Project Proposal – Capstone Project 2

Compared to other traditional approaches such as voice call and email marketing, Twitter has become the new channel for businesses to interact with their customers. This new social networking service allows people to share their experience about company, product or service in a more effective way.
This feedback guides improvements of the customer experience and can empower positive change in any business — even (and especially) when it's negative.

The objective of this project is to apply sentiment analysis, the most common text classification tool, to analyse an incoming message and tell whether the underlying sentiment is positive, negative or neutral. The dataset contains tweets on US Airline of February 2015 classified in positive, negative and neutral tweets. The negative tweets are also classified by the negative reason.

First of all, with the help of exploratory data analysis, we will look into this dataset to find out the best and the worst airlines and identify what is the root cause for negative cases. After we have gained the basic understanding about the dataset, we can start the core part of this project: machine learning algorithms application. In the process of text preprocessing, we will take several steps to complete lowercase transformation, removing stop words, removing special characters and stemming.

We cannot work with text directly when using machine learning algorithms. Instead, we need to convert the text to numbers. A simple and effective model for thinking about text documents in machine learning is called the Bag-of-Words Model, or BoW. We can use the CountVectorizer method of NLTK package to calculate the frequency that each word occurs in the collection of documents. To better evaluate the importance of a word, we then use TF-IDF to consider the word frequency. This is a technique to quantify a word in documents, we generally compute a weight to each word which signifies the importance of the word in the document and corpus.

With all text converted to number, we can apply traditional machine learning algorithms to build classification models. Naïve Bayes multinomial model is good option to start with if we assume that the occurrence of each individual is not related with each other, i.e. mutually independent. To better evaluate the model

performance, we will also try other machine learning algorithms, such as logistic regression, SVM or decision tree.

Human language is more complex than so-called Bag-of-Words because we have to understand the meaning of word in certain “context”. This is why word embedding comes into play. By specifying window size of context and calculating the probability that each word occurs, we will map the word representation of Bag-of-Words into lower dimensional vectors. The word2vec of Gensim package can help us in this regard.

Deep Learning has revolutionized machine learning and data science in general, and NLP is no exception. After we complete word2vec, we will be using neural network to build classification model. Keras, built on the top of Tensorflow, can help us build RRN and LSTM. We will also use dropout and recurrent layer stacking method to reduce overfitting and improve modelling performance.

The deliverables of this project will include code, a paper, and a slide deck.