Classification of Arabic News Articles

Ian Sharff

Agenda

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

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

The Arabic Language

- Right to left; letters joined
- Morphology
  - 3 (or 4) letter roots
  - Irregular plurals
- Verbal forms
  - Up to *5,400 forms* per verb

ك ت ب

book – كِتابٌ
books – كُتُبٌ

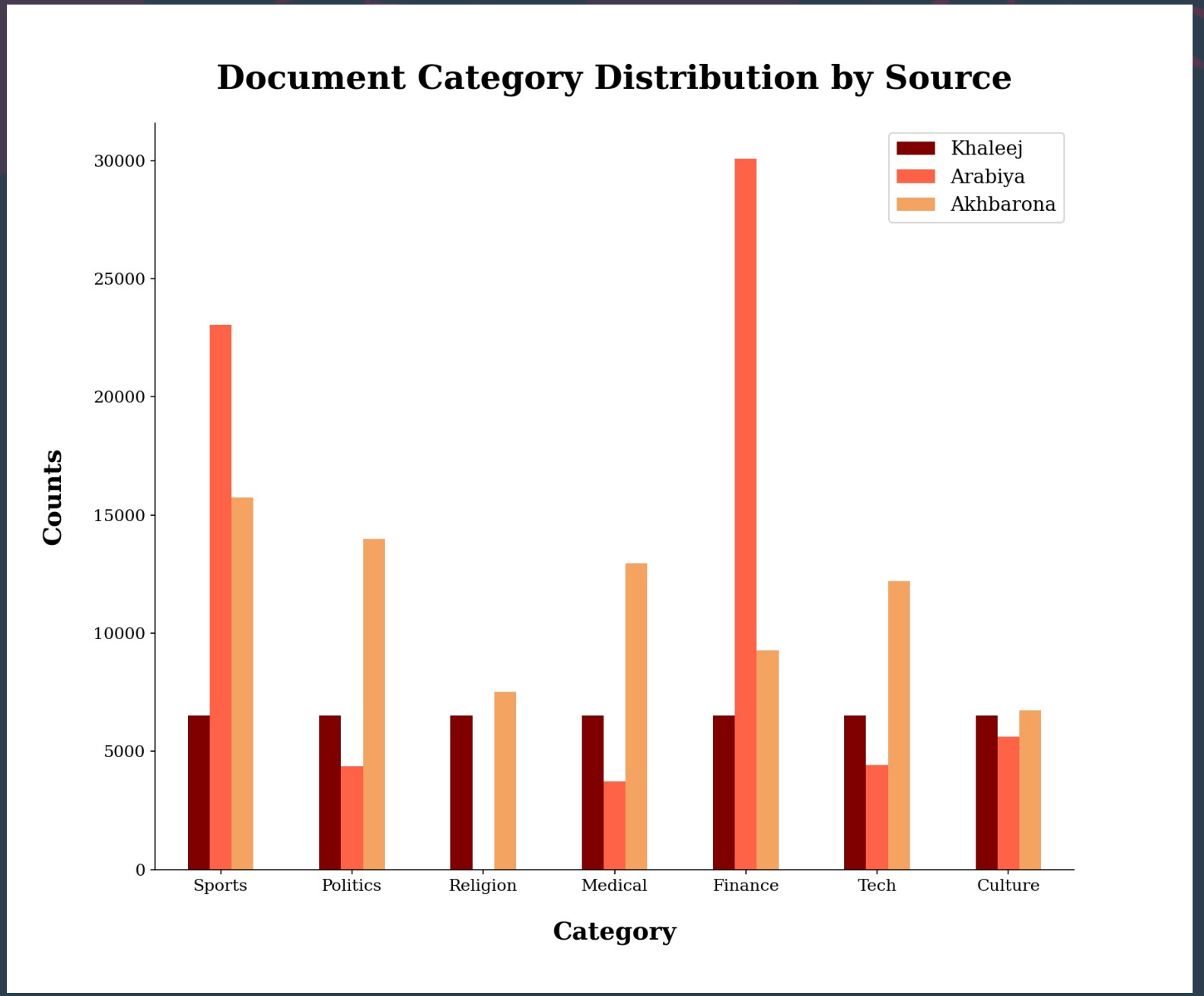
writer – كاتِبٌ
writers – كُتَّاب

office – مَكْتَبٌ
offices – مَكاتِب

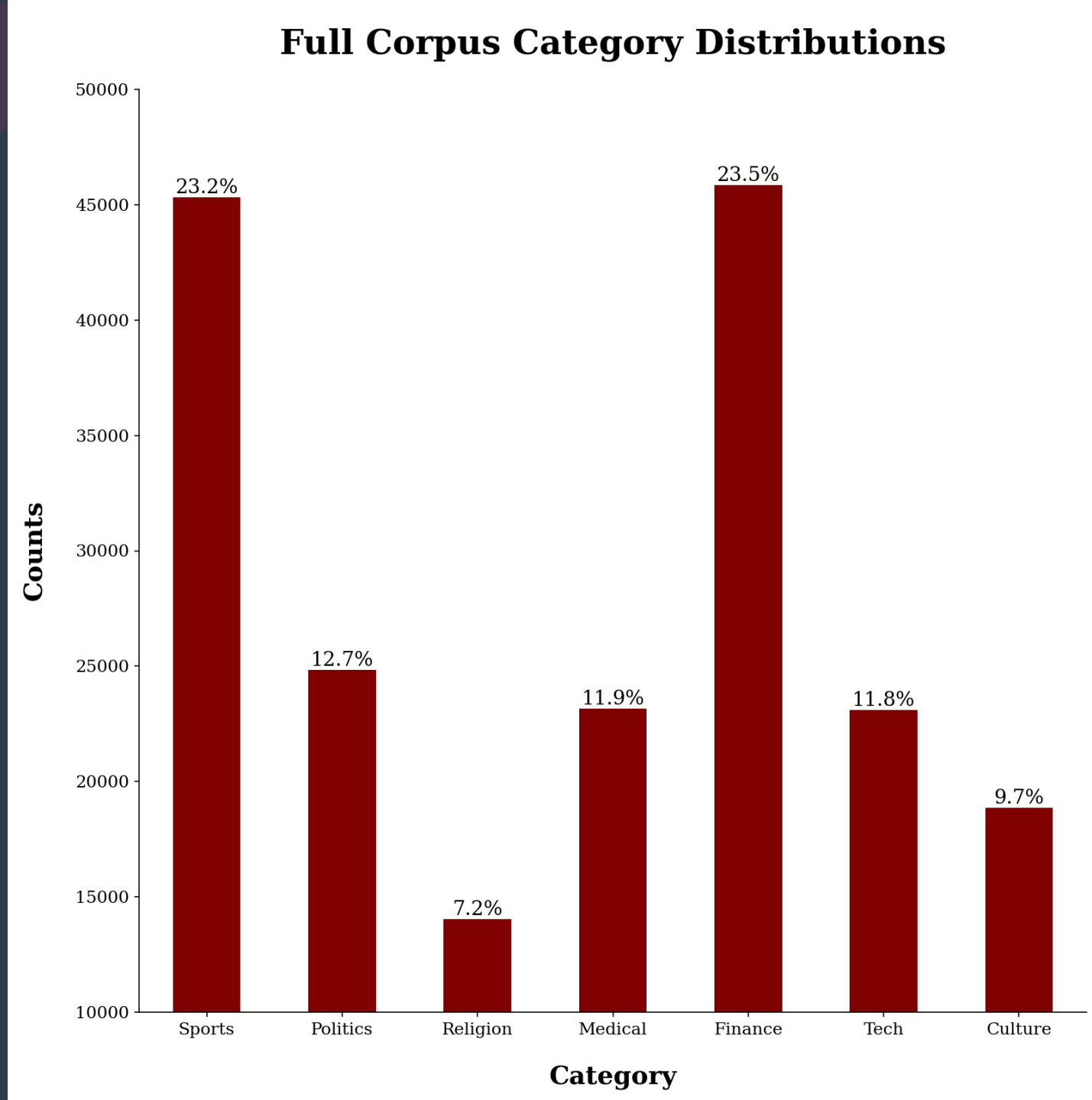
https://towardsdatascience.com/arabic-nlp-unique-challenges-and-their-solutions-d99e8a87893d

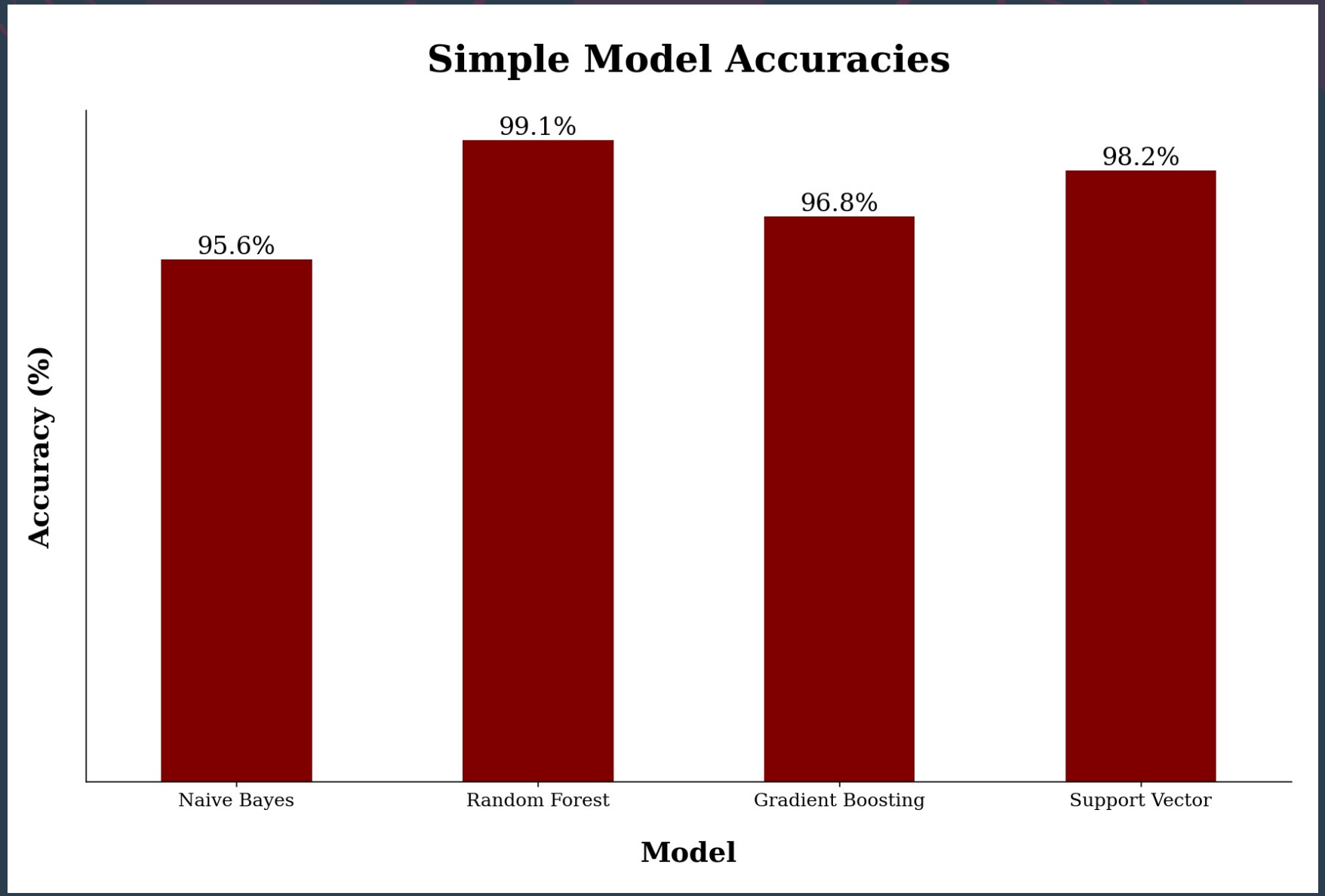
Data Understanding

- SANAD corpus
- Roughly 200,000 articles
- Written and analyzed in Arabic



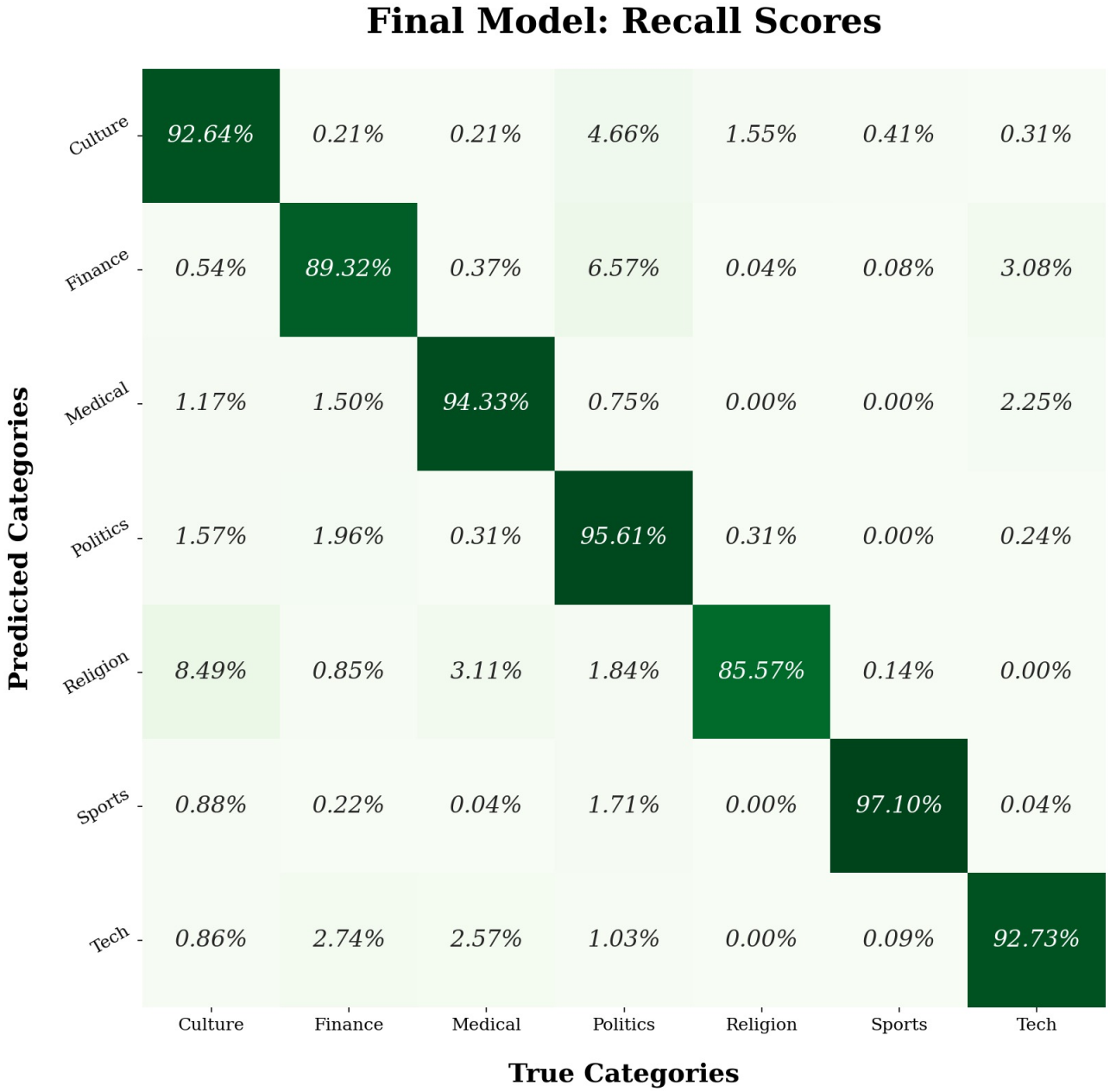
Overall Distribution





Simple Model Evaluations

- Ordered by increasing computational cost



- Confusion in similar categories
  - Religion-Culture
  - Politics-Culture
  - Politics-Finance
- Accuracy: 93.0%

Conclusions and Future Work

- Robust morphological analysis
- Neural networks/word embeddings
- Deployment as a functional application

Acknowledgements

- NYU Abu Dhabi CAMeL Lab
  - https://github.com/CAMeL-Lab/camel_tools
- SANAD Corpus
  - https://data.mendeley.com/datasets/57zpx667y9/2

Thank you, Questions?

Ian Sharff
Email: ips11@georgetown.edu
Github: https://github.com/iansharff/arabic_news
Linkedin: https://linkedin.com/in/iansharff