DM2583 Big Data in Media Technology Fall 2022

Lab 1: Sentiment Analysis

September 16, 2022

Group 10: Senane Zineb, Oumhamed Younes, Oliver Gui

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

Professor: Haibo Li

1 Diagram of Sentiment Analysis Process

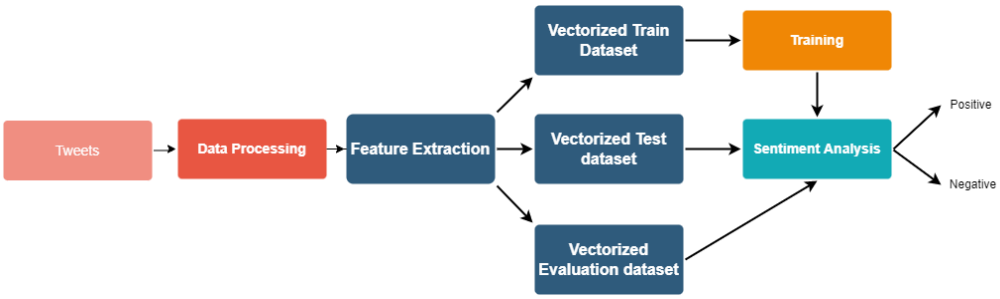

Figure 1: Systematic diagram of Sentiment Analysis Process

2 Data Processing

Some input text data contains HTML entities, hashtags, mentions, emojis and URLs. To clean our dataset we used NLTK library functions along with the Rgex library to remove these expressions. We processed text used as input for model is the cleaned tweets lemmatized and stemmed.

3 Feature Extraction

BoW using Count Vectorizor: The BoW model uses the Count Vectorizor to convert text into a vector of known words, where the number of occurrences of a word is stored.

TF-IDF Vectorizor: The TF-IDF vectorizor converts text into vectors by measuring the importance of words for statistical analysis. As the occurrences of a word in a document increases, TF increases the statistical importance of this word. However, the IDF decreases the statistical importance of a word if it appears in many documents. Hence, common words that appear in most sentences will be given a lower importance, whereas a word that is present many times in a few documents might be more indicative of the context of the document.

4 Model Evaluation

We used the accuracy to evaluate the performance of the model when different vectorization techniques are used. The choice of accuracy was motivated by having a balanced dataset. The highest accuracy is obtained with TFIDF Vectorizer. However we observe that the model doesn't generalize well and overfits.

Accuracy	Training	Testing	Evaluation
Bag Of Words	92.2%	82.3%	80.1%
Tf IDF Vectorizer	93.2%	82.8%	81.0%

5 Results and Limitations

With the evaluation dataset we obtained an accuracy of 81% with TFIdf. To further understand the performance of our models, we observe that the number of false negative and positive is quite high and hence that's confirm the overfitting that our model suffers from. We can ameliorate the model performance by well tuning the parameters and using cross-validation. A higher accuracy can be reached using other models such as Bert.

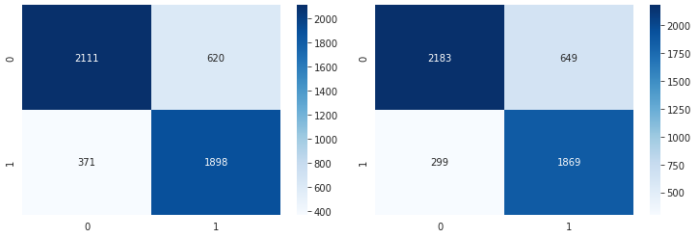

Figure 2: Confusion Matrix of Naive Bayes Classifier with BoW (left) and Tfidf (right)