
DM2583 Big Data in Media Technology Fall 2022

# Lab 2: Sentiment Classification with Support Vector Machine

September 22, 2022

**Group 10: Senane Zineb, Oumhamed Younes, Oliver Gui**

Emails: {senane, oumhamed, ocwgui}@kth.se

Professor: Haibo Li

# 1 Diagram of Sentiment Analysis Process

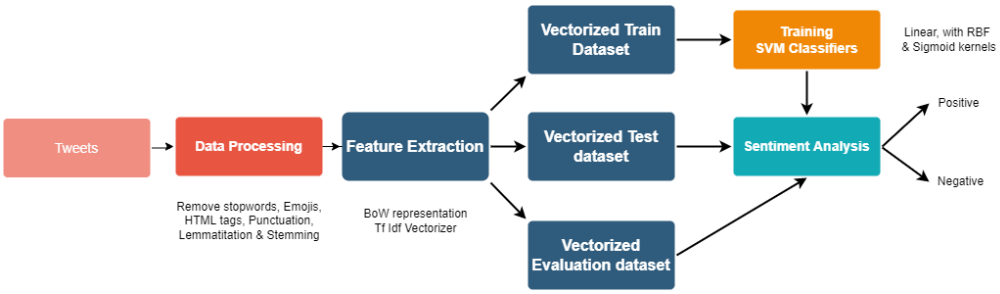


Figure 1: Systematic diagram of Sentiment Analysis Process

# 2 Data Processing and Feature Extraction

We applied the same data processing and feature extraction steps as in Lab 1.

# 3 Model Selection

We used a Support Vector Machine (SVM) to perform sentiment classification on the data. SVMs using a linear kernel can classify linearly separable data, but not non-linearly separable data. To account for this possibility, we used kernels such as the Gaussian Radial Basis Function (RBF) and Sigmoid. These kernels aim to project non-linearly separable data in lower dimensions to linearly separable data in higher dimensions.

# 4 Model Evaluation

As per Lab 1, we used accuracy as the metric for model performance. Additionally, we will use the F1-score. The use case of the model will determine which metric is more relevant, with accuracy more relevant if the cost of false negatives and false positives are similar, and F1-score more relevant otherwise. The highest accuracy and F1-score for test and evaluation data is obtained with the RBF kernel and Tf-IDF.

| | ***Models*** | **Bag Of Words** | | | **Tf IDF Vectorizer** | | |
|---|---|---|---|---|---|---|---|
| | | Training | Testing | Evaluation | Training | Testing | Evaluation |
| *Accuracy* | Naive Bayes | 92.2% | 82.3% | 80.1% | 93.2% | 82.8% | 81.0% |
| | Linear SVM | 99.0% | 81.2% | 80. 2% | 95.6% | 83.8% | 83,4% |
| | SVM with RBF | 99.9% | 59.2% | 50.3% | 98.0% | 84.7% | 84.2% |
| | SVM with Sigmoid | 68.9% | 67.3% | 70.1% | 92.4% | 83.5% | 83.5% |
| *F1 score* | Naive Bayes | 92.2% | 82.2% | 80.1% | 93.2% | 82.8% | 80.9% |
| | Linear SVM | 99.0% | 81.2% | 80.1% | 95.5% | 83.8% | 83.4% |
| | SVM with RBF | 99.9% | 52.1% | 35.0% | 98.0% | 84.7% | 84.2% |
| | SVM with Sigmoid | 68.9% | 67.3% | 69.3% | 92.4% | 83.5% | 83.5% |

# 5 Results and Limitations

Comparing the results of the SVM with the Naïve Bayesian Classifier, we observe that the SVM model applied on data processed with TF-IDF produces a better accuracy and F1-score compared to Naïve Bayesian Classifier. However, even with the models that produce the highest accuracy and F1-scores on the test and evaluation datasets, we noticed a significant drop-off in performance when using the classification model on unseen data. Hence, this shows that our model suffers from overfitting. We can ameliorate the model performance by tuning the hyper-parameters and using cross-validation. Better performance might be obtained using deep learning models such as a word-based Convolutional Neural Network or Long-Short Term Memory Recurrent Neural Network.