# Natural Language: 2nd Mini-Project Report

Lourenço Ponces Duarte
Student nº197023 - Group 62

## 1. Models and Experimental Setup

This paper presents the experiments that were done using the machine learning algorithm Support Vector Machine (SVM) to classify reviews of beauty products. This technique was chosen because it was stated by [1] that for sentiment analysis it presents the best results compared with other approaches and, also, that by utilizing kernel procedures, the training process maximizes the margin between two classes on the feature space.

For the kernel parameter, it was used the Radial Basis Function (RBF) since the problem is non-linear and there is no prior knowledge about the data. Furthermore, RBF Kernel has the advantages of the K-Nearest Neighbors (K-NN) and overcomes the space complexity problem as RBF Kernel SVM just need to store the support vectors during training and not the entire dataset [2].

Initially, the data provided was split 95% into training data and 5% into test data in a random way. And then it was vectorized using the Term Frequency-Inverse Document Frequency (Tf-idf) which prioritized each word by comparing the frequency of a particular term relative to the whole corpus. This method was used because it allows one to take advantage of its inexpensive computations [3].

Much of the code was reused from [4] because it had an implementation that used the desired techniques.

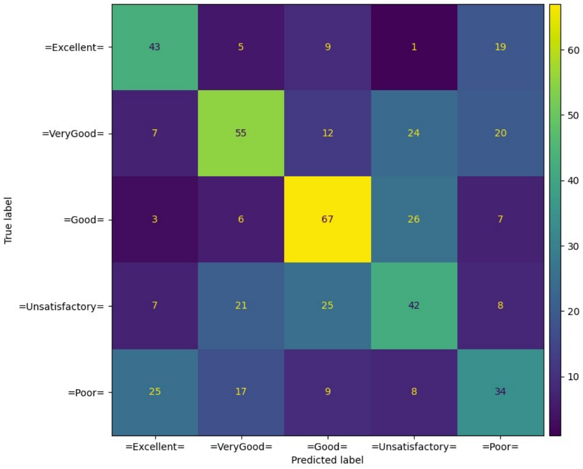

**Figure 1. Confusion matrix**

| Label | Precision | Recall |
|---|---|---|
| =Excellent= | +/- 50% | +/- 55% |
| =VeryGood= | +/- 38% | +/- 36% |
| =Good= | +/- 52% | +/- 46% |
| =Unsatisfactory= | +/- 41% | +/- 40% |
| =Poor= | +/- 54% | +/- 61% |

**Figure 2. The table that evaluates each label with the precision and the recall metrics**

## 2. Results

Depending on the split of the data, i.e. the inputs that are selected for the training set and those that are selected for the test set, the accuracy of the model varies between 44% and 50%. These results beat baselines 1 and 2 for the values of 36.8% and 43% respectively.

The confusion matrix (Figure 1) shows the number of inputs that have been correctly predicted, as well as the inputs that have been incorrectly predicted and for which labels. The vertical axis represents the actual labels, while the horizontal axis represents the predicted labels.

Figure 2, on the other hand, presents 2 metrics [5] for the prediction of each label. The table shows approximate results, again, because the model ran several times and for each time the training data and the test data were different.

- **Precision**: It presents what percentage is truly positive.
- **Recall**: Considering only the positive cases, it presents what percentage are predicted positive.

## 3. Discussion

Before discussing the results, it is necessary to point out that there are some cases where the labels that were assigned may not be 100% correct for their respective reviews. For example, the review ”Absolutely the best!!!!!!!” was assigned the label ”=Poor=” when, probably, a more positive label should have been assigned, in fact, the system assigned the ”=Excellent=” label which, personally, makes more sense.

There are also situations where the review is made in a tone of irony which leads the model to assign a label according to the sentiment of the words present, when in fact it should assign for the real sentiment of the sentence. Most cases this means that the meaning is the opposite of the words. For example, for the review ”Want the biggest zit of your life?” the label ”=Poor=” was assigned, however the system tagged with the ”=Excellent=” label.

In short, since we are working with real data, there are quite a few errors and therefore the accuracy of the system cannot reach more promising values. Furthermore, although SVMs are a good solution to the problem presented, they are not good enough to achieve high accuracies.

## 4. Future work

Other machine learning algorithms could be implemented for future work, such as the Naive Bayes classifier or the K-Nearest Neighbors (K-NN). The system's accuracy could show more promising results if these were used in conjunction with the Support Vector Machine algorithm already implemented.

## References

[1] Tanasanee Phienthrakul et al. “Sentiment classification with support vector machines and multiple kernel functions”. In: *International Conference on Neural Information Processing*. Springer. 2009, pp. 583–592.

[2] Sushanth Sreenivasa. *Radial Basis Function (RBF) Kernel: The Go-To Kernel*. URL: https://towardsdatascience.com/radial-basis-function-rbf-kernel-the-go-to-kernel-acf0d22c798a. (accessed: October 30, 2022).

[3] Anirudha Simha. *Understanding TF-IDF for Machine Learning*. URL: https://www.capitalone.com/tech/machine-learning/understanding-tf-idf/. (accessed: October 30, 2022).

[4] Vasista Reddy. *Sentiment Analysis using SVM*. URL: https://medium.com/@vasista/sentiment-analysis-using-svm-338d418e3ff1. (accessed: October 25, 2022).

[5] Boaz Shmueli. *Multi-Class Metrics Made Simple, Part I: Precision and Recall*. URL: https://towardsdatascience.com/multi-class-metrics-made-simple-part-i-precision-and-recall-9250280bddc2. (accessed: November 01, 2022).