Datasets:

Finding datasets was one of the challenging tasks we faced in the entire process of development of our system, as not much work was done in emotion detection from text as compared to sentiments.

However some work was done by psychologists on classification of emotions, where they had used sentences and words that were classified in to the basic types of emotions.

Some of our findings were,

- 7652 Phrases classified into 7 basic emotions (ISEAR Dataset)

  (Joy, Anger, Fear, Disgust, Sadness, Guilt, Shame)

  https://github.com/bogdanneacsa/tts-master/blob/master/ISEAR/DATA.csv

- 1542 Words Classified into 6 basic emotions (Mining Twitter with R)

  (Joy, Anger, Fear, Surprise, Sadness, Disgust) https://sites.google.com/site/miningtwitter/questions/sentiment/sentiment

- Later we were in a situation to generalize the basic emotion set considered into 3+ve and 3-ve emotions (Joy, Surprise, Love; Anger, Sadness, Fear)
- We didn't have any apt data available for Surprise as an emotion
- This is when we came to know about a project named "We Feel" and the "We Feel Fine" API

  http://wefeel.csiro.au/

  http://wefeelfine.org/

"We feel fine" API:

- This was our temporary solution to the dataset thirst
- We had **2178 emotional words** from we feel fine API
- Given a word the API returns up to 1500 sentences that have that queried word in it
- All those 2178 words were classified in to 6 emotions
- Now we had 2178 * 1500 = **5,42,500 sentences** emotionally classified
- Drawbacks :
  - This API was just a keyword based extractor
  - So this again led to noisy dataset - Due to falsely classified sentences
  - Ex. " I am not happy " was classified under Joy as it had the word "Happy"

- We tried to remove this noise from the dataset and tried out the implantation
- Apart from this we used a 3M word dataset from Google Word2Vec Tool's website for KMeans Clustering

http://word2vec.googlecode.com/

http://word2vec.googlecode.com/svn/trunk/

- (American National Corpus 14M words)OANC was also used for KMeans Clustering
- When it came to LDA implementation we just needed unclassified documents that had some emotion in them, so, we used all datasets normally used for sentiment analysis including

Data Set/Source	
BUILDING RESOURCES	ISEAR
BUILDING RESOURCES	Sentiment dictionaries
EMOTION DETECTION	Live Journals Blogs, Text Affect, Fairy tales, AnnotatEmotion Detection Blogs
EMOTION DETECTION	ISEAR, Emotinet
EMOTION DETECTION	Enron Email corpus
TRANSFER LEARNING	MPQA, RIMDB, CHES
TRANSFER LEARNING	Blogspot, Flicker, youtube, CNN-BBC
FEATURE SELECTION	amazon.com
SENTIMENT ANALYSIS	automotvieforums.com
SENTIMENT ANALYSIS	CNETD
SENTIMENT ANALYSIS	amazon.com, epinions.com, blogs, SNS
SENTIMENT ANALYSIS	ebay.com, wikipEmotion Detectionia.com, epinions.com
SENTIMENT ANALYSIS	amazon.com
SENTIMENT ANALYSIS	amazon.com
SENTIMENT ANALYSIS	Twitter
SENTIMENT CLASSIFICATION	IMDB
SENTIMENT CLASSIFICATION	IMDB, Amazon.com
SENTIMENT CLASSIFICATION	convinceme.net
SENTIMENT CLASSIFICATION	2000-SINA blog data set, 300-SINA Hownet lexicon
SENTIMENT CLASSIFICATION	Reuters 21578
SENTIMENT CLASSIFICATION	amazon.com
SENTIMENT CLASSIFICATION	Twitter

Data Set/Source	
SENTIMENT CLASSIFICATION	Twitter
SENTIMENT CLASSIFICATION	epinions.com