Capstone Proposal

Ashish Kumar Verma

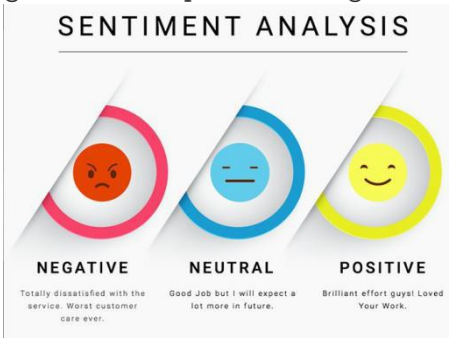
31-March-2019

Title: Sentiment Analysis.

Domain Background

Sentiment Analysis is contextual mining of text which identifies and extracts subjective information in source material, and helping a business to understand the social sentiment of their brand, product or service while monitoring online conversations.

It is the most common text classification tool that analyses an incoming message and tells Whether the underlying sentiment is positive or negative.



Sentiment analysis systems allows companies to make sense of this sea of unstructured text by automating business processes, getting actionable insights, and saving hours of manual data processing, in other words, by making teams more efficient.

Problem Statement

The objective of the task is to understand the underlying sentiment of the reviews and classify them as positive or negative.
Sentiment Analysis can be used as building block in several applications such as :
1.Computing customer satisfaction metrics :- One can get an idea of how happy customers are with your products from the ratio of positive to negative reviews about them.

2.Identifying detractors and promoters :- It can be used for customer service, by spotting dissatisfaction or problems with products.

The problem of automatically identifying underlying sentiment is difficult because of the near infinite number of permutations of words, positions, phrases and so on. It's a really hard problem. This is a well studied problem in Natural Language Processing and more recently an important demonstration of the capability of deep learning.

Data-sets and Inputs

Movie Review Dataset .

The Movie Review Data is a collection of movie reviews retrieved from the imdb.com website in the early 2000s by Bo Pang and Lillian Lee. The reviews were collected and made available as part of their research on natural language processing. The reviews were originally released in 2002, but an updated and cleaned up version was released in 2004, referred to as v2.0. The dataset is comprised of 1,000 positive and 1,000 negative movie reviews drawn from an archive of the rec.arts.movies.reviews newsgroup hosted at IMDB.

Dataset have a directory called txt sentoken with two subdirectories containing the text neg and pos for negative and positive reviews. Reviews are stored one per file with a naming convention from cv000 to cv999 for each of neg and pos. Next, let's look at loading the text data.

SOLUTION STATEMENT

In this project,I will develop a Neural Bag-of-Words + Simple MLP as well as an Embedding + CNN Model For Sentiment Analysis.

BENCHMARK MODEL

For a benchmark model,I will develop a Neural Bag-of-Words + Simple MLP.

For CNN model ,I will use word embedding while training a CNN and try different neural architectures with MaxPooling Layer and with relu activation function as well different optimizers and tune in with different no of epochs to calculate when my model performs better.

EVALUATION METRICS

Accuracy Score -
Accuracy is the ratio of number of correct predictions on the total number of input samples.

$$Accuracy = \frac{Number\ of\ Correct\ predictions}{Total\ number\ of\ predictions\ made}$$

.

Project Design.

1. Load The Data set.
2. Data Cleaning Step which includes splitting it into words and handling punctuation,etc.
3. Implement a Neural Bag-of-Words + Simple MLP and Embedding + CNN as described in the benchmark model section.
4. Try using different optimizers and different epochs to determine the better performance.
5. Check the Model Performance using Evaluation Metrics.

Links and References:
http://www.cs.cornell.edu/people/pabo/movie-review-data/review_polarity.tar.gz

http://www.cs.cornell.edu/people/pabo/movie-review-data/