

COMP30027: Assignment 2 Reviews

Rohan Hitchcock (836598) and Patrick Randell (836026)

Report 1 review

Not exactly sure how to start with summarising what they've done. Something about them tackling sentiment analysis with a mixture of popular linear classifiers, both probabilistic and deterministic? But their motivations for their approach don't exist.

The analysis of selected models for the task is extensive. The authors thorough investigation into their models parameters is reflected by the increase in final prediction accuracy.

This report would benefit from some restructuring. A clear method, with the motivations behind decisions made backed by relevant literature, should precede any results or analysis. The author makes assumptions about the readers understanding at times, resulting in unclear terminology and values. In particular, it is not clear what values the numbers in section 3.2.1 are referring to, or what SGD stands for in section 3.3.2. These kinds of oversights occur throughout.

The underlying machine learning concepts behind Scikit Learn library functions should be used rather than the functions themselves (LinearSVC in particular is a SVM). A closer attention to formatting, including consistent decimal places, spacing, and labelling would improve the overall readability of the report.

- Discussion of literature is non-existent. The only reference which is not the datasets is to a blog post.
- more theoretical analysis everywhere
- Always need to say what the numbers are (are the decimals in 3.2.1 accuracy? fscore? – this occurs throughout)
- Need to explain the reasons for doing things more often (why feature 83 and 30?)
- Using acronyms without explaining them (SGD), names of files, method names from specific implementations is extremely unhelpful (so are terminal screenshots)
- Use expected terminology, Figure 6 is a confusion matrix. SVC is sklearn terminology.
- Formatting problems are distracting
- The number of decimal places in Table 2 is ridiculous, and why accuracy is a tuple is not clear
- Is increasing accuracy from 0.822 to 0.843 a 'huge' increase?

Report 2 review

This report presented an approach to text sentiment analysis which is based on sophisticated feature generation and selection. The approach to feature generation and selection presented in this report is well considered, and appears to be based on a robust statistical understanding of natural language processing. This is reflected in the strong performance of the final classifier.

This report would benefit from more explanation of the theory behind the techniques. While the efficacy of the techniques is clear from the experimental data presented, it is not clear why this is the case. In particular, it would be good to include a discussion of why different classifiers had different performance with the selected features, and why the generated features result in strong model performance. Along these lines, this report would be improved by including a more comprehensive introduction to the feature generation techniques employed. In the reviewers' opinion, the assumption that readers are "familiar with the basic theory of TF-IDF and n-gram vectorised representation of text" is not reasonable, and failing to include a detailed discussion of this makes the remainder of the report less accessible. Additionally, including more references to literature would help guide a reader looking to explore the topics discussed here further.

- Would benefit from more direct explanation at times
- Need to introduce TF-IDF and n-grams
- The relevance of much of the introduction is not clear
- more references
- less figures
- Feature selection is cool