Machine Learning Engineer Nanodegree

Capstone Proposal

Ahmed Bahgat

(ahmedbahgat64@yahoo.com)

February 3rd, 2019

Proposal

Abstract

The project proposal is to create a ML model which accurately apply Sentiment analysis on many movie reviews and can easily classify between positive and negative reviews. This project is a small practical example on what NLP can do and how it can easily ease our daily missions, the field still need more concentration to be able to simulate the human being, However what the field reach till now can be considered as a revolutionary step into the future.

In this project,after cleaning the text data and collecting them, I'll try to reach my goal at first using classical NLP methods (Tfidf—NLTK), then compare what I'll reach with what the owner of the data reaches and publish on the paper corresponding to this data. Then I'll see if I can apply modern methods to reach the same goal like Word2Vec- GloVe and compare between them but I'm not sure If can do this or not.

Keywords: Supervised learning, scikit learn, NLTK, CountVectorizer, Tfidf, word embedding, word2vec,

Domain Background

Natural Language Processing is one of the most promising field in machine learning where already much development has taken place and is currently used in real world application. One of the most important applications based on it and widely used is Chat Bots, and although it may seems scary to tell someone that the one you kept asking and answering you back with exactly what you want is not one of us! it's just some line of codes can accomplish some simpler tasks like sentiment analysis, entity recognition..etc to do the main task which is helping you, The customer is the focal point of any business and is directly proportional to growth of

the business. One area which affect the customers the most is the services provided and the delay in them. In order to reduce the waiting period as well as doing the mission efficiently, the business can use the machine learning algorithms. It will not only help the customers but also let businesses to focus on other things which require regular human intervention. My personal motivation is this fact only that the very first use of machine learning in my mind is to reduce the frequency of redundant tasks, So that we can invest our times in better work and this problem addresses exactly that.

The link to my datasource is: http://www.cs.cornell.edu/people/pabo/movie%2Dreview%2Ddata/

Problem Statement

Applying Sentiment Analysis on some text to see how we can teach the Computer to feel the same when he read the same text data, is one of the oldest and earlier necessary task, which the linguists, ML engineers has been facing, And understanding what the human say, what he feels, and what he wants to say with writing those words can lead us to a brand new era, In this project I will work on 1000 positive and 1000 negative processed movies reviews. Introduced in Pang/Lee ACL 2004 Paper. Released June 2004. and with the help of the ML methods we have, that I mentioned above I will try to correctly classify as correctly as possible between those two categories and compare the results I'll get with what they've got and then try to apply newer techniques on the same data.

As input we are provided with 2000 different reviews differs in the length, the polarity and even the style as they're real data so each review is probably written earlier by different person.

Since our task is to distinguish between two classes, it is essentially a classification task

Data sets and Inputs

review_polarity.tar.gz: contains this readme and data used in the experiments described in Pang/Lee ACL 2004.

Specifically:

Within the folder "txt_sentoken" are the 2000 processed down-cased text files used in Pang/Lee ACL 2004; the names of the two sub-directories in that folder, "pos" and "neg", indicate the true
classification (sentiment) of the component files according to our automatic rating classifier.
File names consist of a cross-validation tag plus the name of the original html file. The ten folds used in the Pang/Lee ACL 2004 paper's
experiments were:
fold 1: files tagged cv000 through cv099, in numerical order
fold 2: files tagged cv100 through cv199, in numerical order
...

fold 10: files tagged cv900 through cv999, in numerical order Hence, the file neg/cv114_19501.txt, for example, was labeled as negative, served as a member of fold 2 Each line in each text file corresponds to a single sentence, as determined by Adwait Ratnaparkhi's sentence boundary detector MXTERMINATOR.
As this was a Real data used earlier in a research paper in the field so it's expected to be accessible. They can be obtained here.

Solution Statement

We want to understand the relationship between the word and the other words in its context, as we can say that each word is a dependent feature, so we want to study carefully the long term relations and also the short term for the words in the same context so that we can hopefully able to reach the true decision, Although it looks straightforward for the human being to do, but unfortunately it isn't so for the computer to teach it how it can accomplish this. There are many frequent words(features) which due to curse of dimensionality, may result in overfitting, so we may have to reduce the features by processing the text data and see which we should keep, and which we should get rid of, also we will see that there's punctuation which should be removed as it'll not provide us more information so it's not important in this task, Then as the only thing our computers can handle is numerical information so we need to change our text data first to an appropriate numerical representation that still capture the related information about its context, finally we can choose between the various classifiers to see which one will give us the best result so we can enhance it using Grid Search technique.

- I'm not sure If our data will need dimensionality reduction techniques like PCA, ICA or not so I'll see what can I do?
- In addition to our main task, I'll try to accomplish the same results using newer methods like Word2Vec but also I'm not sure If I'll be able to do it or not but I'll hardly try for the sake of gaining more knowledge about them.
-

Benchmark Model

As this data was first used in Bo Pang and Lillian Lee,``A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts'' Paper, Proceedings of the ACL, 2004. a benchmark model would be their best score they accomplished in this specific task on this specific version of data which comes in at 86.4% as their best result using NB then comes SVM with 86.15%

note that we would use a part of training data as testing data.
Then we will build our own classifier So we can compare our model with the benchmark model hosted by the paper. A personal goal would be to be near their values even higher if I could

Evaluation Metrics

At the end, we can easily see that this task is a classification task so The model prediction for this problem can be evaluated in several ways. Since I'm not only

interested in precision or the recall but both, I think it’s more appropriate to use f1-score so our model can keep both values in mind and not bias to just one.

Project Design

First method would be using classic NLP methods. To do that , we will read the data, then normalize, and clean , and that will happen through various steps like remove stop words, punctuation, see if using stemming..etc will help or not, then convert the text representation to the appropriate numerical representation so our classifier can easily deal with. Then we will test our model on our test set. So finally we can enhance our classifier with Grid Search, Second method would be to try if we can improve our results using state of the art of NLP using newer word embedding techniques like Word2Vec.

Tools and Libraries used: Python, Jupyter Notebook, pandas, scikit learn, NLTK, string, gensim (in case of Word2Vec). Other libraries will be added if necessary.

References

[1] “A Sentimental Education: Sentiment AnalysisUsing Subjectivity Summarization Based on Minimum Cuts", Proceedings of the ACL, 2004”
http://www.cs.cornell.edu/home/llee/papers/cutsent.pdf

[2] Supervised Learning Wikipedia page

https://en.wikipedia.org/wiki/Supervised_learning