Machine Learning Engineer Nanodegree Capstone Proposal
Maksim Lebedev,
November 2018

Yelp Open Dataset Sentiment Analysis: supervised and unsupervised approach

Domain Background

Yelp's website, Yelp.com, is a crowd-sourced local business review and social networking site. Its user community is primarily active in major metropolitan areas The site has pages devoted to individual locations, such as restaurants or schools, where Yelp users can submit a review of their products or services using a one to five star rating system.
Yelp Dataset is subset of businesses, reviews and user data.

Reviews and comments are one of the most important data sources that business has, especially in a customer-centric, digital market. There are lots of insights that we can get from these information - it could be sentiments, topics extraction, as well as clustering and categorization.
This is a natural language processing project.

Problem Statement

Comments or reviews, which people are leaving in the Internet are really important, especially now, when we live in a digital era. Lots of companies are not using this source of information, some - because they do not think about it, some because they do not have resources. Some companies are still doing that “old” way - when people are reading comments and categorizing them, identifying sentiment as well, which makes this process slow, not giving a business an opportunity to react fast or anticipate some changes in client’s behaviour.
Those companies that are doing this analysis - sometimes spending time and money to prepare datasets and label this datasets. Most successful machine-learning projects are using supervised-learning algorithms, however there are opportunities to use unsupervised-learning algorithms to help addressing some of the issues.

Having labeled dataset is very good, but in most of the companies, it is not the case so, in this project I want to compare usage of supervised learning algorithms with unsupervised learning (or semi-unsupervised).

Datasets and Inputs

The datasets are provided by Yelp and could be found here (https://www.yelp.com/dataset)
In this project I will use:

- Business.json: contains information about businesses. Their location, attributes and categories. File contains more 170k businesses.

- Review.json: Contains full review text data including the user_id that wrote the review and the business_id the review is written for. There are more than 5 million reviews stored in the file.
- User.json: ser data including the user's friend mapping and all the metadata associated with the user.
- Tip.json: Tips written by a user on a business.

Solution Statement

For supervised learning - I will be using word2vec or doc2vec and then compare several algorithms: naive bayes, logistic regression and neural networks.
For unsupervised sentiment analysis: use word2vec or doc2vec together with SentiNet, also look at OpenAI method described here (https://github.com/openai/generating-reviews-discovering-sentiment)

Benchmark Model

Neural networks are working with the best accuracy around 80% (for instance: https://www.kaggle.com/bsivavenu/nlp-using-glove-and-spacy). In my project I aim to achieve this accuracy on supervised learning side,with word embeddings and neural networks.

There are unsupervised models that are working with accuracy higher that 80%, as mentioned in this research paper: http://www.aclweb.org/anthology/P14-1146
For unsupervised part - I will use results of Naive Bayes supervised model as a benchmark.

Evaluation Metrics

As Yelp dataset contains sentiment they will be used to calculate accuracy. For the simplicity reviews with rating higher or equal to 4 - will be positive, 3 - neutral, less than 3 - negative.
For supervised learning I will split dataset into three - training, test and validation.
Precision,recall and F1 score would be used for model evaluation as well
For unsupervised part: it will be accuracy for the sentiment.

As it is a comparison exercise: another metric is time spent to both, as one of the goal of this project is to understand - what could be done with unsupervised learning, and how small businesses could benefit from it. I want to evaluate how much time is needed for unsupervised learning versus supervised, where, in supervised - dataset preparation should be considered as well.

Project Design

I will start with data exploration, may be,I will need to join all 4 datasets. I will also need to choose the data to work with, as working with 5mlns of rows might be slow. I will also need to make sure that negative reviews are represented in training, test and validation dataset.

First - supervised part: expecting that around 60% of my time I will spend on data cleansing and preparation, I will use - nltk/spacy or keras for pre-processing part. I want to use word embeddings, so will evaluate two approaches word2vec and doc2vec. The reason is that I want to understand, which one will work better. Most probably, at the end it will be doc2vec, as we have more than one sentence. I will use gensim for word2vec and doc2vec part.
After I will start with simple algorithms, like Naive Bayes and Logistic Regression to check how they work for this task.
Then I will use keras for neural network creation. The architecture of NN is a product of tries, so I can not say how it will look like at the moment.
For unsupervised part: I will use word2vec and doc2vec - which will be done on the first part of the project, then I will use sentiwordnet, to get the sentiment. After checking the result against Naive Bayes.

Reference

1. https://arxiv.org/ftp/arxiv/papers/1805/1805.00352.pdf
2. http://www.aclweb.org/anthology/P14-1146
3. https://github.com/openai/generating-reviews-discovering-sentiment
4. https://www.cs.york.ac.uk/semeval-2013/task2/
5. https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/viewFile/16334/16010
6. https://arxiv.org/pdf/1408.5882.pdf