Udacity Capstone Project Proposal

Project background:

For some business solution company, it's emerging task to help some of their clients to perform social listening program that enable clients to have idea what people are talking about, what topic they focus on,as well as their sentiment trend towards these topics and comments. These social listening efforts would help them improve service, update product spectrum. This is NLP task in machine learning fields and become more and more important recently.

Project statement:

1) Topic modeling :

Using data offered to explore what caller's main stream topic is,if it's possible to segment transcripts into different categories.

2) Sentiment analysis:

Building customized sentiment analyzer, compare accuracy or precision or recall metric.

Datasets and inputs:

Sentiment 104,[link](http://help.sentiment140.com/for-students) this data is a company that has made their training data available to the public on their site. It is a tool that’s typically used for analyzing sentiments around specific topics, brands, or products that are talked about on Twitter.
This data can be obtained from either standford or google link.
The data is a CSV with emoticons removed. Data file format has 6 fields:

0 - the polarity of the tweet (0 = negative, 2 = neutral, 4 = positive)
1 - the id of the tweet (2087)
2 - the date of the tweet (Sat May 16 23:58:44 UTC 2009)
3 - the query (lyx). If there is no query, then this value is NO_QUERY.
4 - the user that tweeted (robotickilldozr)
5 - the text of the tweet (Lyx is cool)

Solution statement:

I will try use LDA algorithm on topic modeling part and build customized sentiment model to compare with labeled data and benchmark model for sentiment analysis.

Benchmark model:

1) Topic modeling: Pick optimized number of topic,since this pick is quite subjective, so LDA model with different K number of topics would be benchmark model for this case.

2) Sentiment analysis: Use pretrain sentiment analyzer such as textblob, vader to generate sentiment score and category, then compare performance of different models.

Evaluation metric:

Topic modeling is unsupervised learning,and topic selection is relatively subjective, so the model chosen are based on topic separation and topic semantic meaning.During this process, we consider using coherence and perplexity or some variants of perplexity. If the topic distribution is exclusively distributed and topic keywords are more human interpretable thus it can be considered a useful model.

Sentiment analysis can be evaluated upon model accuracy/precision/recall.

Project design:

First this data would go through text preprocess step since it's NLP task, after proper preprocess, text would be transformed into tokens and build corresponding corpus and id2word dictionary. We can compute the coherence score and perplexity to determine the optimal number of topics then if we test the topic words are separatable and understandable, we can distribute topic or likelihood of topic to each single transcript. Since we already obtain the text token, we can feed this into pretrained sentiment analyzer to generate one of the sentiment scores, afterwards, we can build RNN model (since RNN is best to deal with sequence of texts) and feed RNN with preprocessed text to generate score (or category) , I personally prefer category since we have labeled category from raw data, which we can use to validate performance of different sentiment models.

If RNN model is satisfying, then next is to deploy this model on a web app and test with some sample.

For real user case, future monitor and model update should also be covered as it's critical for social listening towards different events.