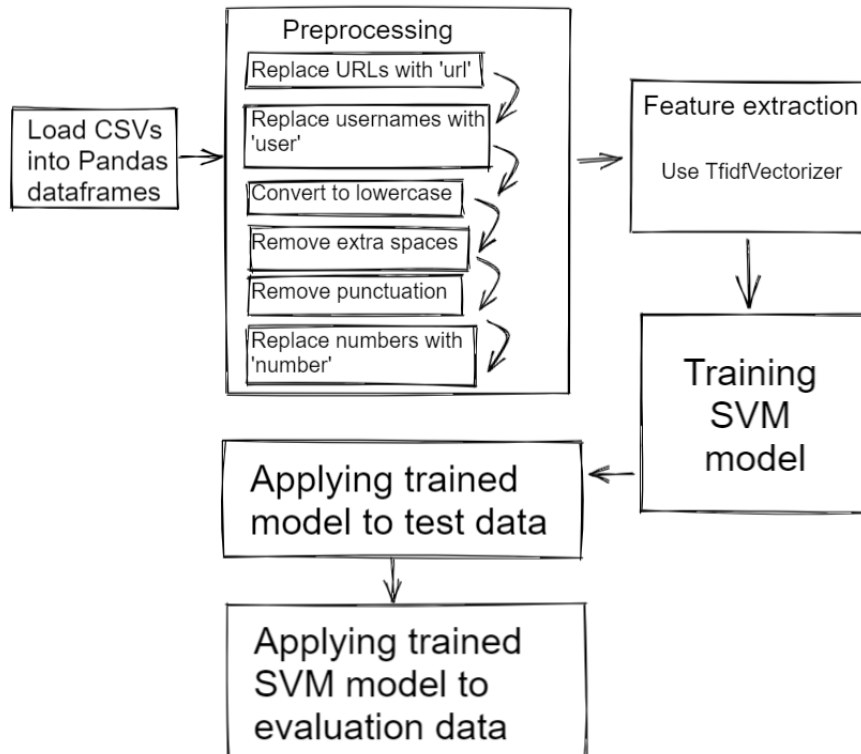


Lab 2: Sentiment Classification with Support Vector Machine

By Group 3: Attila Kádár, Elias Alkvist Cetin, Siyana Ivanova

System diagram of the sentiment analysis process



Feature extraction

After preprocessing the data as described in the system diagram, we apply the `TfidfVectorizer`, which converts a collection of raw data entries into a TF-IDF (term-frequency times inverse document-frequency) matrix of features. This is equivalent to using `CountVectorizer` and then `TfidfTransformer` like we did last time.

Evaluation

In order to evaluate the performance of our SVM model, we ran it on the evaluation dataset and used the sklearn function `classification_report`, which returns a summary of precision, recall, and F1 score per class (in our case, positive or negative).

Results

The screenshot below shows our summary and our training and prediction times.

```
Training time: 14.159907s; Prediction time: 13.584318s
positive: {'precision': 0.856909239574816, 'recall': 0.8397435897435898, 'f1-score': 0.848239579117766, 'support': 1248}
negative: {'precision': 0.8433829287392326, 'recall': 0.860223642172524, 'f1-score': 0.8517200474495848, 'support': 1252}
positive: {'precision': 0.858540562576437, 'recall': 0.8363780778395552, 'f1-score': 0.8473144236572118, 'support': 2518}
negative: {'precision': 0.8382410679230468, 'recall': 0.8601933924254633, 'f1-score': 0.8490753628952078, 'support': 2482}
```

Specifically, our accuracies were as follows.

Test	Evaluation
Positive: ~86%	Positive: ~86%
Negative: ~84%	Negative: ~84%