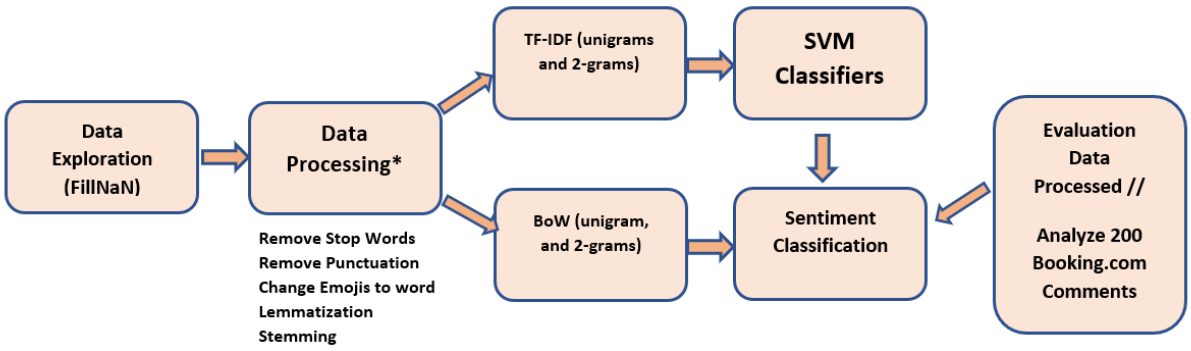
Big Data in Media Technology

Lab 2: Sentiment Classification with Support Vector Machine Classifier

Bosen Cheng - Hamza AJJA - Yassine ELKHEIR (Team 2)

I. Analysis Process Diagram:



II. Features Extraction:

We have used the same feature extraction methods as in the last lab (TF-IDF and BoW), but we developed them by taking into account unigrams (independent single words) and 2-grams (combinations of two words). We avoided taking a larger number of combinations of words into account so as not to fall into the problem of overfitting.

III. Choose Feature Extraction Model and The best Classifier :

Accuracy / F1-Score							Accuracy
Models	SVM (Constraint = 0.1)		Linear SVM		RBF Kernel SVM		Naïve Bayes
	TF-IDF	BoW	TF-IDF	BoW	TF-IDF	BoW	TF-IDF
Train	0.99/0.99	0.99/0.99	0.99/ 0.99	0.99/ 0.99	0.99 / 0.99	0.99 / 0.99	0.9248
Test	0.844 / 0.843	0.814 / 0.813	0.8472 / 0.8475	0.8168 / 0.8167	0.836 / 0.8363	0.55 / 0.46	0.8216

After the analysis of the measures we obtained, we see a big problem which is the overfitting, for all the models tested, even if we only added the 2 grams to TF-IDF. For the compression of the SVM and Naïve Bayes model, we saw that SVM is doing better, but they caused the overfitting. In order to test this hypothesis, we have to use a statistical testing hypothesis technique to get the statistical significance of this result (due to the limited time, we could not do this, but we plan to integrate it in the final project).

IV. Result :

We have reached an accuracy of 83.00% on evaluation data using TF-IDF (2-grams) and Linear SVM. Our accuracy has increased from slightly less than 0.81 using TF-IDF(unigram) and Naïve Bayes to 0.83 using Linear SVM. So as mentioned in the last lab, our assumptions were correct, using a more sophisticated model results in a higher accuracy for testing data, but we still suffer from the overfitting problem which is a crucial problem that we have to deal with. In the project, we aim to build a model using the word2vec word embeddings technique that tries to encapsulate the meaning of a given word to reduce the overfitting.