Machine Learning Engineer Nanodegree

Capstone Proposal

Samujjwal Bhandari

November 2, 2018

Proposal

Sentiment Analysis from Tweets

Domain Background

Twitter ((https://twitter.com/)) is a micro-blogging platform that is being used for several purposes such as, expressing beliefs, connecting people, promoting the business, and as such. The messages that we post on Twitter are called tweets. In this project we will be focusing on tweets that can be categorized as contents with positive or negative sentiments.

Sentiment analysis is a concept that helps predicting an opinion of a user from the context being presented. The context, in general, is represented as a text as an unstructured information. By the application of sentiment analysis on such textual information the opinion about domain such as, products, services, brands, and politics can be derived. Such results of analysis can be used in many applications including marketing analysis, public relations, product reviews, product feedback, and customer service [1]. Past works in semantic analysis has explored various approaches to predict the intention of the text such as, using Naive Bayes, Maximum Entropy, SVM [2], and Recursive Neural Tensor Network [3].

Problem Statement

In this project, the sentiment analysis from tweets to find out whether the tweet has a positive or a negative sentiment is a binary classification task. There are several ways to approach the problem such as, neural networrk defined in [3], Naive Bayes and SVM defined in [2]. Also in conjunction to the learning approach, the feature creation also plays a very crucial role in text analysis. For instance part of speech tagging [2] to explore the text using language semantics can be used during feature extraction.

The goal of this project is to classify given text as a positive or negative based on the trained model using the collection of tweets labeled as postive and negative. To accomplish the goal, this project focuses on feature reduction/extraction in order to improve the outcome of learning and the use of machine learning algorithms to develop a sentiment classification model. Some details on intended feature extraction and learning approaches are discussed in later *solution statement* section.

Datasets and Inputs

This project uses the dataset described in [2], where the authors discuss how the data data was automatically created, as opposed to manual annotation of tweets. The tweets were collected with keyword search and the labeling were done based on the emoticons such as :) for the positive sentiment and :(for the negative sentiment. All the emoticons were removed from the data after labelling them.

The data is a collection of 6 tuple. Each tuple is represented by the following information in given order:

- the polarity of the tweet (0 = negative,4 = positive)
- the id of the tweet
- the date of the tweet
- the query. If there is no query, then this value is NO_QUERY.
- the user that tweeted
- the text of the tweet

There are 1.6 milliions data points in total, where 800,000 are positive tweets and the remaining 800,000 are negative tweets. Out of these 20% of the data points will be extracted for using as a test data. Finally, the remaining 1.28 millions data points will have 80-20 split to get training and validation sets.

Solution Statement

The proposed problem is a binary classification of texts. In order to develop a classification model, this project uses tweets dataset and converts them into feature data considering the concepts such as, defining unigram/multigram feature vectors and using stemmers to reduce the feature space. Once the feature is extracted we have planned to train classification models using Naive Bayes, Support Vector Machine, Logistic Regression, and Neural Network algorithms. Along with the classification models using different algorithms, this project also compares the performance of those alogrithms on classifying sentiments from the texts.

Benchmark Model

We will use the polarity lexicon from [4] and classify the tweets based on the occurrences of the words from the lexicon in the tweets. If we have equal number of occurrences of both positive and negative words we will consider the tweet to be positive.

Evaluation Metrics

The predictions from classification models are evaluted againsts the classification accuracy measure. Since the dataset for this project has a balanced distribution of positive and negative tweets (~50% for each class after uniform random sampling of training/validation sets), for this binary classification problem accuracy can be a good measure of evaluation. By definition, Accuracy = (TP+TN)/total, where TP (True Positive) is the number of tweets classified as positive that are

actually positive, TN (True Negative) is the number of tweets classified as negative that are actually negative, and total is the number of total tweets used in prediction. We will use the same accuracy measure to compare different models discussed in this proposal against the benchmark model.

Project Design

The proposed project will start with background research on solving the sentiment analysis problem so that we get the further insights on the problem domain and its nuances. As this project will be using different learning models we will perform the following steps to compare the models performances.

1. Feature reduction using techniques such as, tokenizing user identifiers and URL from tweets, word stemming, and cleaning repetitive characters appearing more than twice in words as defined in [2], for example, 'hellooooo' will get converted into 'hello'.
2. Feature vector construction using unigram and bigram models.
3. Define train/test/validation sets from the data set.
4. Implement train-validate-test pipleline to use Naive Bayes, Logistic Regression, SVM, and Neural Network algorithms.
5. Tune the parameters and evaluate the performance of each model using the accuracy metrics.

Reference

- [1] https://monkeylearn.com/sentiment-analysis/#sentiment-analysis-use-cases-and-applications
- [2] Go, A. (2009). Sentiment Classification using Distant Supervision.
- [3] Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C.D., Ng, A.Y., & Potts, C. (2013). Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. EMNLP.
- [4] https://github.com/felipebravom/StaticTwitterSent/blob/master/extra/polarity-lexicon.txt