Group Project Poster

By Alex de Roode and Dexter Chen

Group 28

Task Description

This project implements and evaluates three fundamental Natural Language Processing (NLP) techniques: Named Entity Recognition and Classification (NERC), Sentiment Analysis, and Topic Analysis. The implementation uses both rule-based and machine learning approaches.

Data Description

The data we used were 3 different sources: 1. the my_tweets.json file: This file contains tweets sampled from the web with sentiment annotation. 2. Tsentiment-topic-test.tsv. This test sets contains 18 annotated sentences with sentiment labels and topic categories.
3. NER-test.tsv: This test set contains the tokenized sentences using the BIO tagging scheme for NERC.

Approach Description

Regarding the sentiment analysis, we compared 2 different approaches. The first method consisted of using VADER, whilst the second approach was using an SVM to compare the ML-side. Regarding NERC, we used SpaCy's pre-trained English model. And for the topic analysis, we implemented an ML-approach using TF-IDF features and an SVM.

Potential Applications:

1. **Social media monitoring**
2. **Customer Feedback Analysis**
3. **Content Categorization**
4. **Information Extraction from Unstructured Text**
5. **News Article Classification & Entity Extraction**

Division of Labour

When considering the division of labor, we chose to do split equally the work for both the poster and coding/analysis. Alex focussed on the preprocessing part as well as the lower 3 parts of the poster, while Dexter did the remaining work

Results Analysis

Regarding sentiment analysis, our results for VADER we had a 55.6% accuracy with balanced precision and recall across all classes. The SVD did much worse compared to vader, with only a 22.2% accuracy on the test set. So VADER did way better in this regard.Regarding NERC, SpaCy's model achieved a respectable 81% overall accuracy, with particularly good performance on LOC and PERSON entities. And regarding topic analysis, the results were a measly 38.9% accuracy using cross-validation

Conclusions & limitations

To conclude our research, it is safe to say that VADER performed much better than the SVM model with respect to sentiment analysis. However, we also suspect that this could be due to the SVM having a low amount of training data, causing that problem to bottelneck its potential. Regarding NERC, spacy's model performed well with an respectable 81% accuracy. We suggest finetuning the model to perform better in the future. And regarding topic analysis, we need an increase in size for the training dataset. Also , domain adaptation could be used to improve results