Genre Classification

Michael Rios
Dept. of Computer & Info. Science
Fordham University
Bronx, New York, USA
mrios27@fordham.edu

***Abstract*— The researcher attempted to classify books by genre based on a summary and genre classes provided by Wikipedia. The results were mixed at best. We have some consistency if we simplified the problem or used classes based on clustering algorithms instead of the classes provided by wikipedia. Unfortunately the computer generated classes don't necessarily have much meaning to humans. What the researcher could conclude was that a half a page summary is probably not sufficient information to base comparisons on, and that when trying to set classes, genres, the curation of data is extremely important.**

Keywords—test analysis, classification, naive bayes, svm

I. Introduction

The author is a librarian and wanted to try and use text analysis techniques on genre classification. It is very clearly his first time using these techniques so the conclusions are not very earth shattering but it was extremely illuminating to him.

II. Experimental methodology

A. Data set

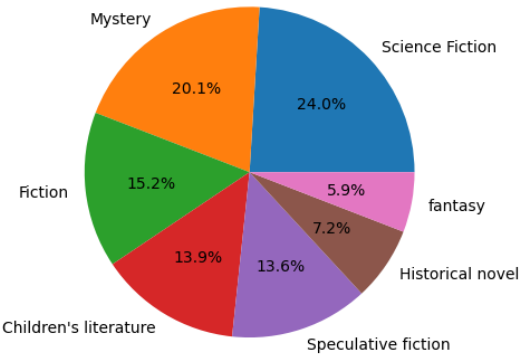
The data was compiled by David Bamman, Assistant Professor School of Information University of California, Berkeley. It consists of 16,559 books extracted from Wikipedia. Entries include a book title, author, publication date, genres (here referred to as classes, often multiple) and a plot summary. An example can be seen here https://www.cs.cmu.edu/~dbamman/booksummaries.html.

About three thousand of the entries had no class, genre, given these were dropped leaving us with 12,841 books. From those we kept the first class listed, if there was more than one, in hopes that this would be the most prominent class. This still left us with 180 classes, many of which only had one entry, which is essentially an impossible number to classify given the relatively small amount of data we had.

To overcome this the researcher hand clustered classes. Mystery, Crime Fiction, Thriller, Spy fiction and Detective fiction, were all merged into Mystery. Fiction and Novel were combined into just Fiction. Children's literature and Young adult literature into just Children's literature. Finally, Historical fiction, Alternate history, and War novel were all added to Historical novel

After this process was complete we were left with 7 classes.

Class	Number of entries
Science Fiction	2556
Mystery	2139
Fiction	1622
Children's literature	1480
Speculative fiction	1444
Historical novel	770
Fantasy	625

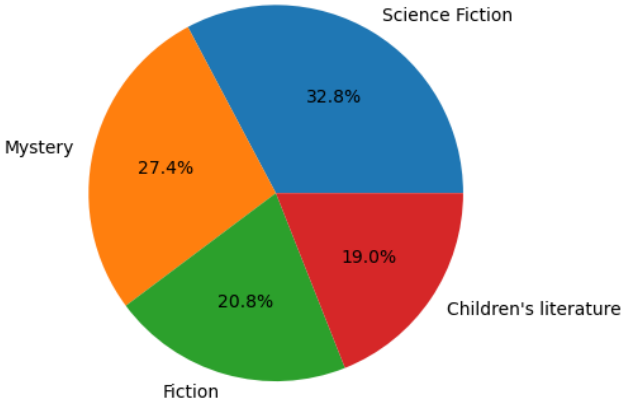


The smaller classes were simply cut, this left us with 10,636 books

This was far from perfect. Science fiction is the bulk of our data, we have a nebulous class of Fiction which does not supply us with much information as all our books are fiction, and we also have speculative fiction which seems like a close cousin to science fiction. Also Children's literature is not really the same type of category as mystery or science fiction, it's just the age level of the material which could presumably still be another class like science fiction. Still it was a place to start.

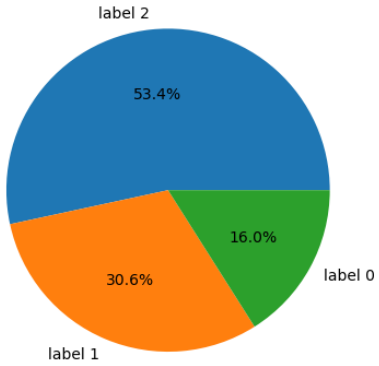
Once we experimented with this data we tried simplify the problem a bit by just keeping our top four categories

Science Fiction	2,556
Mystery	2,139
Fiction	1,622
Children's literature	1,480

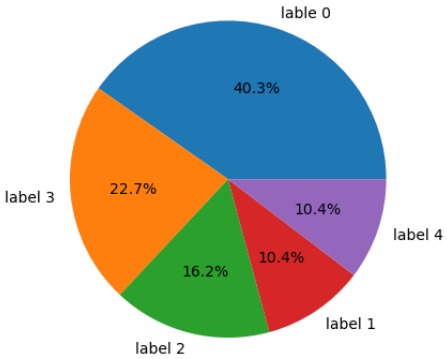


After this we used k-means clustering to reduce the 4 groups to three which actually made our data more unbalanced. It also needed to be investigated to see if these new categories have any meaning to humans.

Label 0	1,249
label 1	2,386
Label 2	4,162



We also clustered all the data original data with classes using k-means into 5 groups



Label 0	5,172
label 1	1,340
Label 2	2,079
Label 3	2,920
Label 4	1,330

B. algorithms

Following the advice of Sklearn , we used Naive Bayes and support vector machine (SVM) algorithms. SKlearn Gidsearch was used to check our parameters. Most of the parameters were related to how we should vectorize the plot summaries, though the alpha parameter of the SVM algorithm was also checked.

Random over sampling, ROS, was also used as a means of dealing with the class imbalance.

For Naive Bayes we searched through these possible, mostly, vector parameters

'vect__lowercase': (True, False),

'vect__stop_words': ('english', None),

'vect__min_df': [2, 5, 10, 15],

'vect__max_features': [2000, 3000, 4000, 5000],

'vect__ngram_range': [(1, 1), (1, 2), (1,3)],

'tfidf__use_idf': (True, False)

For the the unbalanced data these were the chosen parameters

vect__lowercase: True

vect__stop_words: 'english'

vect__min_df: 5

vect__max_features: 4000

vect__ngram_range: (1, 3)

tfidf__use_idf: True

For the ROS data these parameters were chosen

vect__lowercase: True

vect__stop_words: 'english'

vect__max_features: 5000

vect__min_df: 2

vect__ngram_range: (1, 1)

tfidf__use_idf: True

For SVM the possible choices of parameters were

'vect__lowercase': (True, False),

'vect__stop_words': ('english', None),

'vect__min_df': [2, 5, 10, 15],

'vect__max_features': [2000, 3000, 4000, 5000],

'vect__ngram_range': [(1, 1), (1, 2), (1,3)],

'tfidf__use_idf': (True, False),

'clf__alpha': (1e-1, 1e-2, 1e-3),

With unbalanced data these were chosen

vect__lowercase: True

vect__stop_words: 'english'

vect__min_df: 2

vect__max_features: 5000

vect__ngram_range: (1, 3)

tfidf__use_idf: True

clf__alpha: 0.001

With ROS these were chosen

vect__lowercase: True

vect__stop_words: 'english'

vect__min_df: 2

vect__max_features: 5000

vect__ngram_range: (1, 1)

tfidf__use_idf: True

clf__alpha: 0.001

C. Evaluation procedure

we stuck with the default metrics displayed in classification report, precision, recall, f1-score, their macro and weighted averages, a heat map to help us visualize our data, and finally the accuracy as it makes it much easier to get a quick overview of how the algorithm did

Ideally our test data would have been divided into validation and test. Unfortunately I ran out of time.

III. Results

For the sake of ease we will first just show how the accuracy scores.

The results for our first data set in which we hand clustered the data into 7 categories

Algo	data	parameters	Accuracy
Naive Bayes	Unbalanced	Default	.378
Naive Bayes	Unbalanced	Chosen	.597
Naive Bayes	ROS	Default	.63
Naive Bayes	ROS	Chosen	.614

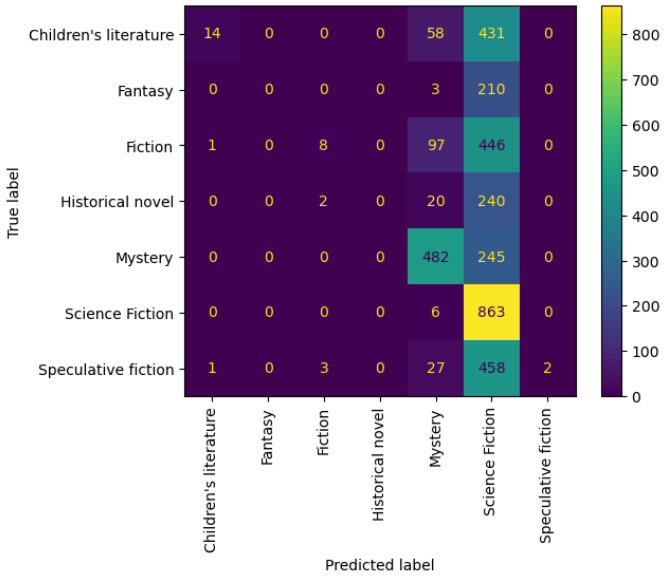
SVM	Unbalanced	Default	.648
SVM	Unbalanced	Chosen	.596
SVM	ROS	Default	.651
SVM	ROS	Chosen	.599

We see that using naive bayes with unbalanced data gets us the worst score and that SVM, works much better right off the bat but we cannot really improve on it.

For a more complete look at the data see the following page

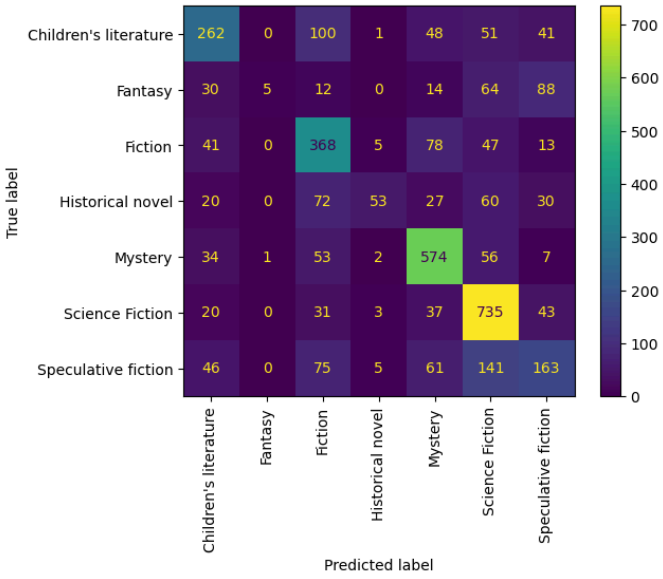
Algo: Naive Bayes | Data: Unbalanced | Parameters: Default |

	precision	recall	f1-score	support
Children's literature	0.88	0.03	0.05	503
Fantasy	0.00	0.00	0.00	213
Fiction	0.62	0.01	0.03	552
Historical novel	0.00	0.00	0.00	262
Mystery	0.70	0.66	0.68	727
Science Fiction	0.30	0.99	0.46	869
Speculative fiction	1.00	0.00	0.01	491
accuracy			0.38	3617
macro avg	0.50	0.24	0.18	3617
weighted avg	0.56	0.38	0.26	3617



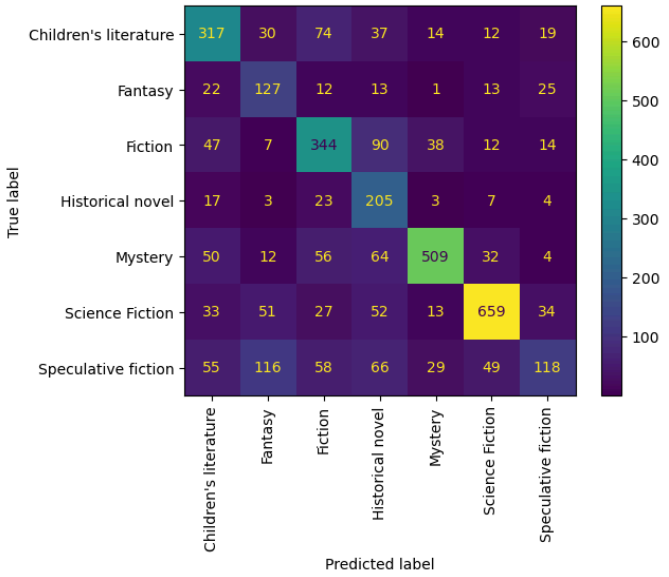
Algo: Naive Bayes | Data: Unbalanced | Parameters: Chosen | A

	precision	recall	f1-score	support
Children's literature	0.58	0.52	0.55	503
Fantasy	0.83	0.02	0.05	213
Fiction	0.52	0.67	0.58	552
Historical novel	0.77	0.20	0.32	262
Mystery	0.68	0.79	0.73	727
Science Fiction	0.64	0.85	0.73	869
Speculative fiction	0.42	0.33	0.37	491
accuracy			0.60	3617
macro avg	0.63	0.48	0.48	3617
weighted avg	0.61	0.60	0.56	3617



Algo: Naive Bayes | Data: ROS | Parameters: Default | Accurac

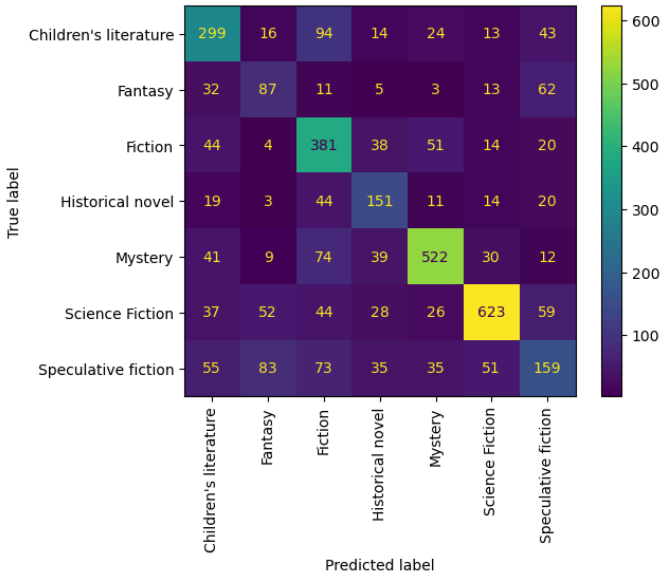
	precision	recall	f1-score	support
Children's literature	0.59	0.63	0.61	503
Fantasy	0.37	0.60	0.45	213
Fiction	0.58	0.62	0.60	552
Historical novel	0.39	0.78	0.52	262
Mystery	0.84	0.70	0.76	727
Science Fiction	0.84	0.76	0.80	869
Speculative fiction	0.54	0.24	0.33	491
accuracy			0.63	3617
macro avg	0.59	0.62	0.58	3617
weighted avg	0.66	0.63	0.63	3617



Algo: Naive Bayes | Data: ROS | Parameters: Chosen | Accuracy:

---

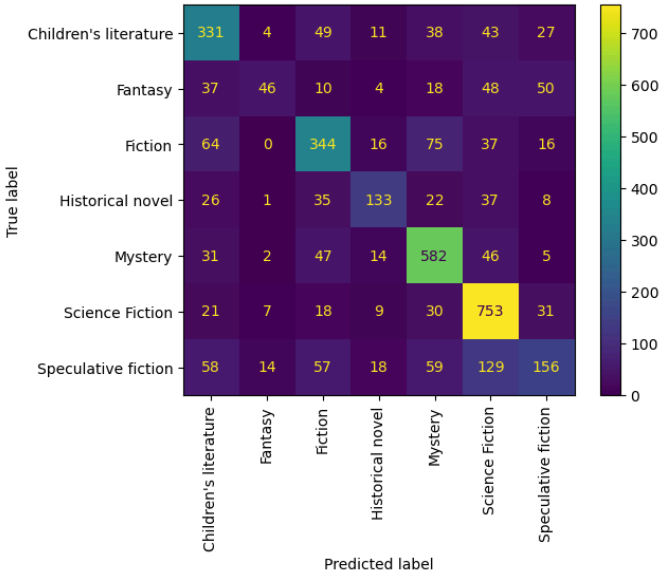
	precision	recall	f1-score	support
Children's literature	0.57	0.59	0.58	503
Fantasy	0.34	0.41	0.37	213
Fiction	0.53	0.69	0.60	552
Historical novel	0.49	0.58	0.53	262
Mystery	0.78	0.72	0.75	727
Science Fiction	0.82	0.72	0.77	869
Speculative fiction	0.42	0.32	0.37	491
accuracy			0.61	3617
macro avg	0.56	0.58	0.57	3617
weighted avg	0.63	0.61	0.62	3617



Algo: SVM | Data: Unbalanced | Parameters: Default | Accuracy

---

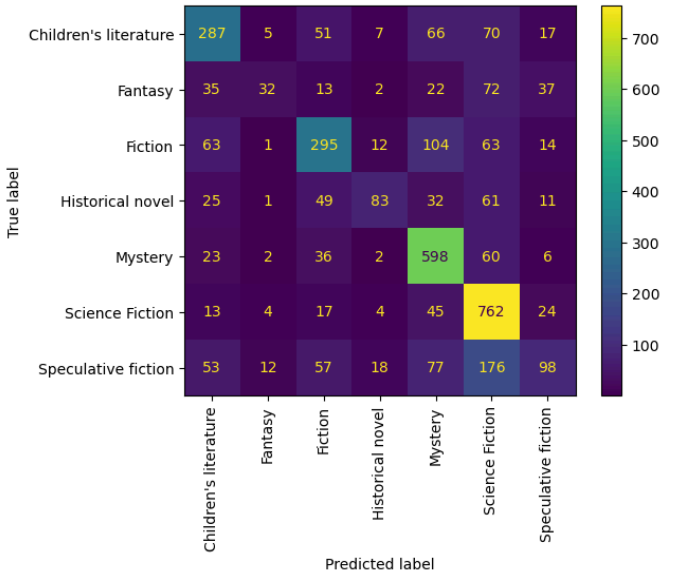
	precision	recall	f1-score	support
Children's literature	0.58	0.66	0.62	503
Fantasy	0.62	0.22	0.32	213
Fiction	0.61	0.62	0.62	552
Historical novel	0.65	0.51	0.57	262
Mystery	0.71	0.80	0.75	727
Science Fiction	0.69	0.87	0.77	869
Speculative fiction	0.53	0.32	0.40	491
accuracy			0.65	3617
macro avg	0.63	0.57	0.58	3617
weighted avg	0.64	0.65	0.63	3617



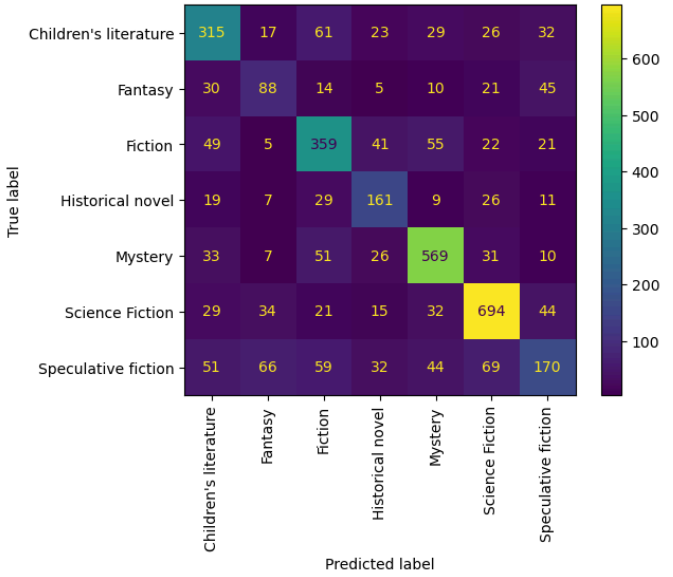
Algo: SVM | Data: Unbalanced | Parameters: Chosen | Accuracy:

---

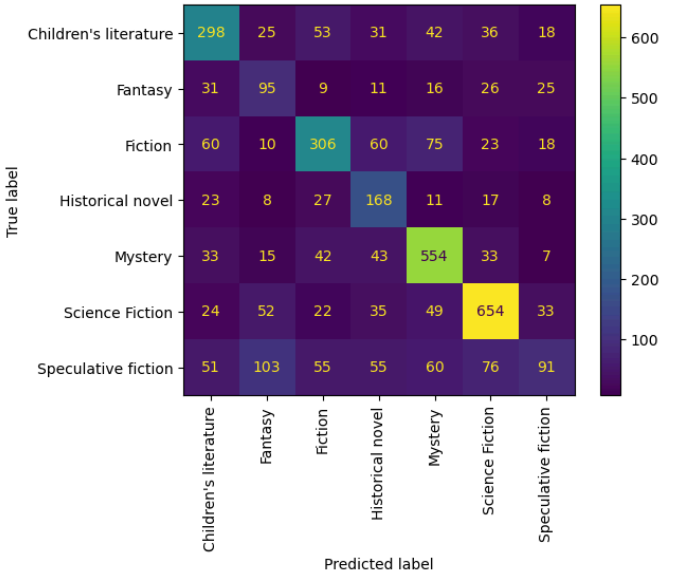
	precision	recall	f1-score	support
Children's literature	0.58	0.57	0.57	503
Fantasy	0.56	0.15	0.24	213
Fiction	0.57	0.53	0.55	552
Historical novel	0.65	0.32	0.43	262
Mystery	0.63	0.82	0.72	727
Science Fiction	0.60	0.88	0.71	869
Speculative fiction	0.47	0.20	0.28	491
accuracy			0.60	3617
macro avg	0.58	0.50	0.50	3617
weighted avg	0.58	0.60	0.56	3617




```
----------------------------------------------------------------
Algo:  SVM | Data:  ROS | Parameters:  Default | Accuracy:  0.65
----------------------------------------------------------------
                      precision    recall  f1-score   support

Children's literature       0.60      0.63      0.61       503
              Fantasy       0.39      0.41      0.40       213
              Fiction       0.60      0.65      0.63       552
     Historical novel       0.53      0.61      0.57       262
              Mystery       0.76      0.78      0.77       727
      Science Fiction       0.78      0.80      0.79       869
Speculative fiction         0.51      0.35      0.41       491

             accuracy                           0.65      3617
            macro avg       0.60      0.60      0.60      3617
         weighted avg       0.65      0.65      0.65      3617
```




```
----------------------------------------------------------------
Algo:  SVM | Data:  ROS | Parameters:  Chosen | Accuracy:  0.59
----------------------------------------------------------------
                      precision    recall  f1-score   support

Children's literature       0.57      0.59      0.58       503
              Fantasy       0.31      0.45      0.36       213
              Fiction       0.60      0.55      0.57       552
     Historical novel       0.42      0.64      0.51       262
              Mystery       0.69      0.76      0.72       727
      Science Fiction       0.76      0.75      0.75       869
Speculative fiction         0.46      0.19      0.26       491

             accuracy                           0.60      3617
            macro avg       0.54      0.56      0.54      3617
         weighted avg       0.60      0.60      0.59      3617
```



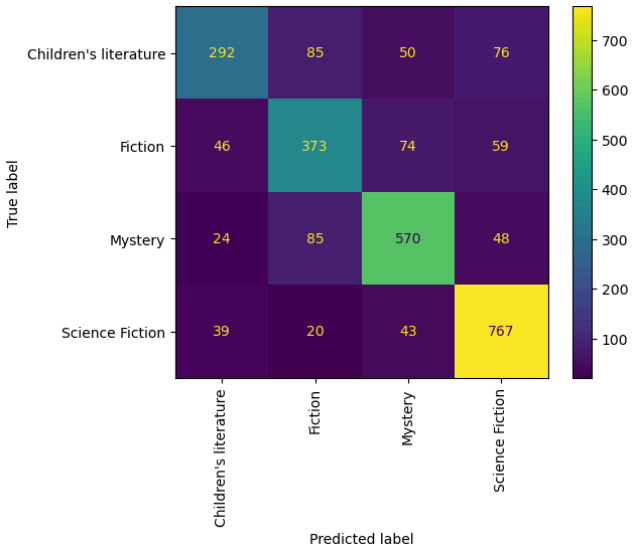
If you look more closely at this data, what sticks out is that our smallest three categories, speculative fiction, fantasy and historical fiction, tend to fare the worst. Often just being gobbled up by Science fiction. The large eat the small so to speak, even with ROS. So we tried with just the top for categories, science fiction, mystery, fiction, and children's literature. We just used two of the well performing algorithms from our first batch.

We do see immediate improvement

Algo	data	parameters	Accuracy
Naive Bayes	Unbalanced -Simplified	Chosen	.755
SVM	Unbalanced -Simplified	Default	.776

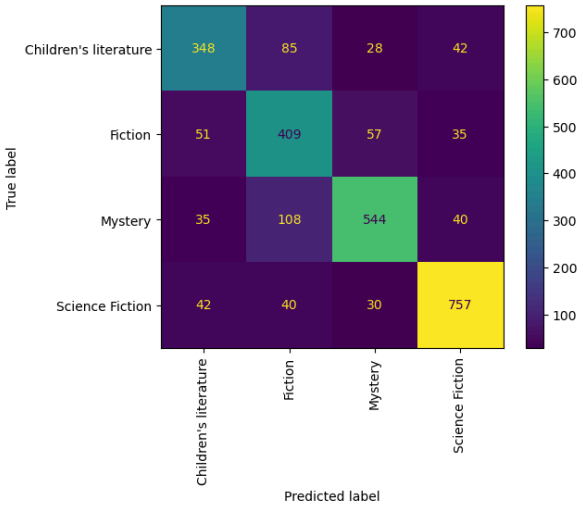
Algo: Naive Bayes | Data: Unbalanced and Simplified | Parameters

	precision	recall	f1-score	support
Children's literature	0.73	0.58	0.65	503
Fiction	0.66	0.68	0.67	552
Mystery	0.77	0.78	0.78	727
Science Fiction	0.81	0.88	0.84	869
accuracy			0.76	2651
macro avg	0.74	0.73	0.73	2651
weighted avg	0.75	0.76	0.75	2651



Algo: SVM | Data: Unbalanced and Simplified | Parameters: Def

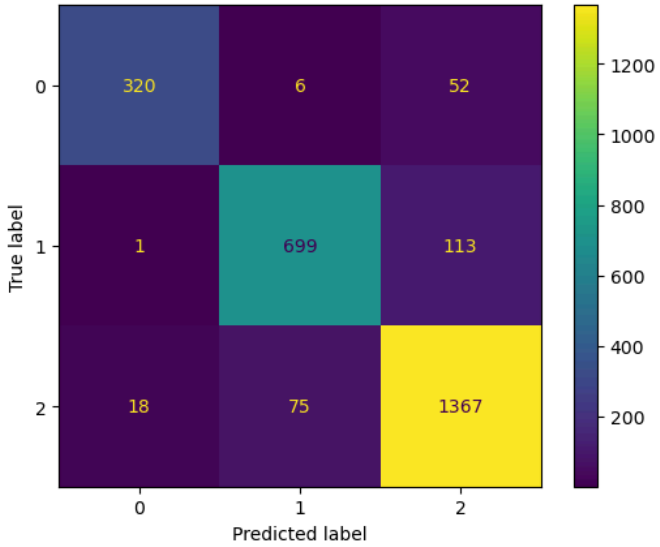
	precision	recall	f1-score	support
Children's literature	0.73	0.69	0.71	503
Fiction	0.64	0.74	0.69	552
Mystery	0.83	0.75	0.78	727
Science Fiction	0.87	0.87	0.87	869
accuracy			0.78	2651
macro avg	0.76	0.76	0.76	2651
weighted avg	0.78	0.78	0.78	2651



In an attempt to further simplify the problem and improve results we then used the K-means to cluster these four groups into three

Algo: SVM | Data: Simple Clustered | Parameters: De

	precision	recall	f1-score	support
0	0.94	0.85	0.89	378
1	0.90	0.86	0.88	813
2	0.89	0.94	0.91	1460
accuracy			0.90	2651
macro avg	0.91	0.88	0.89	2651
weighted avg	0.90	0.90	0.90	2651



This gets us the best results we have seen, but do the categories mean anything to a human. That is harder to say our smallest category seems to be essentially all science fiction so that is good.

Class: 0
Total entries: 1112

Science Fiction	1066
Mystery	21
Children's literature	17
Fiction	8

Our largest is mostly mystery and science fiction.
Class: 2
Total entries: 4294

Mystery	1603
Science Fiction	1283
Fiction	766
Children's literature	642

Finally the middle one is mostly fiction and children's literature with very little science fiction:

Class: 1
Total entries: 2391

Fiction	848
Children's literature	821
Mystery	515
Science Fiction	207

So we have a small, essentially pure science fiction class, a larger mostly non science fiction category, that hopefully skews younger and then the largest group which has books that are a combination of mystery and science fiction.

If we take a closer look at some of the actual titles classified as Children's literature from class 1 we see such titles as Emma, and Persuasion both by Jane Austen, the Scarlet Letter by Nathiel Hawthorne, as well as something more typically thought of as children's literature such as the Phantom Tollbooth by Norton Juster. While the fiction from class 1 includes titles such as To Kill a Mockingbird by Harper Lee, The Great Gatsby by F. Scott Fitzgerald and East of Eden by John Stienbeck.

We are seeing what are essentially considered classic books that are often assigned reading in high school and/or college on both lists, perhaps this being what ties them together. A good title for this class might be required reading and some children's books.

A look at class 2 shows titles such as Dracula by Bram Stoker, A Scanner Darkly, by Phillip K. Dick and Dirk Gently's Holistic Detective agency by Douglas Adams all of which could be seens as being on the edge of the science fiction genre, and thus put with the mysteries, which includes a lot of James Bond books. So class 2 seems to hold books that are not pure science fiction like those of class 0 but also have more action than books in class 1.

There is value to these classes but it needs to be teased out.

We also used the k-means algorithm to cluster all 12K books with a given class into 5 clusters.

We still get one that is relatively small and mostly science fiction, we still get a large one that is about 40% of our books and is dominated by science fiction, crime fiction, mystery and speculative fiction, and we still get one that has relatively little science fiction. So not surprisingly some of the same trends emerge.

Largest seems similar to the largest one, Class 2 from our first clustering.

Class: 0
Total entries: 5172
Top entries

Science Fiction	767
Crime Fiction	529
Mystery	520
Speculative fiction	508
Children's literature	461

Mostly science fiction like class 0 from previous clustering.

Class: 1
Total entries: 1340

Science Fiction	1056
Speculative fiction	69
Alternate history	34
Fantasy	20
Thriller	19

We also get.

Class: 2
Total entries: 2079
Top entries

Science Fiction	349
Novel	241
Fiction	225
Speculative fiction	206

This is unlike anything from the previous clustering. If we look at a few science fiction titles we see the Handmaid Tale by Margaret Atwood, It can't happen here by Sinclair Lewis and Cat's cradle by Kurt Vonnegut, so a bit more literary than your average science fiction novels and perhaps what ties it to the novel and fiction section

This seems like class 1 from the previous clustering, the least science fiction we see in any category.

Class: 3
Total entries: 2920
Tope entries

Children's literature	492
Novel	350
Fiction	333
Speculative fiction	232
Science Fiction	162

Again a new class that seems different from what we saw in the first clustering.

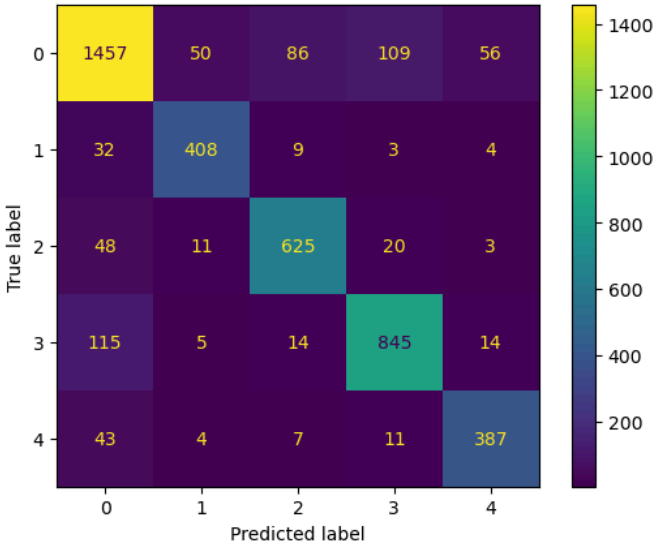
Class: 4
Total entries: 1330

Speculative fiction	429
Fantasy	234
Science Fiction	222
Children's literature	141

Here we see all the wizard of Oz books and C.S. Lewis Lion, the Witch and the Wardrobe series marked as speculative fiction, which seems to be the link to your fantasy and children's literature.

When we run our classification model we do pretty good, not as good as with the clustered and simplified data but pretty good.

```
-------------------------------------------------------
Algo:  SVM | Data:  All Clustered | Parameters:  Default
-------------------------------------------------------
              precision    recall  f1-score   support

           0       0.86      0.83      0.84      1758
           1       0.85      0.89      0.87       456
           2       0.84      0.88      0.86       707
           3       0.86      0.85      0.85       993
           4       0.83      0.86      0.84       452

    accuracy                           0.85      4366
   macro avg       0.85      0.86      0.86      4366
weighted avg       0.85      0.85      0.85      4366
```



This makes sense if they have been clustered by k-means; their vectors have to be fairly close, so it should be easier to predict which goes where.

Though that still leaves us with the problem of extracting meaning from the clusters. Which we did but is a much messier and subjective process than one would like. Also since we are just going on plot summaries, we are not really comparing books just however someone has summarized the plot which is arguably its own genre of writing.

IV: Conclusion

Classification can be done, even for genre classifications in fiction but ideally one is given more data than half page summaries. Reading in larger chucks of the actual text would have surely helped us build better models.

Also your model will be built around the data you give it. I would never have put the Wizard of Oz books in speculative fiction. I would call them fantasy so a more thorough checking of the classes in the beginning would have better informed me of what I should expect in terms of overlap amongst genres.

Still by simplifying the problem we were able to return better results. So one of the lessons learned is simply what useful question can you ask of your data, in this case is this a science fiction book or not may have been a more realistic goal.

Also much like the SKlearn documentation stated, SVM algorithm seems to work better than Naive Bayes.

References

[1] De Luca, E. W., et al. "Teaching an Algorithm How to Catalog a Book." Computers, vol. 10, no. 11, Nov. 2021. EBSCOhost, https://doi-org.avoserv2.library.fordham.edu/10.3390/computers10110155.

[2] Yelton, Andromeda. "A Simple Scheme for Book Classification Using Wikipedia." Information Technology & Libraries, vol. 30, no. 1, Mar. 2011, pp. 7–15. EBSCOhost, https://doi-org.avoserv2.library.fordham.edu/10.6017/ital.v30i1.3040.

[3] Working With Text Data. Scikit Learn, https://scikit-learn.org/stable/tutorial/text_analytics/working_with_text_data.html, accessed 10 December 2023.

[4] Ganguly, Sirshendu. "Book-Genre-Prediction", Github, https://github.com/chikne97/Book-Genre-Prediction/tree/master accessed 11 December 2023