Machine Learning Engineer Nanodegree

Capstone Proposal

Proposal

Domain Background

Millions of people express their opinion, preferences, likes, dislikes and complaints nowadays through social networks. That means that a lot of useful data is available and organisations like companies, institutions and governments can take advantage of this. Companies, for example, can get a powerful feedback, but for this to be a reality, information and knowledge must be extracted from data. This must be done as fast as possible in order organisations be able to react on time: for example, in social networks a complaint can become viral and produce a great damage if it is not kept under control.

To monitor social networks like Twitter can provide a significant advantage for companies helping them to have a better knowledge of their clients and their market. Humans are good understanding raw messages and extracting information and actionable knowledge from them, but the volume of data is so big and is generated so fast that it would be a very slow and inefficient process. In order organisations be able to react on time it is necessary to have an automatized process. Here Artificial Intelligence, and Machine Learning particularly, can be very useful.

Problem Statement

To illustrate how Machine Learning can help to obtain quick knowledge from Social Networks, we are going to apply it to a particular case: we are going to use a set of tweets where U.S. airlines are mentioned, and we are going to build a Machine Learning model able to interpret if a tweet has a negative sentiment associated. This model will be helpful to do sentiment analysis of future tweets in the same domain providing a quantifiable measurement of people consideration about each airline brand and its competitors.

Datasets and Inputs

We are going to use the Kaggle dataset "Twitter US Airline Sentiment" (https://www.kaggle.com/crowdflower/twitter-airline-sentiment) which was generated by Crowdflower. This dataset contains tweets with mention to 5 of the major U.S. airlines and was collected in February of 2015. Each tweet has been manually labelled indicating if it expresses a positive, a negative or neutral sentiment.

The dataset contains 14640 rows and is a bit skewed, because people tend to express more negative opinions: about 63% are negative, 21% neutral and 16% positive.

Solution Statement

We are going to try different supervised models to predict if a tweet has associated a negative sentiment. Particularly we will try some "classic" models like Logistic Regression, SVM, Naive Bayes and more "modern" ones like Random Forest, xgboost and neural networks. We will evaluate the performance of each model (with the metrics defined below) for different data codifications: "bag of words" and embeddings. So, we will be able to compare and to know the best solution.

Benchmark Model

Random guessing keeping the same proportion of negative sentiment predictions that the observed in the dataset. So, if 60% of the tweets express a negative sentiment, our benchmark model will choose randomly 60% tweets from the testing set as negative and the rest as non-negative (neutral or positive sentiment). Tweets text input will be codified as "bag of words" (https://en.wikipedia.org/wiki/Bag-of-words_model).

Evaluation Metrics

This is a classification problem with two possible classes (negative or non-negative sentiment) We are going to evaluate the following metrics: accuracy, AUC and f1. We will use the same metrics for all the models included the benchmark model.

Project Design

We are going to follow these steps:

1. Exploratory Data Analysis:

we are going to explore the dataset to understand it better: we will obtain some statistics, and we will see if there are missing data or any other kind of problem. We will clean the dataset if it is necessary and we will check the data we have can help us to solve the problem.

2. Data preprocessing

we will apply NLP preprocessing techniques as tokenization in order to prepare the dataset to build models. We will split the dataset into a training and a testing set.

3. We will build different supervised models to evaluate them and compare their performance. At least:

- Logistic Regression
- SVM
- Naive Bayes
- Random Forest
- Xgboost
- Neural network

We will try different data input codification: "bag of words" and embeddings. In each case, we will try to find good hyperparameters.

4. We will try to build a stack with some of the previous models to see if we can achieve better performance.

5. Write conclusions.