Twitter Sentiment analysis

Naor Maoz, Eliyahu Peretz

Prof. Lee-ad Gottlieb

41

1.Project goal

Sentiment Analysis is part of the field of Natural Language processing, where the subjective context is learned from text. Our goal was to combine a Machine learning model which was constructed both from processed text and other data keys to achieve a better prediction for a sentiment analysis label.

2. Introduction

Natural language processing has different approaches for text processing ,each one with its advantages and disadvantages.

Experimenting with the traditional algorithms as well as using more advanced algorithms can bring a new perspective on the way we use deep learning algorithms.

3. Text Processing

Stemming – Convert words that have the same meaning to their root form. For example, the same word in different verb tense will be replaced with the same root word in present tense.

Word vectorization – Convert words to numbers, which also converts sentences to vectors. This facilitates the use of several different machine learning algorithms which work on numerical vectors

4. Deep learning Algorithms

Transformers – A transformer is a deep learning model that adopts the mechanism of self-attention, differentially weighting the significance of each part of the input data. It is used primarily in the field of natural language processing and in computer vision.

BERT – Bidirectional Encoder Representations from Transformers is a transformers-based machine learning technique for natural language processing pre-training developed by Google.

GPT 2 – Generative Pre-trained Transformer, published by OpenAI. An unsupervised learning model that its main ability is to predict the next value in an arbitrary sequence.

RoBERTa - Robustly Optimized BERT. Is an improvement of the BERT algorithm, a Pre-training model. Uses large-scale data to perform pre-training.

5. Data

The data consist of 1,600,000 time-stamped tweets over a three month period.

Further additional information was extracted from the 'date' from each time-stamped tweet to gain more features for each record.

6. Selected Approach

- We begin with processing our text, removing unnecessary tokens such as Hashtags, URL's, Emoticons, etc.
- The next stage is vectorizing the data using one of three models: GloVe,Tf-Idf,Word2Vec.
- We then add the timestamp data to the model (Day in month, Day in week, Part of day), that has more impact than an arbitrary data.
- Finally, we train the Machine learning models on a sample of our data and conclude by predicting new untrained data on the models.