Movie Reviews

Fabio Zamboni

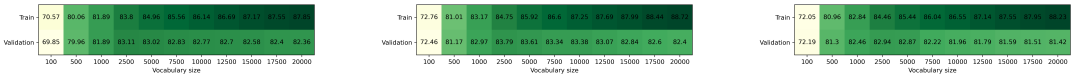
02-04-2023

1 Introduction

In this report we talk about movie reviews, I am not a film expert but I hope my model understands more than I do. The goal is to train a model for classifying reviews into positive and negative. We are going to use a Naïve Bayes and a SVM classifiers, both trained on a bag of word (BOW) features vector. In the first phase, vocabulary creation and feature extraction, we are going to try to discard a set of very common words and apply the stemming technique.[1]

2 Results

Starting whit the Naïve Bayes Classifier we tried different vocabulary sizes, vocabulary building and feature extraction technique, we can see the results in Fig.(1). Looking at the results we



(a) Classic (b) Using Stop Word (c) Using Stop Word and Stemming

Figure 1: Comparison between different feature extraction.

see an improvement in accuracy, in train and validation with respect to the straight method, for the same vocabulary size using stop words, Fig.(1a) and Fig.(1b). Adding stemming, Fig.(1c), we notice that accuracy converges faster in terms of vocabulary size but peaks smaller overall than in case Fig.(1b). Due to the good results on accuracy, the following results will be from the model trained whit only the stop word technique and a vocabulary size of 2500. This, in fact, is the model with the best accuracy in validation. Evaluating on the test set we reach 83.27% of accuracy. Fixed the model now we can investigate the words that most influence the decision, with the following figure we can easily see them Fig.(2).

[1] "Stemming is the process of removing a part of a word, or reducing a word to its stem or root." https://towardsdatascience.com/stemming-of-words-in-natural-language-processing-what-is-it-41a33e8996e2

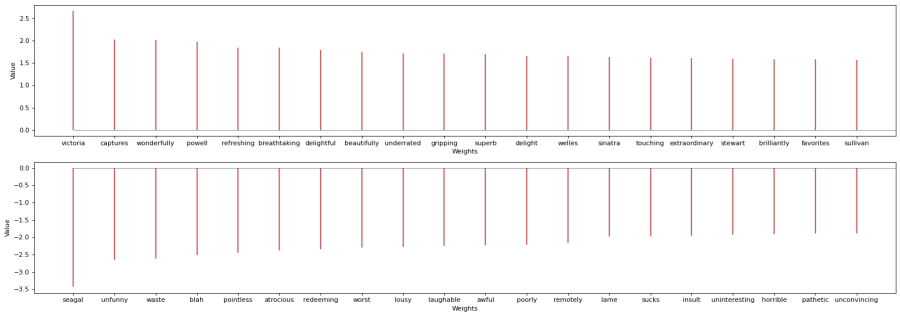
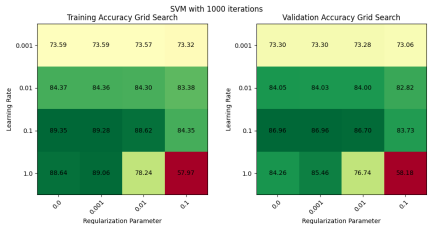


Figure 2: Most relevant weights

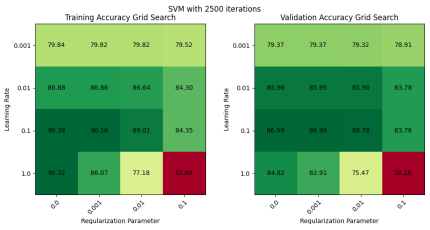
We can also check the worst errors committed by the model, here the top 5 distances, from the decision boundary, of the False Positive and False Negative

$$FP_{Top5} = [-33.72, \quad -31.38, \quad -30.26, \quad -26.46, \quad -26.37]$$
$$FN_{Top5} = [44.45, \quad 38.61, \quad 36.58, \quad 34.51, \quad 32.80]$$

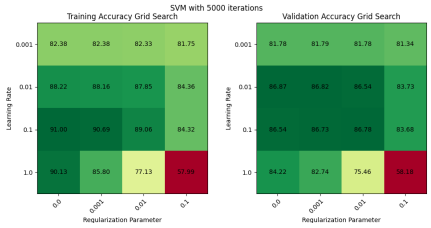
Now it's SVM turn, let's see the result performed with grid search, in particular we can see that the combination $iter = 2500$, $\lambda = 0$ and learning rate $\gamma = 0.1$ give us the best accuracy on the validation data. Evaluating on the test sets we reach 86.90% of accuracy.



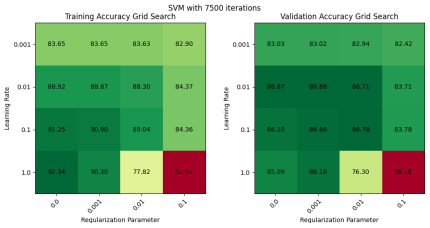
(a) 1000 iter



(b) 2500 iter



(c) 5000 iter



(d) 7500 iter

Figure 3: Grid Search results.

I affirm that this report is the result of my own work and that I did not share any part of it with anyone else except the teacher.