BAYESIAN CLASSIFICATION OF RESEARCH PAPERS

Huilin Chang(hc5hq) , Yihnew Eshetu (yte9pc), Celeste Lemrow (ctl7t)

DS6014 Project

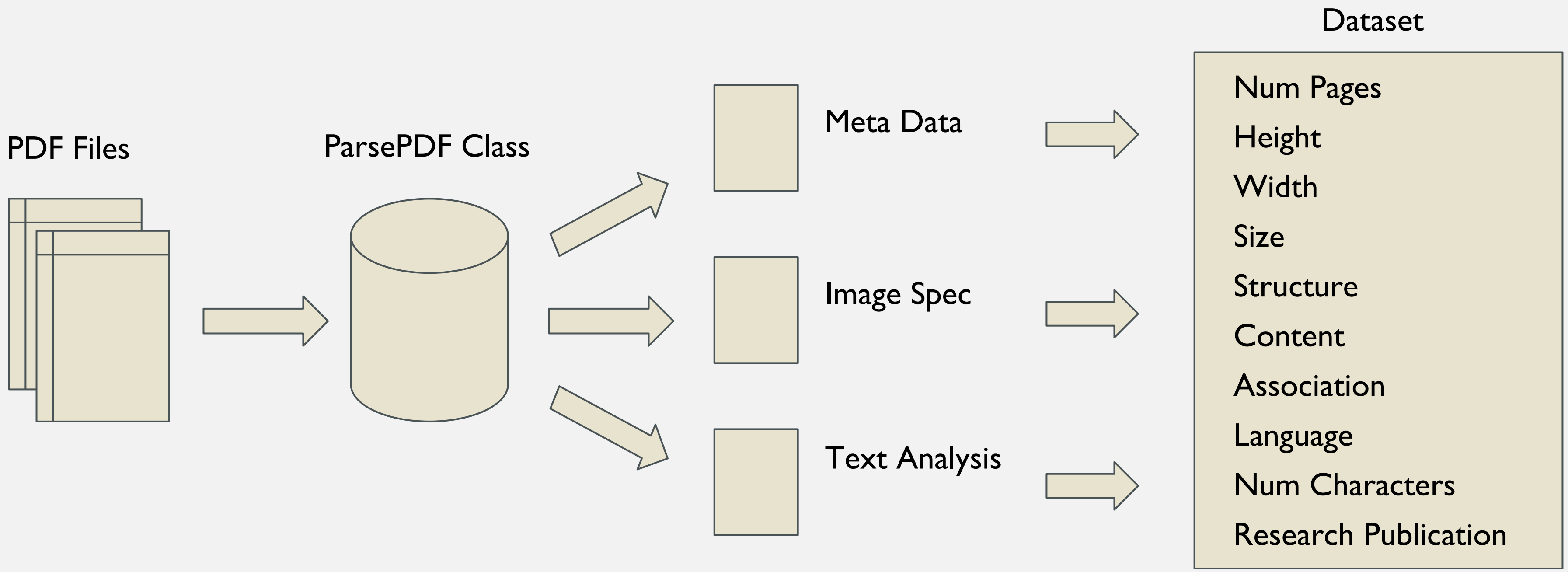
TABLE OF CONTENTS

- Introduction
- Data Extraction
- Exploratory Analysis
- Methods
- Results
- Conclusions

INTRODUCTION

- One of the Internet Archive’s mission areas is “Universal Access to All Knowledge”, which is an attempt to collect and provide access to the “scholarly web”: the public record of research publications and datasets available on the world wide web.
- Our project aims to help this mission by implementing a fast PDF identification tool, which will score files on their likelihood of being a research publication.
- Given the volume of PDF documents in the Internet Archive’s (IA) repository, a classifier is needed to determine which are legitimate research documents
- Our group constructed several Bayesian logistic models with different feature combinations and compared results on accuracy of identification of research papers

DATA EXTRACTION



PDF Files

ParsePDF Class

Meta Data

Image Spec

Text Analysis

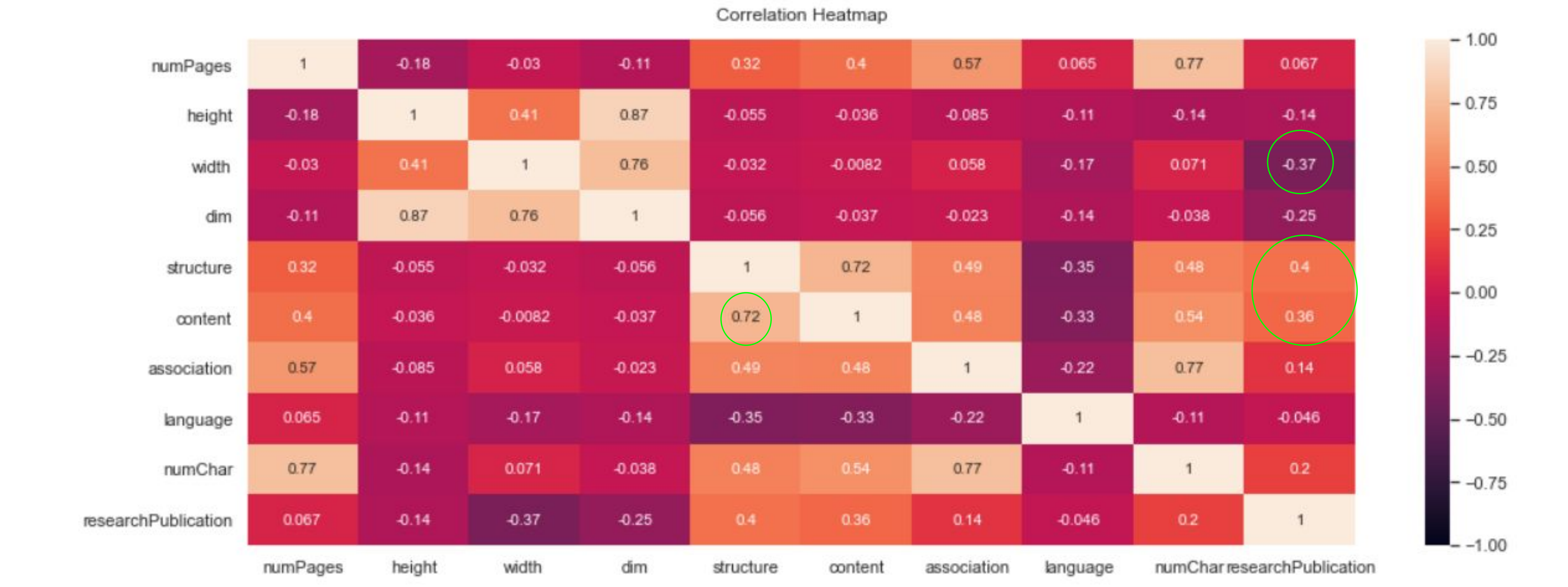
Dataset

- Num Pages
- Height
- Width
- Size
- Structure
- Content
- Association
- Language
- Num Characters
- Research Publication

Language: English, Romance, and other
Structure: Words that represent the structure of a paper
{abstract, introduction, conclusion, reference, table of content}
Content: Words that represent the content of a paper
{research, analyze, result, table, investigation, explain, theory, study, paper, data, perform}
Association: Words that represent association
{journal, association, organization, doi, university, school, board}

EXPLORATORY ANALYSIS

Correlation matrix : correlation between the research publication variable and selective features

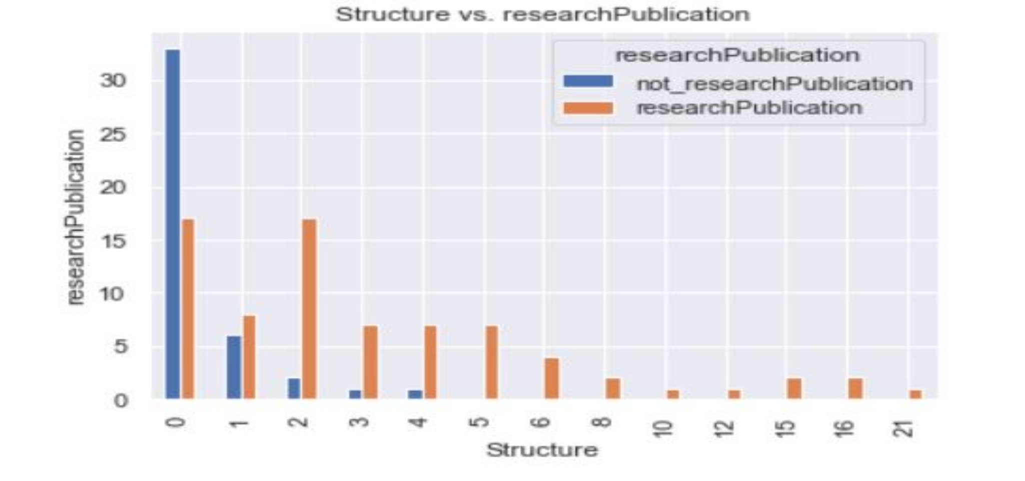


Correlation Heatmap

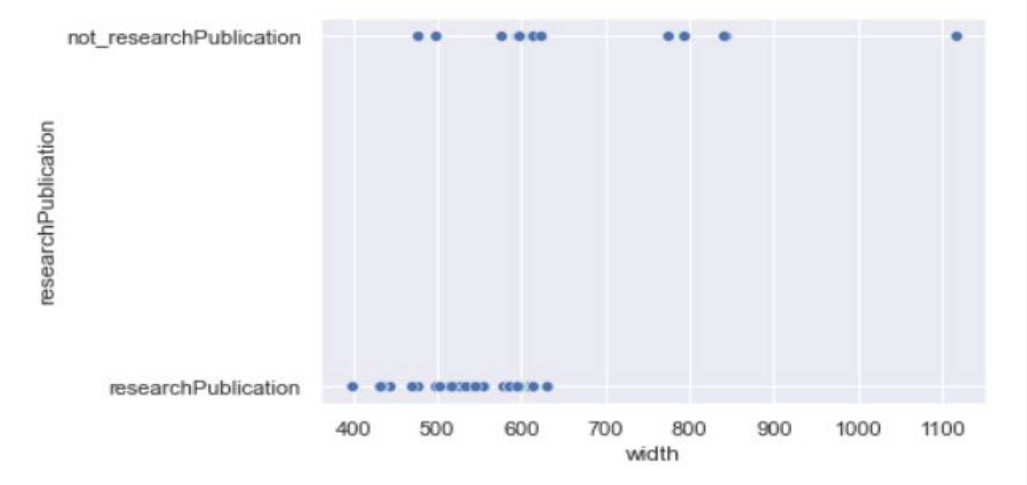
	numPages	height	width	dim	structure	content	association	language	numChar	researchPublication
numPages	1	-0.18	-0.03	-0.11	0.32	0.4	0.57	0.065	0.77	0.067
height	-0.18	1	0.41	0.87	-0.055	-0.036	-0.085	-0.11	-0.14	-0.14
width	-0.03	0.41	1	0.76	-0.032	-0.0082	0.058	-0.17	0.071	-0.37
dim	-0.11	0.87	0.76	1	-0.056	-0.037	-0.023	-0.14	-0.038	-0.25
structure	0.32	-0.055	-0.032	-0.056	1	0.72	0.49	-0.35	0.48	0.4
content	0.4	-0.036	-0.0082	-0.037	0.72	1	0.48	-0.33	0.54	0.36
association	0.57	-0.085	0.058	-0.023	0.49	0.48	1	-0.22	0.77	0.14
language	0.065	-0.11	-0.17	-0.14	-0.35	-0.33	-0.22	1	-0.11	-0.046
numChar	0.77	-0.14	0.071	-0.038	0.48	0.54	0.77	-0.11	1	0.2
researchPublication	0.067	-0.14	-0.37	-0.25	0.4	0.36	0.14	-0.046	0.2	1

EXPLORATORY ANALYSIS

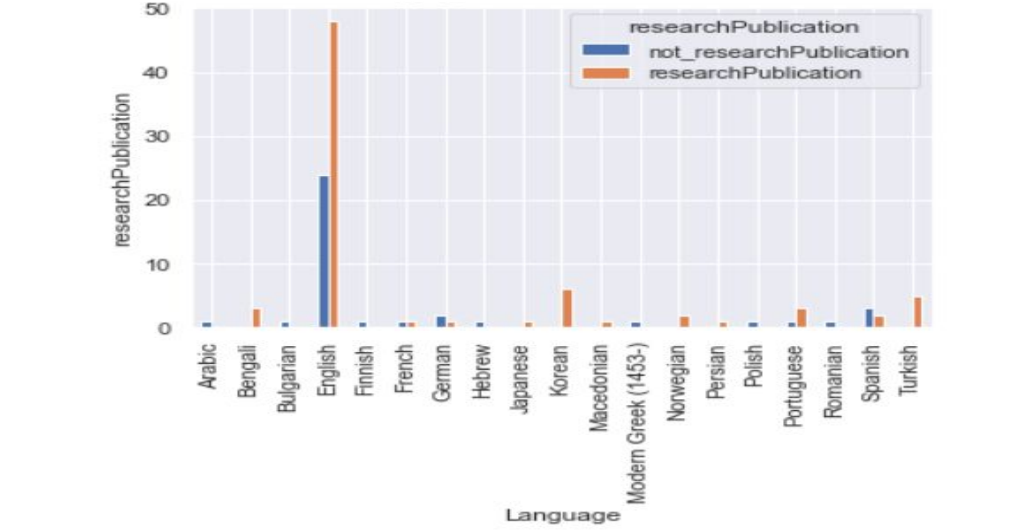
Structure vs. Research Publication



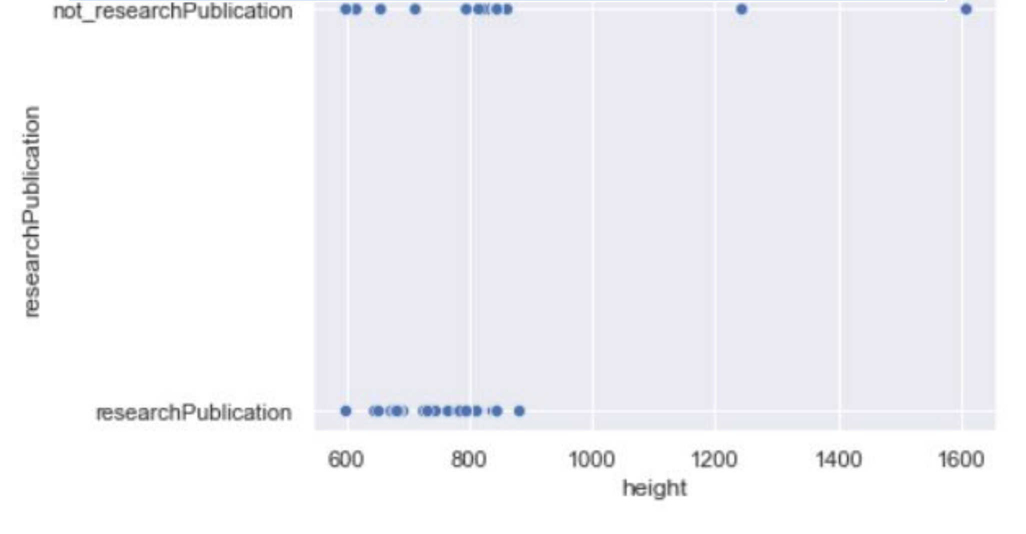
Width vs. Research Publication



Language vs. Research Publication



Height vs. Research Publication



BAYESIAN STATISTICS

- Algorithm: Bayesian logistical regression, using PYMC3

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)}$$

- Mathematical connection : the likelihood is the product of n Bernoulli trials,

$$\coprod_{i=1}^{n} p_i^{y}(1-p_i)\ (1-p_i)^{1-y_i}\text{, where } p_i = \frac{1}{1+e^{-z_i}}$$

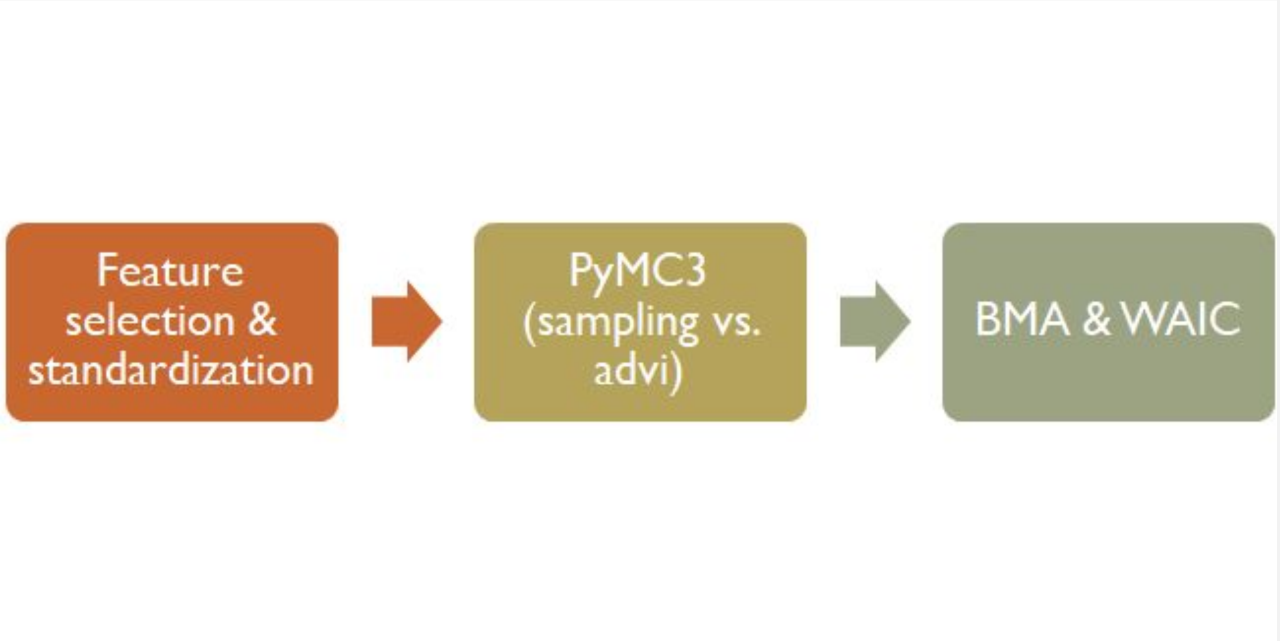
$$y_i = \beta_0 + \beta_1(\text{numPages})_i + \beta_2(\text{height})_i + \beta_3(\text{width})_i + \beta_4(\text{dim})_i + \beta_5(\text{structure})_i + \beta_6(\text{content})_i + \beta_7(\text{association})_i + \beta_8(\text{language})_i + \beta_9(\text{numChar})_i$$

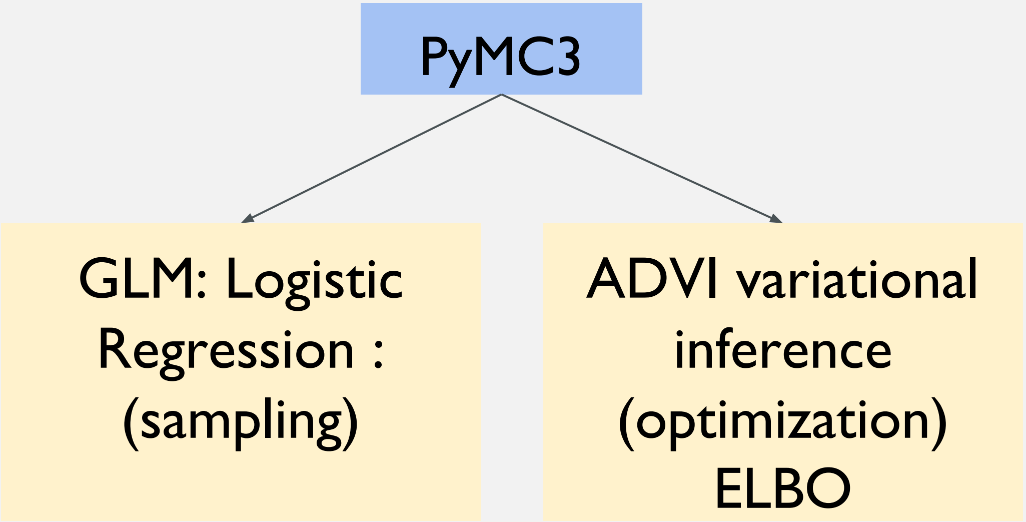
Where y_i = 1 if researchPublication and y_i = 0 otherwise

- Priors : default $p(\theta)$ = N(0, 10^{12}I)

METHODS

- Total set of features considered: number of pages, height, width, dimensions of page, structure, content, association, language, number of characters
- How likely is it a **research publication** based on the selective features?
- Model comparison approach – compared different sets of features and accompanying accuracy. Given parameters for the capstone project, including speed, prioritizing a balance of the smallest number of features with acceptable accuracy is a key objective


Feature selection & standardization
PyMC3 (sampling vs. advi)
BMA & WAIC

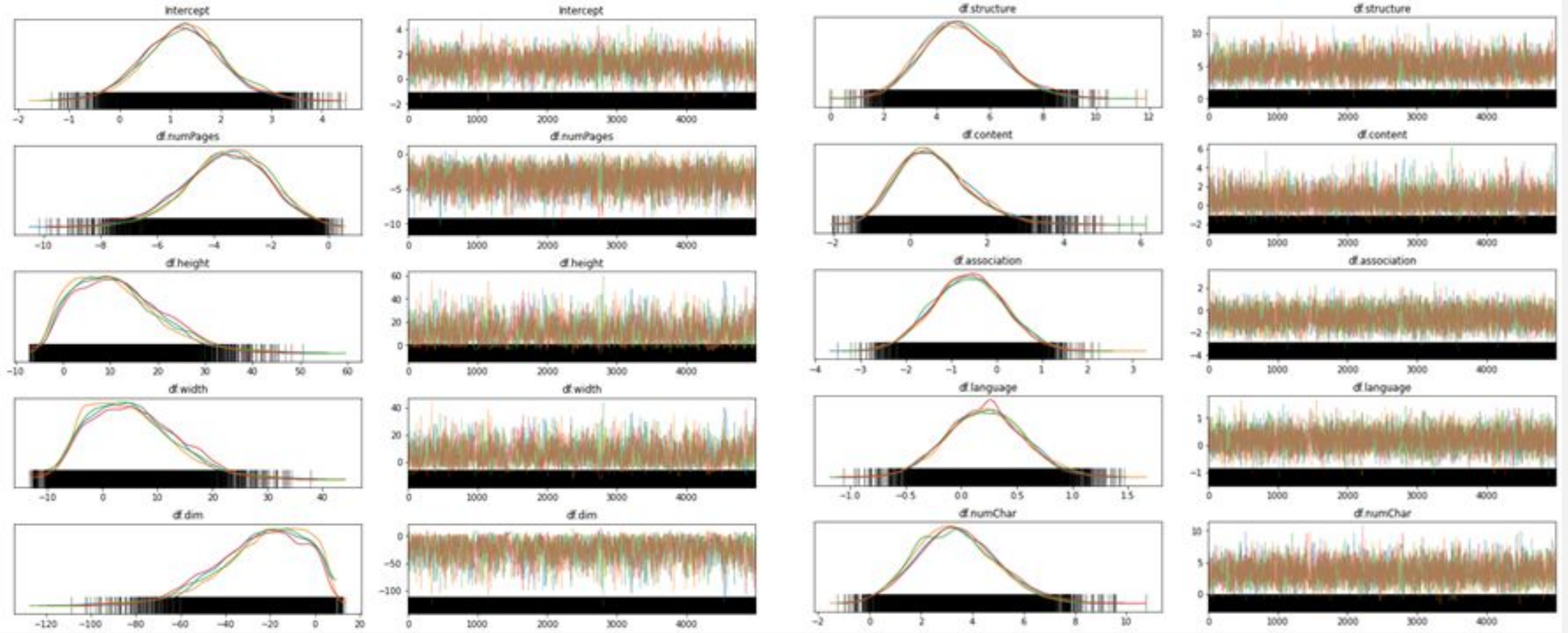


PyMC3

GLM: Logistic Regression : (sampling)

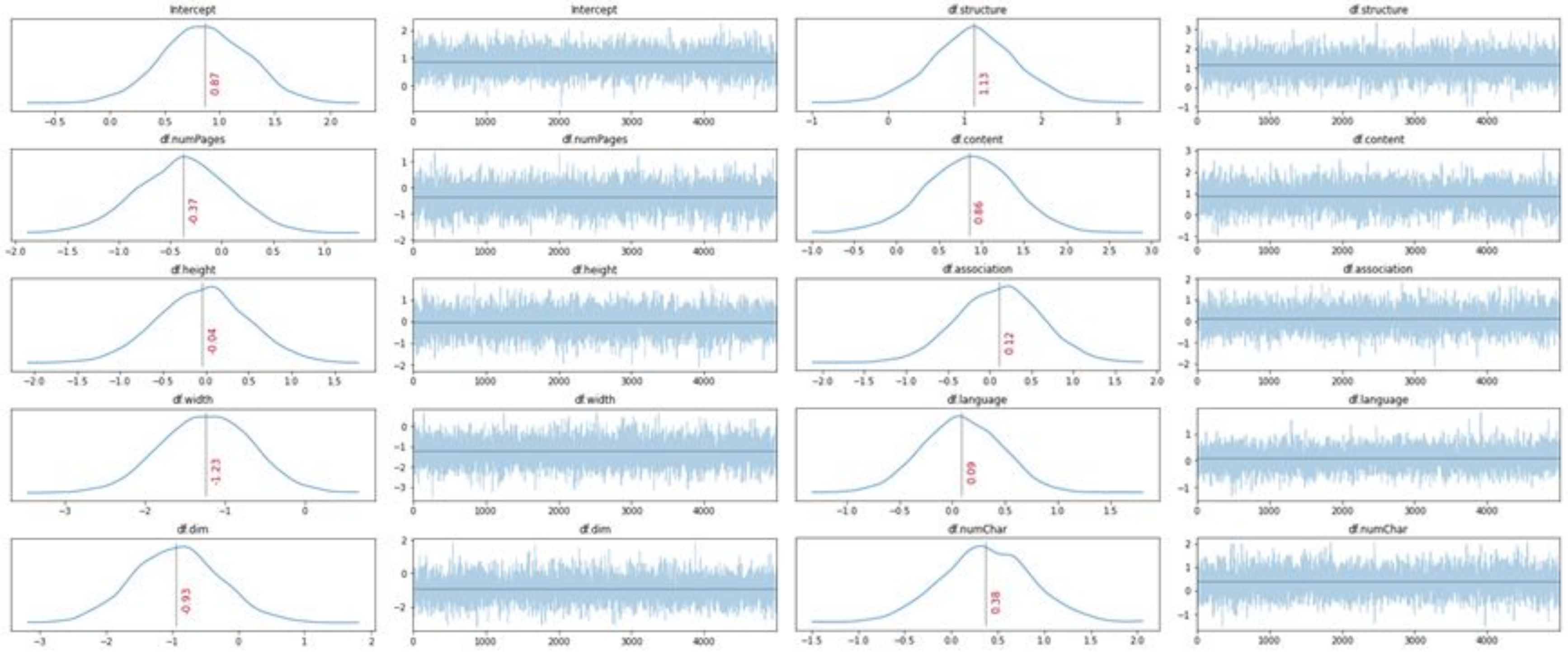
ADVI variational inference (optimization) ELBO

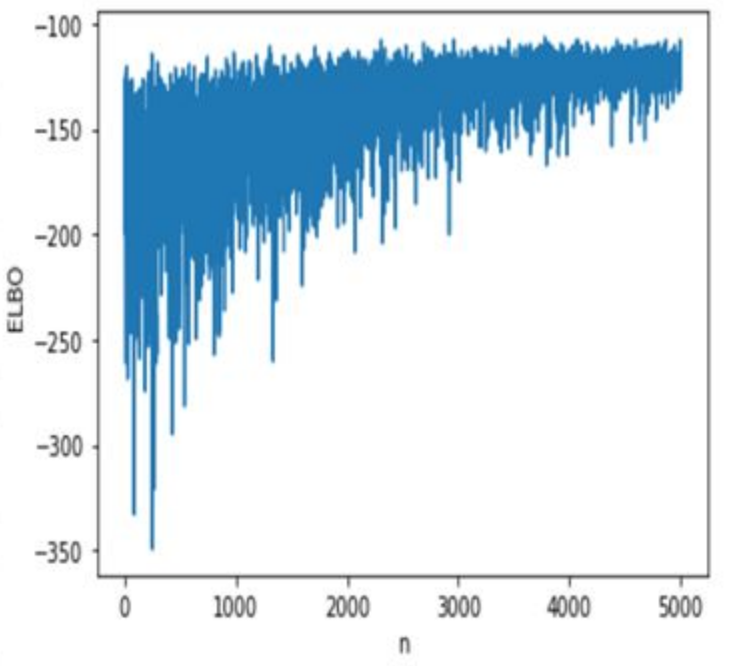
RESULTS - Full Model (Sampling)



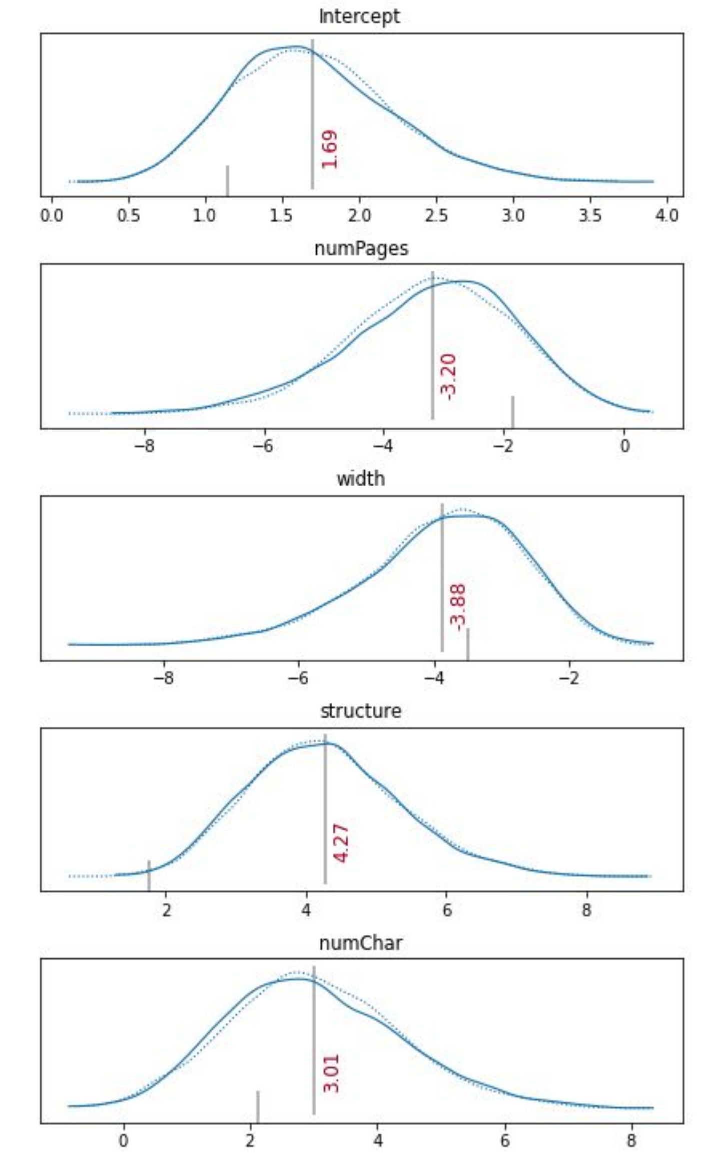
RESULTS - Full Model (ADVI)

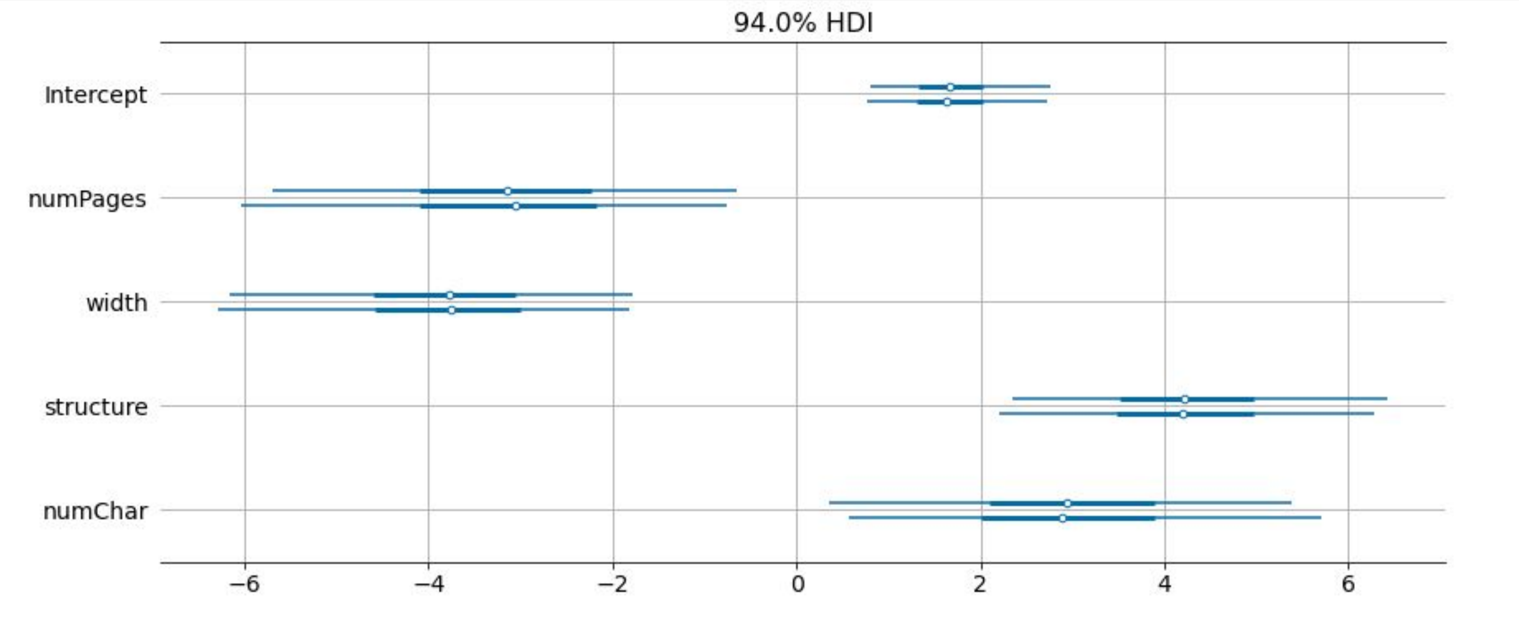
Variational Inference: ADVI



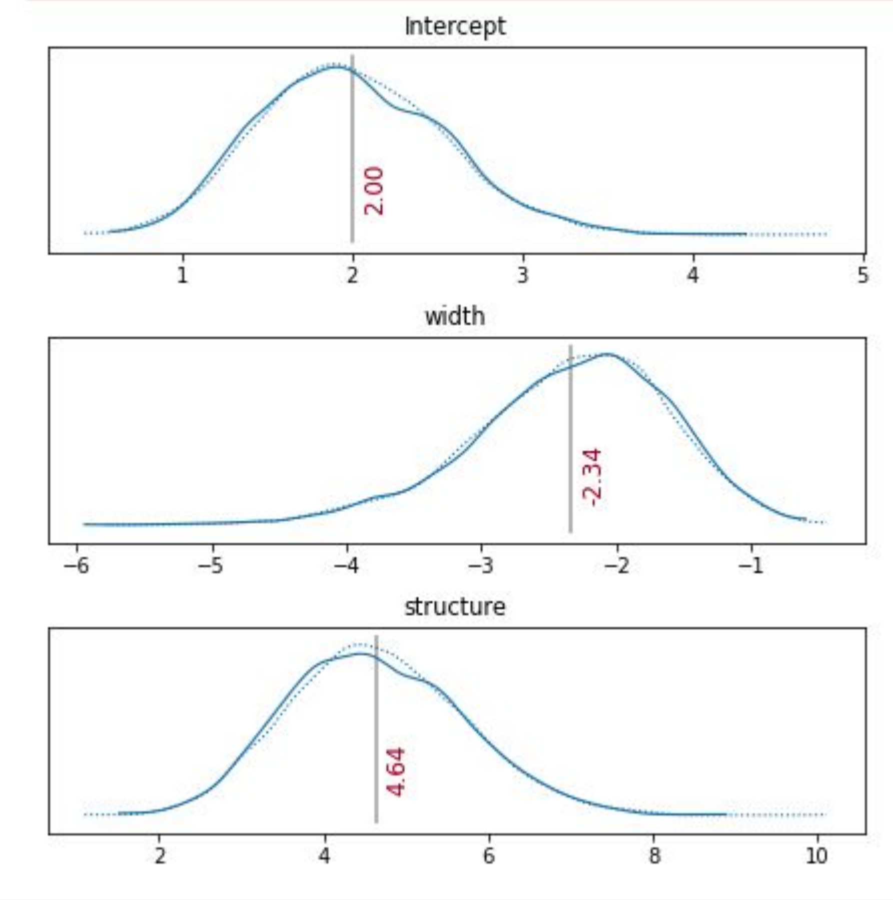


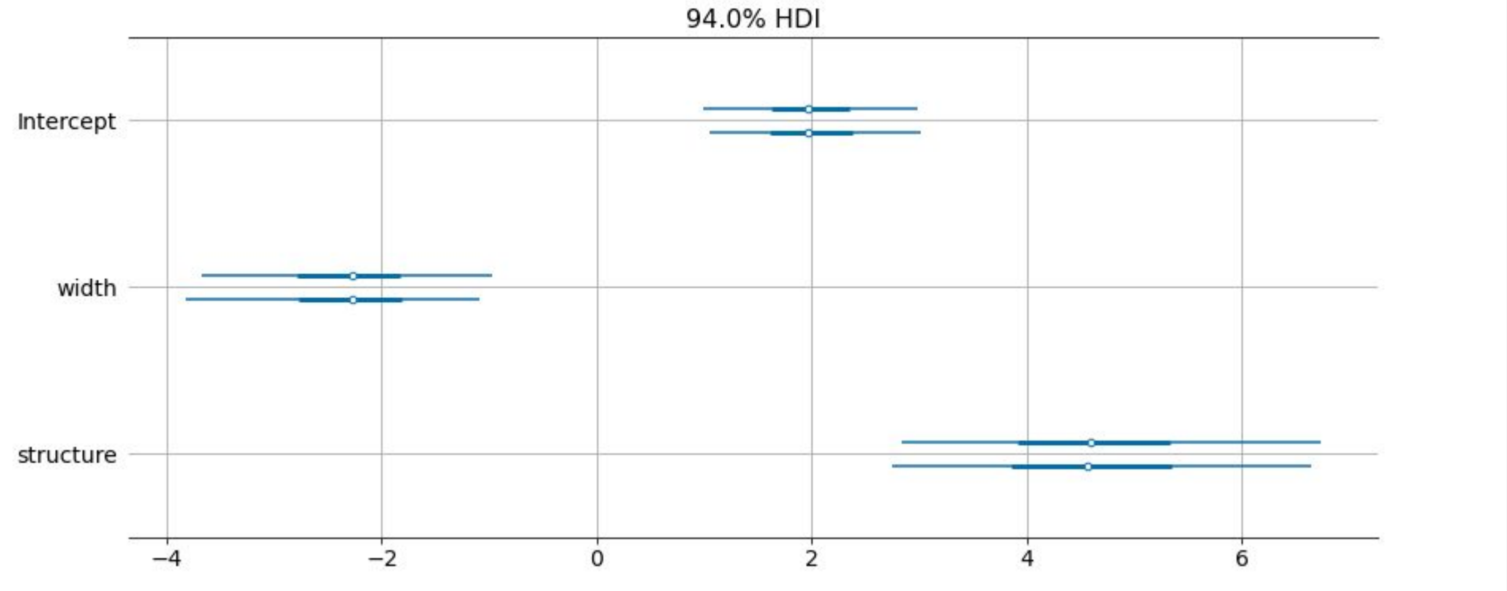
RESULTS - Model 1



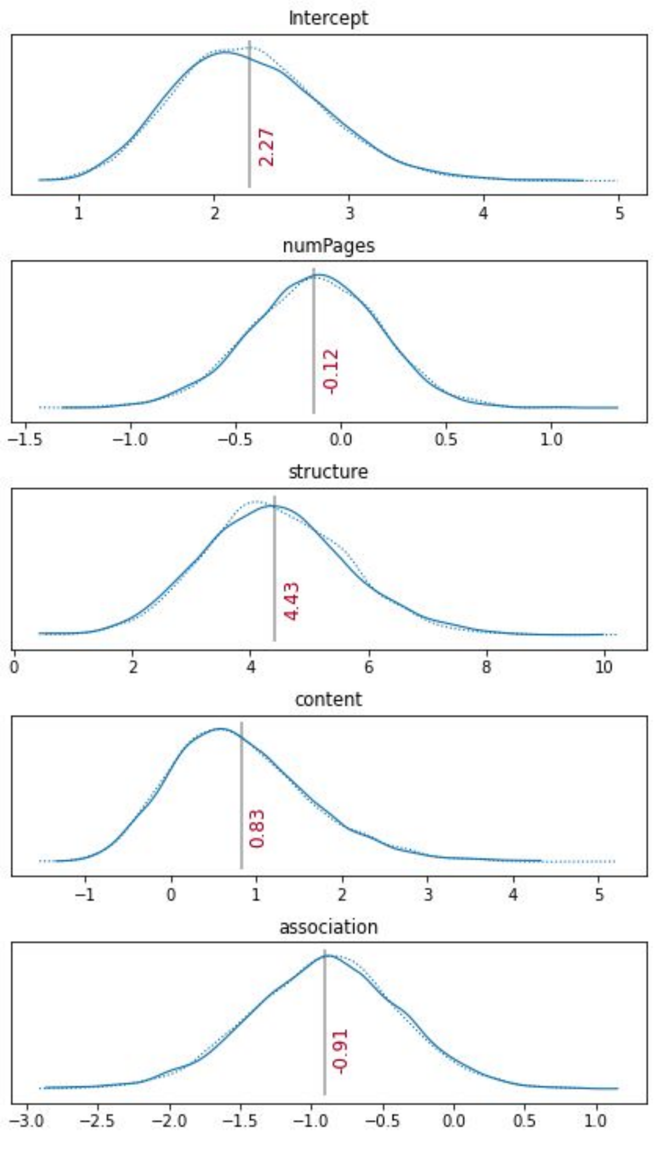


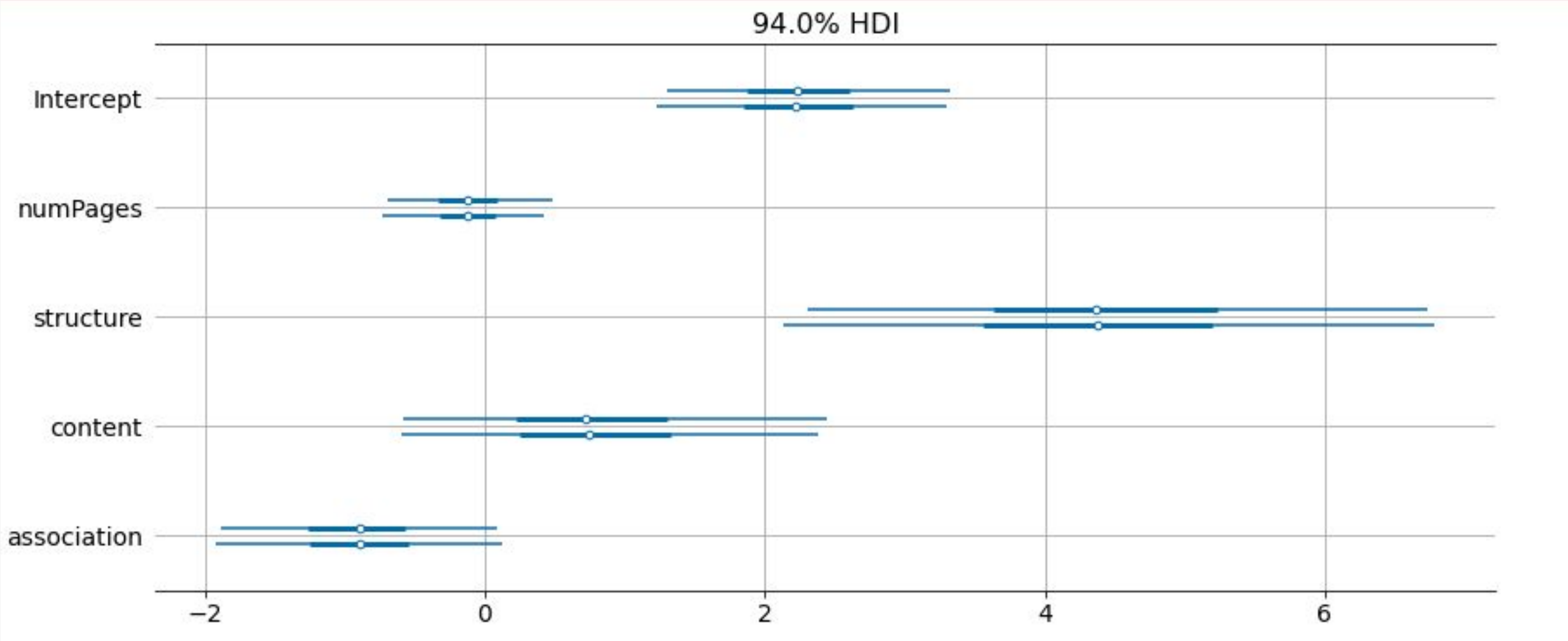
RESULTS - Model 2





RESULTS - Model 3





RESULTS - BMA - Top Models

Model	Features	Likelihood
1	number of pages, width, structure, number of characters	1.03e-23
2	width, structure	6.20e-23
3	number of pages, structure, number of characters	5.71e-24

RESULTS - WAIC

Model	Features	WAIC
1 - Top model from BMA	number of pages, width, structure, number of characters	95.0114
2 - 2nd best model from BMA	width, structure	95.291
3 - Selected out of curiosity	structure, content, association	119.888
4 - Full model	number of pages, height, width, dimension, structure, content, association, language, number of characters	95.1717
5 - Simple model (also selected out of curiosity)	structure	95.1717

PREDICTION COMPARISON

Model	**Features**	**Accuracy**
1 - Top model from BMA	number of pages, width, structure, number of characters	78.3%
2 - 2nd best model from BMA	width, structure	75%
3 - Selected out of curiosity	structure, content, association	80%
4 - Full model	number of pages, height, width, dimension, structure, content, association, language, number of characters	77%
5 - Simple model (also selected out of curiosity)	structure	76.7%

CONCLUSIONS

- Key findings:
  - Top feature - structure
  - Additional key features - number of pages, width
  - Format matters, in addition to content
  - Simple model may be reasonable, given comparable accuracy to others, to prioritize speed

- Future work:
  - Cross-validation for further comparison
  - Additional investigation into models focused on the language model (structure, content, association)
  - Use LDA dimension reduction to assess whether it generates improved results
  - Expand analysis to larger dataset
  - Address non-English language factors in the model