Machine Learning Engineer Nanodegree

Capstone Project

Peng Wu

April 2nd, 2018

I. Proposal

Domain Background

As personal computers and mobile computing platforms have become virtually ubiquitous over the past decade, people have largely adopted the use of online documents as a way of expressing their feelings and thoughts on the world around them. As this online method of expression has continued to gain traction, it has become a valuable task to attempt to categorize these submissions for the purpose of gaining information about public sentiment. There are many reasons why an individual or organization would have an interest in public sentiment, but one important motivation is commercial: by being able to gain an understanding of how members of the public are reacting to a product or announcement, companies and retailers can adapt their behaviors and business strategies to increase positive sentiment surrounding their products, while minimizing sources of negative sentiment.

Problem Statement

The primary challenge inherent to processing online submissions is at the very thing that makes these submissions so attractive as a representation of public sentiment: the amount of data involved is often extremely large. In order to process text information in these quantities, it has become necessary to turn to the use of automated processing systems. Although the task of automatically extracting information from text, broadly referred to as natural language processing, has resulted in the establishment of procedures for classifying text based on objective markers such as subject matter,

procedures for classifying text based on subjective markers of sentiment are yet unestablished. While humans are well conditioned to interpret emotion from speech and writing, there are no obvious and well established markers of sentiment that can easily be identified by a computer. Several heuristic methods have been suggested for identifying features or markers of a particular sentiment, and various papers have been published detailing attempts at performing automated sentiment analysis on large bodies of online text.

Datasets and Inputs

The first dataset we use is from nltk corpus “movie reviews”. This corpus is from paper of Bo Pang and Lillian Lee[1]. It contains 2k movie reviews with sentiment polarity classification. The reason that I used it in this project because it is processed and easy to used as in nltk corpus. It has 1K positive reviews and 1K negative reviews. Most the reviews are long, have more than 1000 words.

Another corpus used is from this website[2]. Just as it described in the website, our main goal is to do Twitter sentiment. Usually the twitter is very short, if we use the corpus in nltk, the result is not what we will expect. Hence, short movie reviews maybe better to fit for the twitters from Twitter.

Solution Statement

In this paper, we will explore some of those feature extraction methods as well as machine learning methods that might be applied to them to successfully perform automated textual sentiment analysis.

Our final goal is to create a model that can do Twitter sentimental analysis and the tasks involved are the following:

- Process the short movie data
- Train the traditional classifier that can determine the sentiment of a sentence
- Train a model using a deep learning method that can determine the sentiment of a sentence
- Make a Twitter application that extract the twitters and analyse its sentiment by a specific word.

Benchmark Model

From this website [3] , I find that it's accuracy is around 67%. Using the same corpus from this website, so I hope get a better accuracy. Using traditional classifiers, such as NaiveBayesClassifier, BernoulliNB_classifier, LogisticRegression_classifier and so on, the accuracy I hope will be around 70%. Another way is used by deep learning, I wish the accuracy will be around 75%.

Evaluation Metrics

Accuracy is the common metric for binary classifiers.[4] .

We can clearly see that the accuracy means from the figure: the proportion of the total number of predictions that were correct. In general, we consider one of the defined metrics in the figure. For example, in a pharmaceutical company, they will be more concerned with minimal wrong positive diagnosis. So, they will be more use the metrics of Specificity. In our binary classifier problem, we need to predict if the sentences if classified in the correct sentimental value. Hence, we only concerned the metrics of Accacy.

Confusion Matrix		Target			
		Positive	Negative		
Model	Positive	a	b	*Positive Predictive Value*	a/(a+b)
	Negative	c	d	*Negative Predictive Value*	d/(c+d)
		Sensitivity	*Specificity*	**Accuracy** = (a+d)/(a+b+c+d)	
		a/(a+c)	d/(b+d)		

Project Design

The workflow of the project is as followed:

1. Data exploration

First, we should have a thorough understanding of the data. Some questions should be solved. What kind of the dataset? How many of them? What kind of words in the corpus? What the most common words in the corpus?

2. Data processing

This is an important part. Should know how to extract the features as the data are words which is different from the number dataset.

3. Train the model using traditional classifier such as NaiveBayes.
4. Train the model using deep learning, such as CNN. Choose the paraments that can get a better accuracy.
5. Apply our model to do Twitter sentimental analysis and make a visualization to show the tendency.