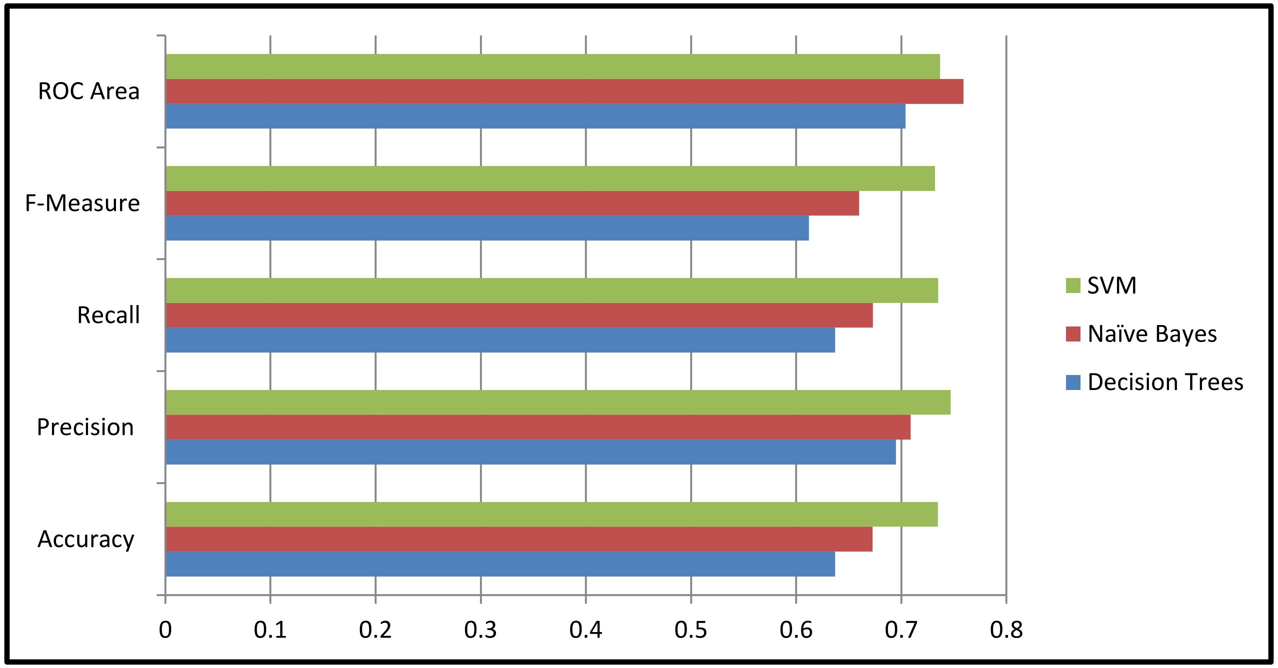
- *Positive Clusters*: day,good,today, posemoti, finally,love,life,heart, today, happy, smile, birthday.

- *Negative Cluster*: negemoti, tonight, miss, don't, people, feel, time, sad, week.

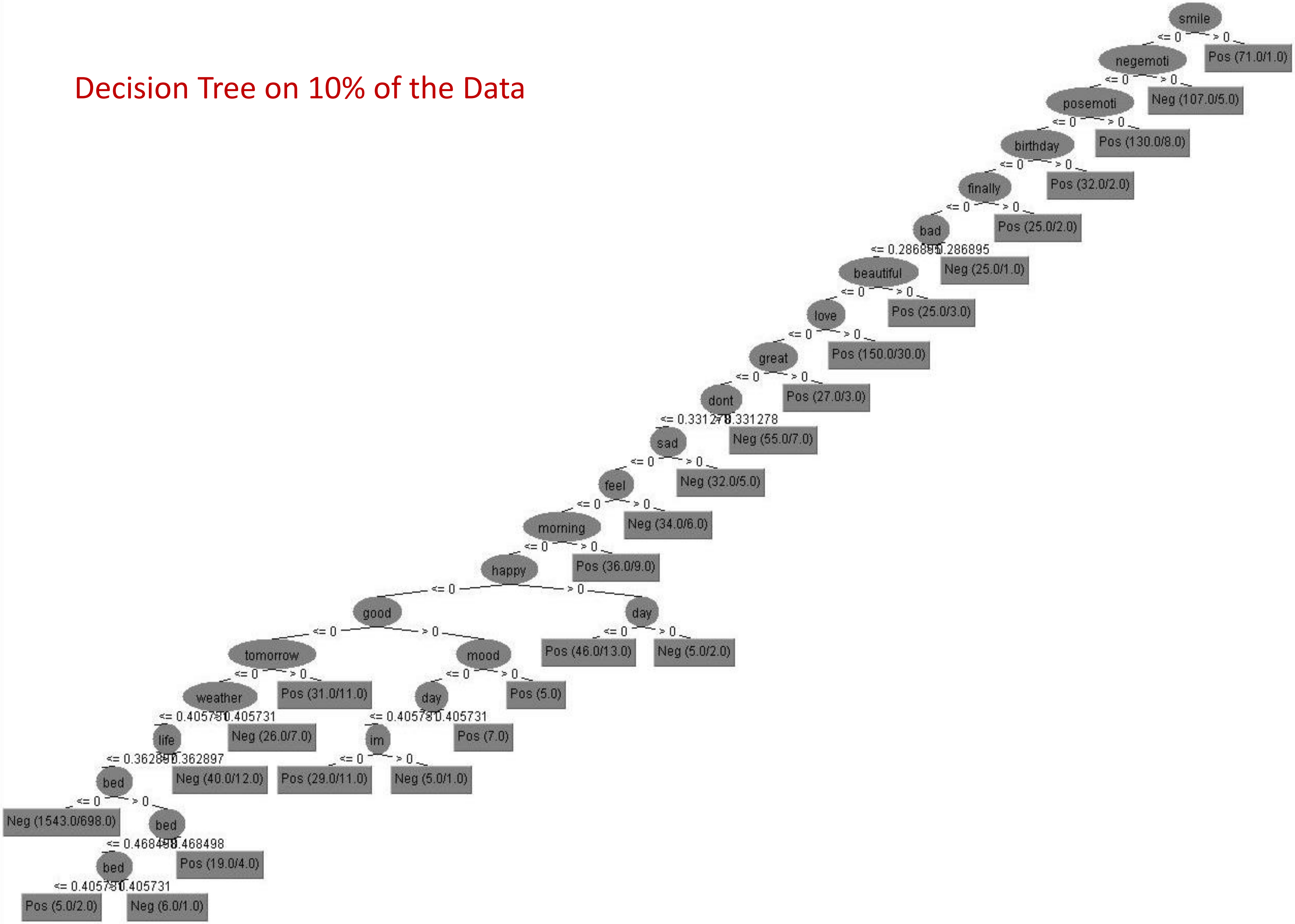
- *Neutral Cluste*r: days,work,tomorrow, im,back,home.

Classification(Subset of Data 10%)

Classifier	Accuracy	Precision	Recall	F-Measure	ROC Area
Decision Trees	0.637	0.695	0.637	0.612	0.704
Naïve Bayes	0.6727	0.709	0.673	0.66	0.759
SVM	0.7347	0.747	0.735	0.732	0.737



Decision Tree on 10% of the Data



```
smile <= 0
|   negemoti <= 0
|   |   posemoti <= 0
|   |   |   birthday <= 0
|   |   |   |   finally <= 0
|   |   |   |   |   bad <= 0.286895
|   |   |   |   |   |   beautiful <= 0
|   |   |   |   |   |   |   love <= 0
|   |   |   |   |   |   |   |   great <= 0
|   |   |   |   |   |   |   |   |   dont <= 0.331278
|   |   |   |   |   |   |   |   |   |   sad <= 0
|   |   |   |   |   |   |   |   |   |   |   feel <= 0
|   |   |   |   |   |   |   |   |   |   |   |   morning <= 0
|   |   |   |   |   |   |   |   |   |   |   |   |   happy <= 0
|   |   |   |   |   |   |   |   |   |   |   |   |   |   good <= 0
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   tomorrow <= 0
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   weather <= 0.405731
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   life <= 0.362897
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   bed <= 0: Neg (1543.0/698.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   bed > 0
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   bed <= 0.468498
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   bed <= 0.405731: Pos (5.0/2.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   bed > 0.405731: Neg (6.0/1.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   bed > 0.468498: Pos (19.0/4.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   life > 0.362897: Neg (40.0/12.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   weather > 0.405731: Neg (26.0/7.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   tomorrow > 0: Pos (31.0/11.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   good > 0
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   mood <= 0
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   day <= 0.405731
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   im <= 0: Pos (29.0/11.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   im > 0: Neg (5.0/1.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   day > 0.405731: Pos (7.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   mood > 0: Pos (5.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   happy > 0
|   |   |   |   |   |   |   |   |   |   |   |   |   |   day <= 0: Pos (46.0/13.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   day > 0: Neg (5.0/2.0)
|   |   |   |   |   |   |   |   |   |   |   |   morning > 0: Pos (36.0/9.0)
|   |   |   |   |   |   |   |   |   |   |   feel > 0: Neg (34.0/6.0)
|   |   |   |   |   |   |   |   |   |   sad > 0: Neg (32.0/5.0)
|   |   |   |   |   |   |   |   |   dont > 0.331278: Neg (55.0/7.0)
|   |   |   |   |   |   |   |   great > 0: Pos (27.0/3.0)
|   |   |   |   |   |   |   love > 0: Pos (150.0/30.0)
|   |   |   |   |   |   beautiful > 0: Pos (25.0/3.0)
|   |   |   |   |   bad > 0.286895: Neg (25.0/1.0)
|   |   |   |   finally > 0: Pos (25.0/2.0)
|   |   |   birthday > 0: Pos (32.0/2.0)
|   |   posemoti > 0: Pos (130.0/8.0)
|   negemoti > 0: Neg (107.0/5.0)
smile > 0: Pos (71.0/1.0)
```

Experimentation Full Dataset Using Naïve Bayes

Correctly Classified Instances	18964	62.7989 %
Incorrectly Classified Instances	11234	37.2011 %
Kappa statistic	0.2577	
Mean absolute error	0.372	
Root mean squared error	0.6085	
Relative absolute error	74.4068 %	
Root relative squared error	121.695 %	
Coverage of cases (0.95 level)	63.3155 %	
Mean rel. region size (0.95 level)	50.5315 %	
Total Number of Instances	30198	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.319	0.061	0.841	0.319	0.463	0.329	0.781	0.759	Neg
	0.939	0.681	0.578	0.939	0.715	0.329	0.779	0.750	Pos
Weighted Avg.	0.628	0.370	0.710	0.628	0.589	0.329	0.780	0.754	

=== Confusion Matrix ===

a	b	<-- classified as
4839	10317	a = Neg
917	125	b = Pos

Experimentation SVM on Full Dataset

Correctly Classified Instances	29692	78.6585 %
Incorrectly Classified Instances	8056	21.3415 %
Kappa statistic	0.5531	
Mean absolute error	0.2134	
Root mean squared error	0.462	
Relative absolute error	43.6811 %	
Root relative squared error	93.4678 %	
Total Number of Instances	37748	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.663	0.122	0.8	0.663	0.725	0.77	Neg
	0.878	0.337	0.779	0.878	0.826	0.77	Pos
Weighted Avg.	0.787	0.246	0.788	0.787	0.783	0.77	

=== Confusion Matrix ===

```
    a     b   <-- classified as
 10620  5401 |     a = Neg
  2655 19072 |     b = Pos
```

SVM Weka Results(Balanced Dataset)

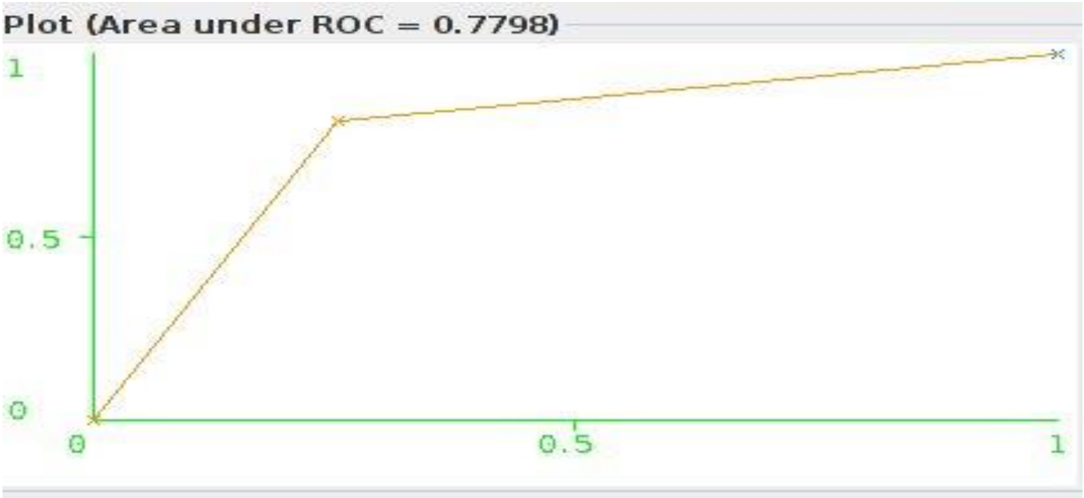
Correctly Classified Instances	29822	79.0029 %
Incorrectly Classified Instances	7926	20.9971 %
Kappa statistic	0.58	
Mean absolute error	0.21	
Root mean squared error	0.4582	
Relative absolute error	41.9945 %	
Root relative squared error	91.6455 %	
Total Number of Instances	37748	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.823	0.243	0.773	0.823	0.797	0.79	Neg
	0.757	0.177	0.809	0.757	0.783	0.79	Pos
Weighted Avg.	0.79	0.21	0.791	0.79	0.79	0.79	

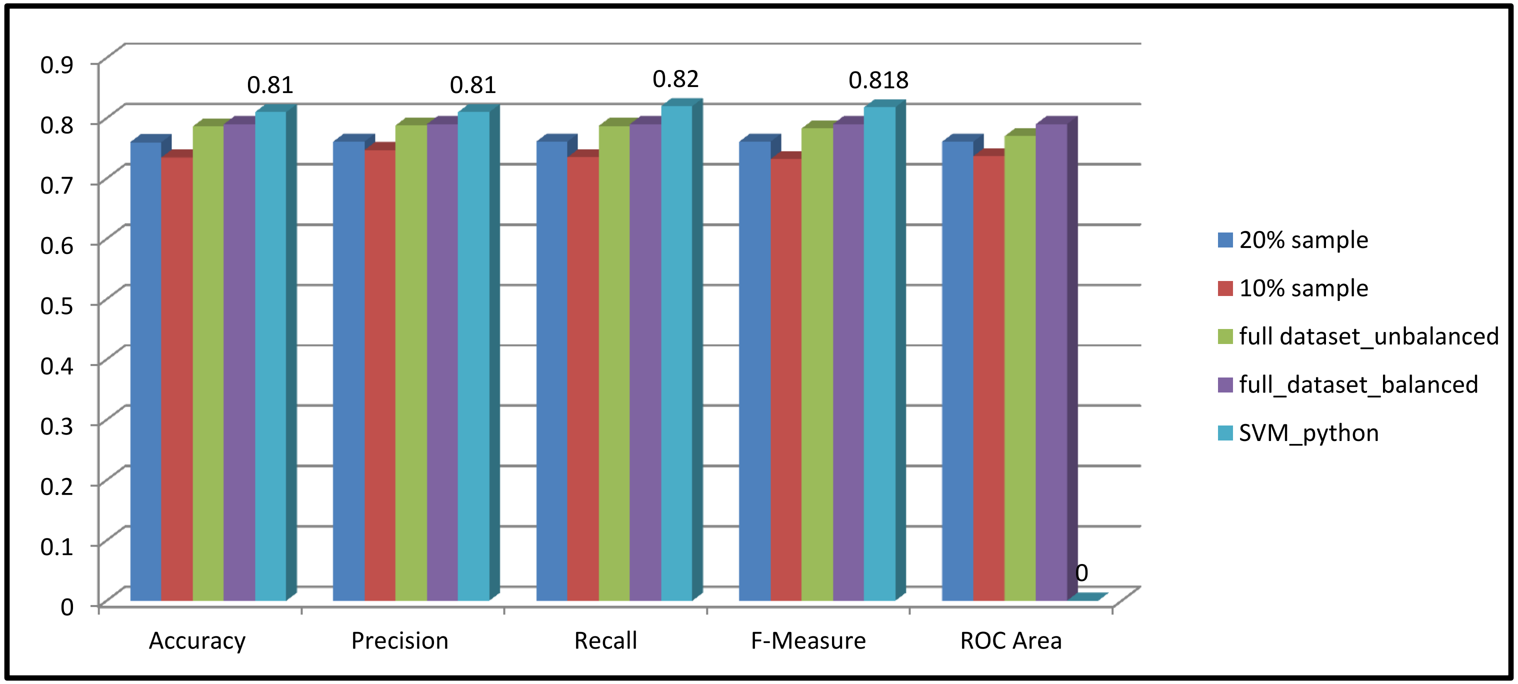
=== Confusion Matrix ===

a	b	<-- classified as
15564	3356	a = Neg
4570	14258	b = Pos



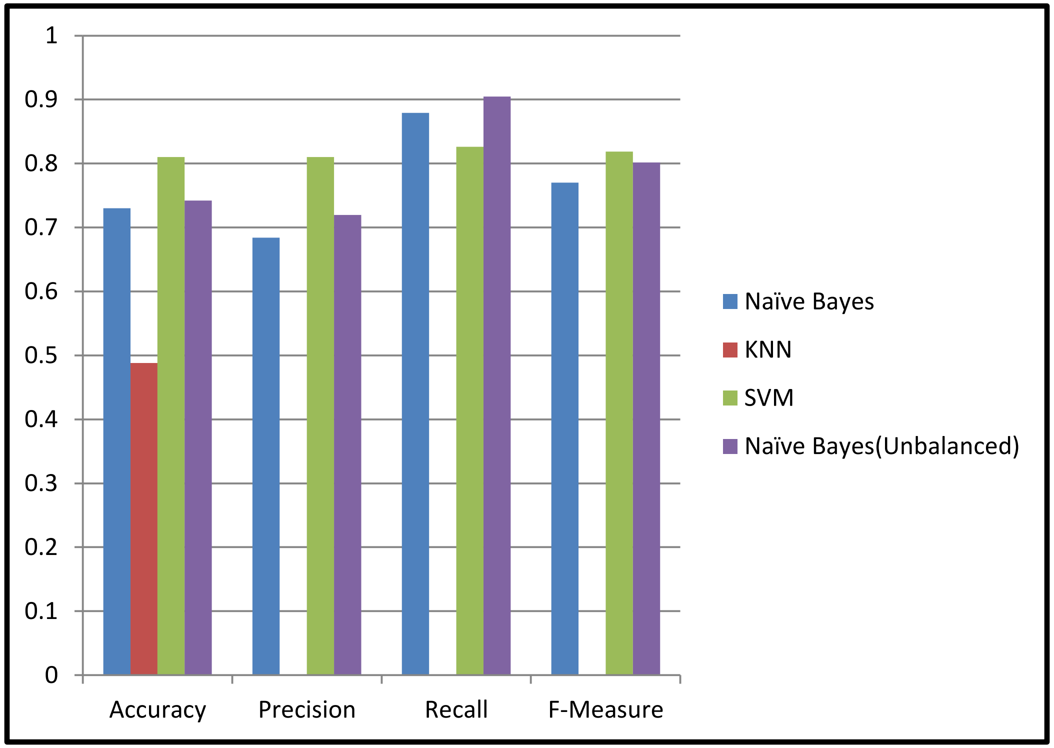
Results Using SVM

Clasifier	Accuracy	Precision	Recall	F-Measure	ROC Area
20% sample	0.76	0.761	0.761	0.761	0.761
10% sample	0.7347	0.747	0.735	0.732	0.737
full dataset_unbalanced	0.786	0.788	0.787	0.783	0.77
full_dataset_balanced	0.79	0.79	0.79	0.79	0.79
SVM_python	0.81	0.81	0.82	0.818	-



Classification(Full Dataset using Python)

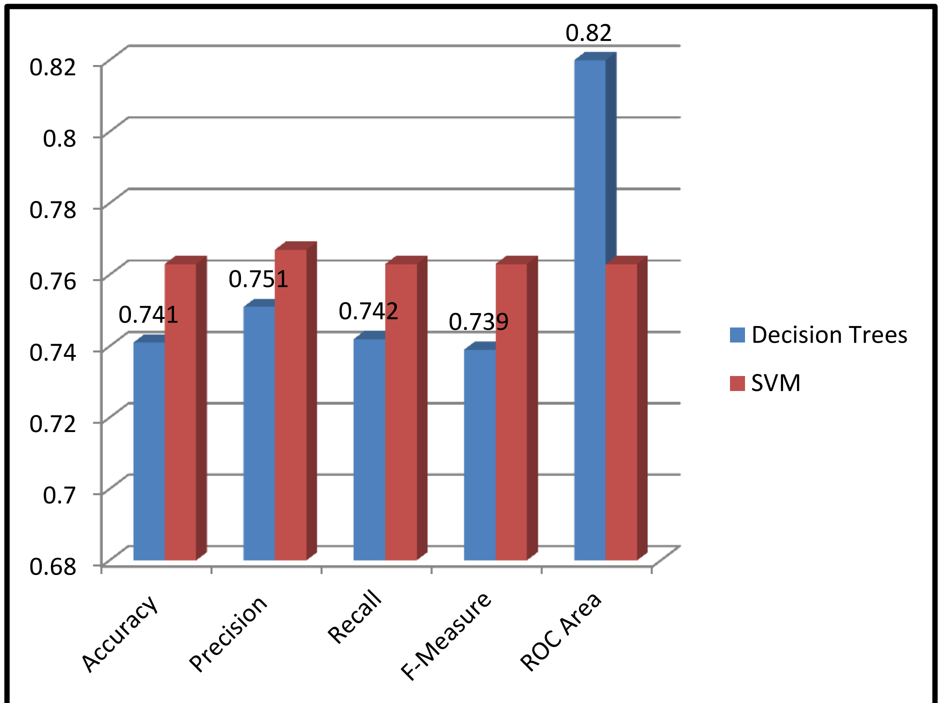
- Number of Negative Tweets: 16021
- Number of Positive Tweets: 16805
- number of documents: 32826
- number of words: 25098
- number of words (average): 5.33244988728



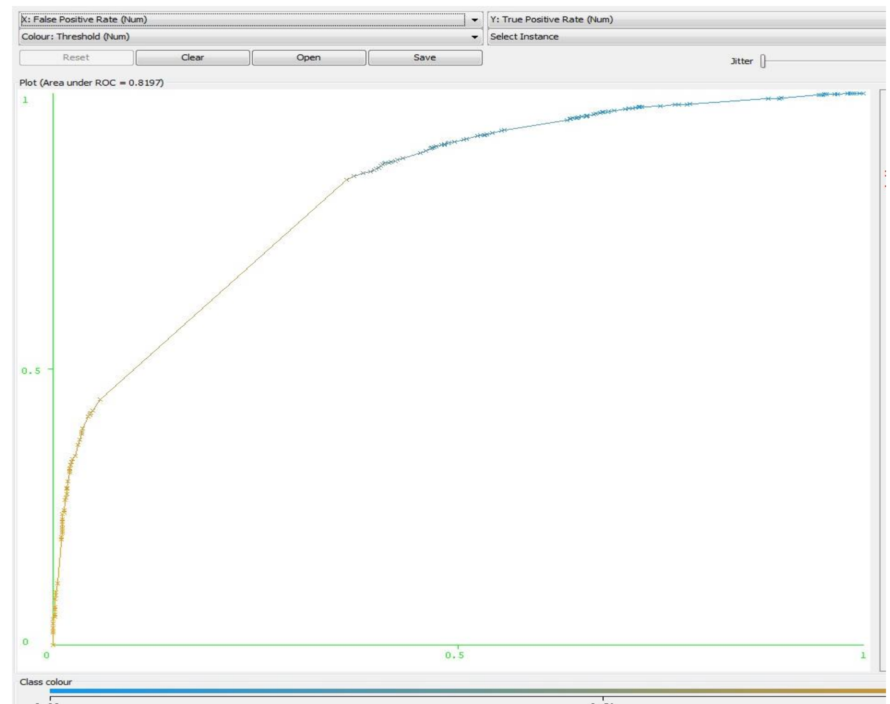
Classifier	Accuracy	Precision	Recall	F-Measure	Parameters
Naïve Bayes	0.73	0.684	0.879	0.77	
KNN	0.488	-	-	-	K=20,top300
SVM	0.81	0.81	0.826	0.8187	Linear Kernel
Naïve Bayes	0.742	0.7197	0.9046	0.8016	Unbalanced Data

Results Using Binary Model for Documents

Classifier	Accuracy	Precision	Recall	F-Measure	ROC Area
Decision Trees	0.741	0.751	0.742	0.739	0.82
SVM	0.763	0.767	0.763	0.763	0.763



Decision trees vs SVM on Binary model



ROC curve for Decision Trees

Test Data Set

Preprocessed Tweet	Class
in london again cant wait to see my girlfriend negemoti	0
negemoti	0
days till prom still no date	1
bored of this focusing on my work plan already gone weeks without a drop of alcohol and ive had enough passmethejd	0
mins ago i was crying because i didnt wanna go work now im crying because my company has shut down and i dont have a job	0
annoo i want to go soo frickin bad shitweather	0
heady highminded lovers of pleasures more than lovers of god sad lonely Christians	1
having a form of godliness but denying the power thereof from such turn away sad lonely Christians	1
and everyday it feels like im losing you all over again missyousomuch	0
listening to magic makes me tour depressed even though they didnt sing it waa	0
prototype proton supported by sidney samson	1
heart this once again robs working with amazing actors and director mtts	1
heart squee vermont in one week for work of course but it still feels like a mini vacation to me craftbeer	1
posemoti heart sums up my whole mood	1
weeks from today ill be going home yay excited	1
fallinhard yourthebest	1
whole days to myself	1
followers on tumblr	0
on my math achievement test today	1
birthday prezies from my daughter posemoti my first chane	1

Conclusion

- In this project, a proof of concept was implemented aimed at detecting emotions from tweets.
- The ability of SVM to classify high dimensional data was evident by it obtaining an accuracy of ~81% on 10-fold crossvalidation on the entire corpus, Naïve bayes was able to produce an accuracy of only ~74%, while Decision trees was took an enormous amount of time to compute and had to ultimately be shut down.

- When considered the Binary model for documents, the Decision tree algorithm performed much better than its counterparts. Its accuracy rose to ~74% from its previous value of ~68%. This is because the decision tree algorithm works much better on binary and nominal values than continuous values.

- Although SVM was able to produce high accuracy on this data if a tweet contains words which the model has not yet seen , its performance cannot be . Also, the same words could be used to denote very different meanings and emotions, for example people use the term sad and happy in the same tweet. Also the model is currently not equipped for identifying Neutral tweets. This would be an interesting task for the future work.