

Classification of Arabic News Articles

Ian Sharff

Agenda

- Background
- Business Context
- Data Understanding
- Modeling and Evaluation
- Conclusions and Future Work

The Arabic Language

- Complex morphology compared to English

Business Context

- One of the 6 official languages of the UN
- Rich vocabulary, complex grammar, difficult to master
- Tool for Arabic educators and their students

Data Understanding

- Data from the SANAD corpus
- Roughly 200,000 articles
- No intermediate translation

Model Building and Evaluation

- Naïve Bayes: 95 %
- Random Forest: 99 %
- Gradient Boosting: 97 %
- Support Vector Machine: 98 %

Conclusions and Future Work

- Implementation and deployment
- Model improvement (neural networks/word embeddings)

The background features a dark blue-grey gradient with a subtle radial blur effect. Overlaid on this are several large, semi-transparent circles in varying shades of grey. Some circles contain abstract patterns of thin, light brown lines forming concentric arcs or radiating from a central point. The overall aesthetic is clean and modern, with a focus on geometric shapes and organic patterns.

Questions?