Austen Revisited:

A Computational Approach to Authorship Attribution

Teresa Whitesell

Capstone Presentation

Nashville Software School – Data Science Cohort 7

June 8, 2024

Introduction

> It is a truth universally acknowledged that a newly trained data scientist must be in want of a dataset.
>
> – Jane Austen, Pride and Prejudice (sort of)

Authorship Attribution

- The task of identifying the author of a given document
- Applications
    - Forensics analysis
    - Plagiarism detection
    - Copyright infringement

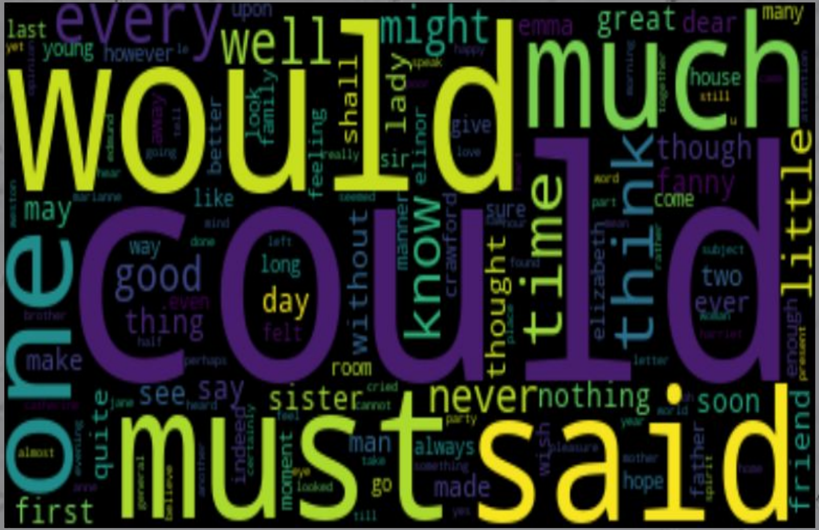
Experimental Design and Data Sources

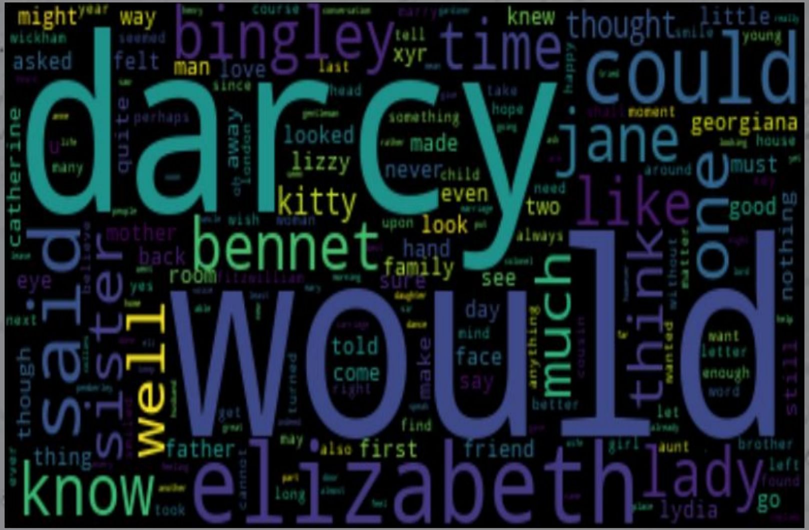
- Jane Austen texts
    - Well authenticated and available
    - 6 published novels
    - Project Gutenberg
- Texts from fan authors
    - Deliberately imitating Austen’s style
    - Stories based on Austen’s novels
    - Large corpus of work to choose from
    - Challenges: inconsistent formatting, inaccurate tagging

Data Collection and Processing

Text Preparation

- Metadata preparation
- Functions for reading files, tokenizing





Tokenization

- Use sent_tokenize to create list of sentences from a text
- Function to iterate through sentences and create sections no longer than the set character limit
- Returns token texts and assorted metadata
- Added to dataframe with text id and metadata from data source
- Created binary classification - Austen and Not Austen

Models

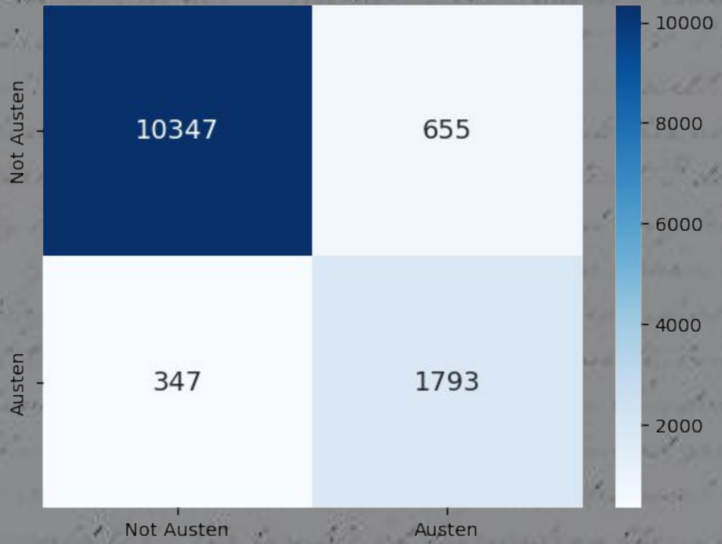
Model Training

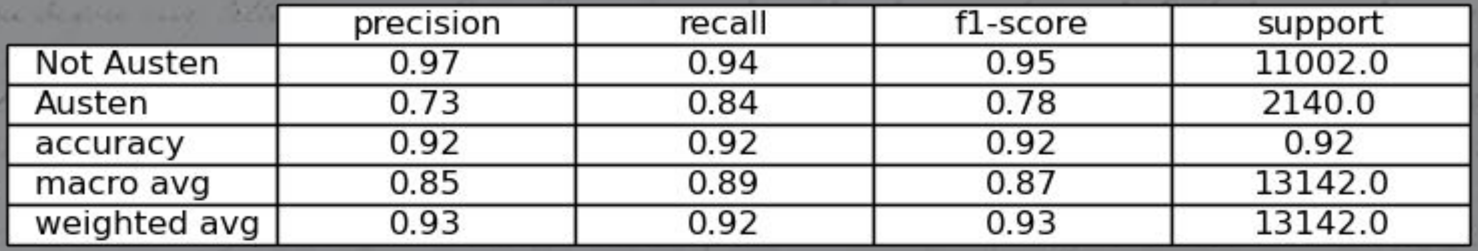
- “Leave one out” approach
  - Series of models
  - Train on all but one observation, test on the one
  - Predictions for each book based on the maximum amount of training data
- Model types
  - Multinomial Naive Bayes
  - XGBoost
- Vectorizers
  - Bag of words
  - TF-IDF

Results

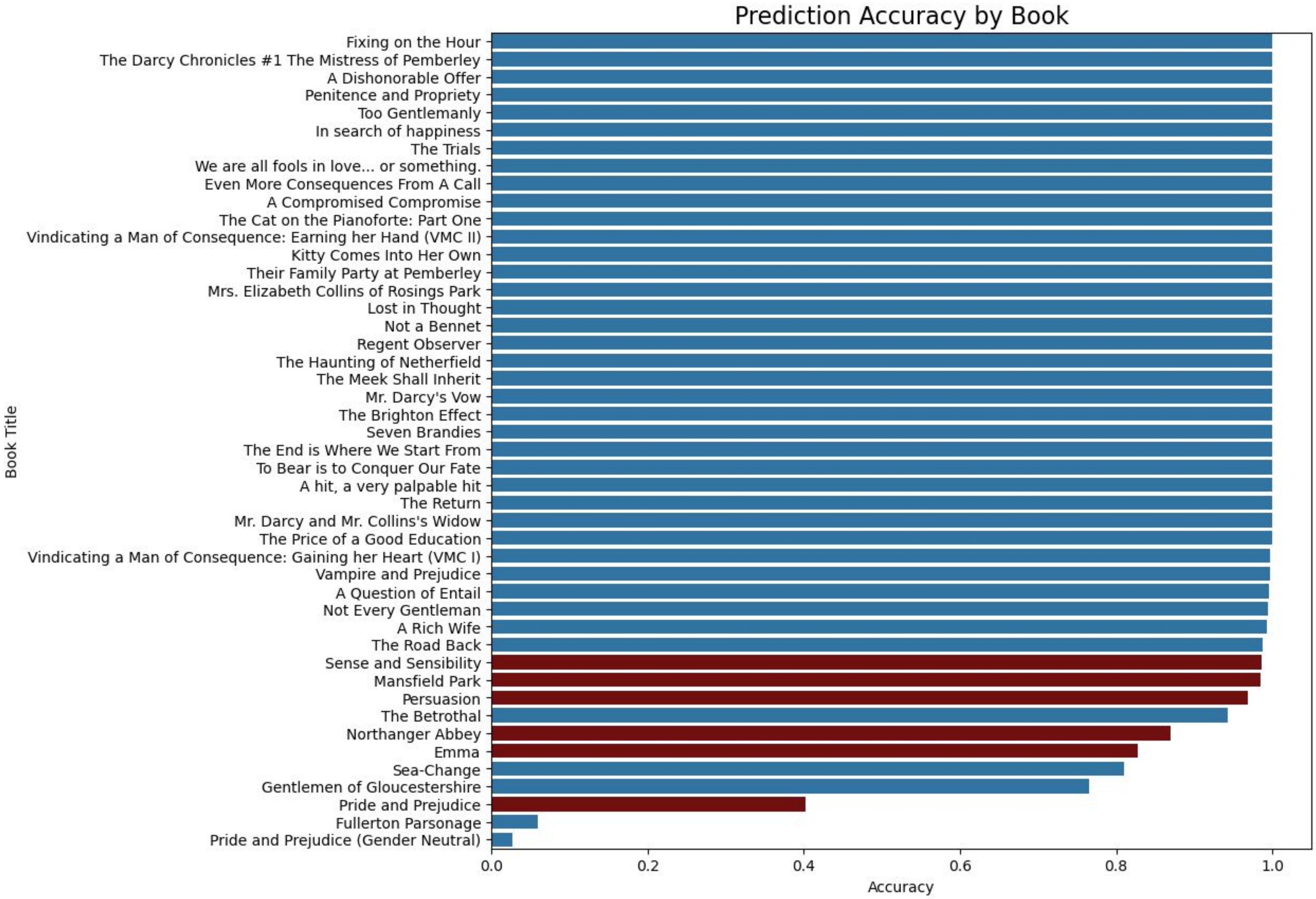
Results – Multinomial Naive Bayes

- Accuracy: 92.4%
- CountVectorizer()
- No text preprocessing



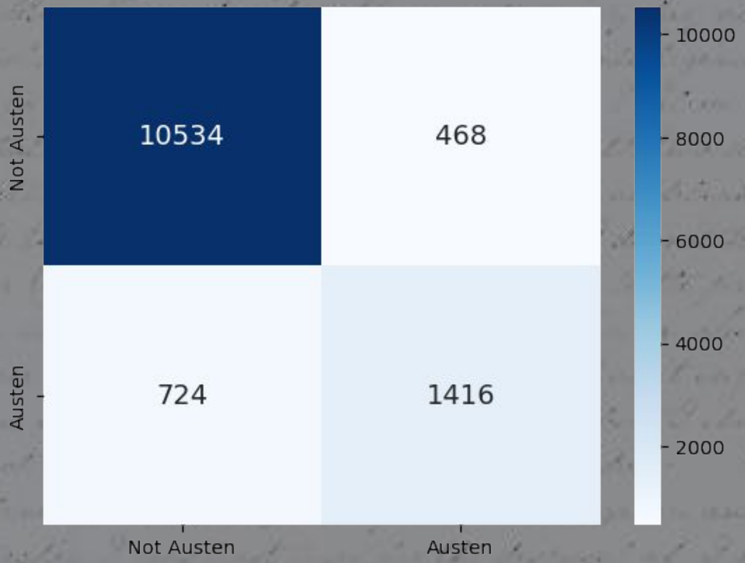


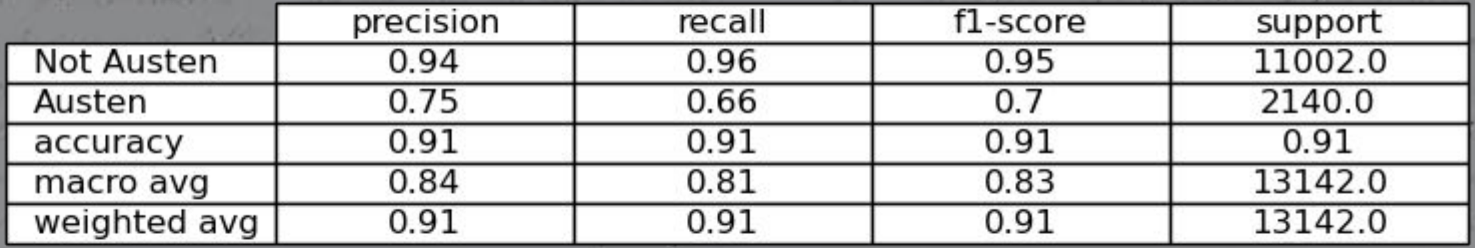
	precision	recall	f1-score	support
Not Austen	0.97	0.94	0.95	11002.0
Austen	0.73	0.84	0.78	2140.0
accuracy	0.92	0.92	0.92	0.92
macro avg	0.85	0.89	0.87	13142.0
weighted avg	0.93	0.92	0.93	13142.0


Prediction Accuracy by Book
Book Title
Fixing on the Hour
The Darcy Chronicles #1 The Mistress of Pemberley
A Dishonorable Offer
Penitence and Propriety
Too Gentlemanly
In search of happiness
The Trials
We are all fools in love... or something.
Even More Consequences From A Call
A Compromised Compromise
The Cat on the Pianoforte: Part One
Vindicating a Man of Consequence: Earning her Hand (VMC II)
Kitty Comes Into Her Own
Their Family Party at Pemberley
Mrs. Elizabeth Collins of Rosings Park
Lost in Thought
Not a Bennet
Regent Observer
The Haunting of Netherfield
The Meek Shall Inherit
Mr. Darcy's Vow
The Brighton Effect
Seven Brandies
The End is Where We Start From
To Bear is to Conquer Our Fate
A hit, a very palpable hit
The Return
Mr. Darcy and Mr. Collins's Widow
The Price of a Good Education
Vindicating a Man of Consequence: Gaining her Heart (VMC I)
Vampire and Prejudice
A Question of Entail
Not Every Gentleman
A Rich Wife
The Road Back
Sense and Sensibility
Mansfield Park
Persuasion
The Betrothal
Northanger Abbey
Emma
Sea-Change
Gentlemen of Gloucestershire
Pride and Prejudice
Fullerton Parsonage
Pride and Prejudice (Gender Neutral)
0.0
0.2
0.4
0.6
0.8
1.0
Accuracy

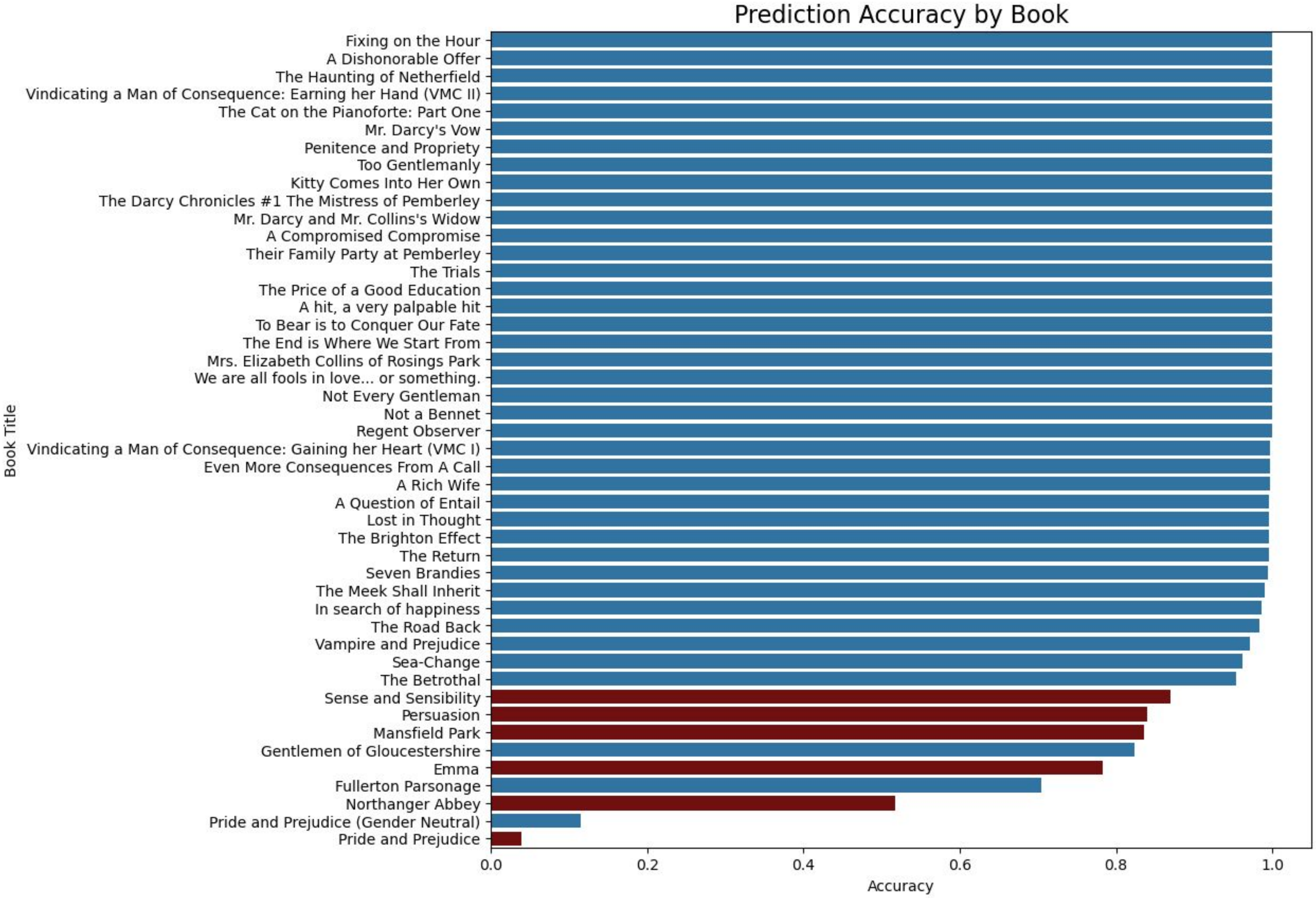
Results – XGBoost

- Accuracy: 90.9%
- TF-IDF
- No text pre-processing





	precision	recall	f1-score	support
Not Austen	0.94	0.96	0.95	11002.0
Austen	0.75	0.66	0.7	2140.0
accuracy	0.91	0.91	0.91	0.91
macro avg	0.84	0.81	0.83	13142.0
weighted avg	0.91	0.91	0.91	13142.0


Prediction Accuracy by Book
Book Title
Accuracy
Fixing on the Hour
A Dishonorable Offer
The Haunting of Netherfield
Vindicating a Man of Consequence: Earning her Hand (VMC II)
The Cat on the Pianoforte: Part One
Mr. Darcy's Vow
Penitence and Propriety
Too Gentlemanly
Kitty Comes Into Her Own
The Darcy Chronicles #1 The Mistress of Pemberley
Mr. Darcy and Mr. Collins's Widow
A Compromised Compromise
Their Family Party at Pemberley
The Trials
The Price of a Good Education
A hit, a very palpable hit
To Bear is to Conquer Our Fate
The End is Where We Start From
Mrs. Elizabeth Collins of Rosings Park
We are all fools in love... or something.
Not Every Gentleman
Not a Bennet
Regent Observer
Vindicating a Man of Consequence: Gaining her Heart (VMC I)
Even More Consequences From A Call
A Rich Wife
A Question of Entail
Lost in Thought
The Brighton Effect
The Return
Seven Brandies
The Meek Shall Inherit
In search of happiness
The Road Back
Vampire and Prejudice
Sea-Change
The Betrothal
Sense and Sensibility
Persuasion
Mansfield Park
Gentlemen of Gloucestershire
Emma
Fullerton Parsonage
Northanger Abbey
Pride and Prejudice (Gender Neutral)
Pride and Prejudice
0.0
0.2
0.4
0.6
0.8
1.0

Model Comparison Summary

Model	Overall Accuracy	Austen Accuracy	Training time
XGB, CV	91.6%	70%	2 min, 34 sec
XGB, TF-IDF	90.9%	66%	7 min, 42 sec
XGB, CV, text proc	90.4%	64%	1 min 39 sec
MNB, CV	92.4%	84%	1 min, 10 sec
MNB, CV, text proc	91.8%	81%	59 sec
XGB, TF-IDF, text proc	89.0%	54%	6 min 28 sec
MNB, TF-IDF, text proc	83.5%	0%	1 min 2 sec
MNB, TF-IDF	83.7%	0%	1 min, 46 sec

Conclusion

Conclusions and Next Steps

- Simple models can be surprisingly accurate

- Test TF-IDF with chi2 to extract most meaningful features
- Part-of-speech tagging
- Proper noun removal/replacement to determine role of character names
- Fine-tuning Huggingface model to compare performance

Thank you!