Sentiment Analysis

TP App Auto en langues

1 Description

Sentiment Analysis involves estimating identifying and categorizing opinion expressed in a piece of text. In this context, we will train a system to estimate the opinion of a body text. For this purpose, we will automatically determine if a body text, input to our model, is belonging to one of the following classes :

— 0 : negative sentiment
— 1 : positive sentiment

2 Data Description

The dataset is a corpus of sentences with their appropriate class label. The developed models are evaluated with classification accuracy (The percent of labels that are predicted correctly) on a held-out validation set including a body text without class labels.

We will use 3 training/evaluation sets belonging to three different domain.

1. imdb
2. amazon_cells
3. yelp

The sentences come from three different websites/fields : imdb.com, amazon.com and yelp.com. For each website, there exist 500 positive and 500 negative sentences subdivided as follows :

Corpus	train set	valid set	test set
imdb	900	50	50
amazon	900	50	50
yelp	900	50	50

TABLE 1 – Size (#sentences) of train/valid and test sets for each corpus

3 Project Roadmap

1. Study and investigate the training data
2. Use the Pytorch framework to Train and evaluate three different models (RNNs) using the training data independently.
3. Evaluate each model on the in-domain and out of domain data.
4. Use all the training data to create one model. Evaluate this model using the three evaluation sets
5. Compare the results and comment.

For the full datasets look :

1. imdb : Maas et. al., 2011 "Learning word vectors for sentiment analysis"
2. amazon : McAuley et. al., 2013 'Hidden factors and hidden topics : Understanding rating dimensions with review text'
3. yelp : Yelp dataset challenge http ://www.yelp.com/dataset_challenge