Literary Period Classification

June 4, 2010

Caitlin Colgrove colgrove@stanford.edu
Sheldon Chang sheldonc@stanford.edu
Phumchanit (Yiam) Watanaprakornkul yiam@stanford.edu

I. Introduction

Our project deals with the issue of genre classification for English language fictional texts. We focused on attempting to classify these works, with an emphasis on novels, into the literary traditions that they represent. The four literary movements we worked with are: Augustan, Romantic, Naturalist, and Modernist. These periods were chosen because they are quite distinct in terms of their content and style, and there are plenty of available authors and texts within each epoch from our corpus, Project Gutenberg.

We classified texts by using a variety of features, many of which were taken from the literature on authorship attribution and stylometry, since this style of literary genre classification had not been done before (see the "Related Work" section for more details). We then used these features and trained a Maximum Entropy classifier on texts from each era, and then used a test set of other texts to evaluate performance. In selecting our features, we initially largely avoided features that would result in a classifier that evaluated by topic. However, we ended up including some features that would classify by topic (such as the unigram features) because so much of literary genre classification has to do not only with the style of the text, but also the subject matter and content (see table below).

In order to evaluate the performance of our classifier as a literary *genre* classifier, and not just an authorship attributer, we ran various experiments that attempted to control for the author, such as eliminating a Modernist author completely from the training set for modernism and then running the classifier on texts by this author.

II. Corpus

We limited ourselves to texts from Project Gutenberg, which restricted the time period for texts to before the early to mid 20th century. We also posed some additional constraints. We chose texts by author, selecting authors who are seen as representative or leaders of their movements. We allowed both American and British authors, however we did not use any works which were not written originally in English because we felt that the stylistic interpretation of the translator would have a significant affect on the translated text. We also did not include authors whose movement was ambiguous or who were from the same time period but not a part of the movement (as it would most likely be described from a literary standpoint).

	Augustan	**Romantic**	**Naturalist**	**Modernist**
Time Period	18th century	Early 19th	Late 19th, early 20th	Early 20th
Representative Authors	Alexander Pope, Jonathan Swift, Daniel Defoe, Samuel Richardson,	Edgar Allen Poe, Nathaniel Hawthorne, Mary Shelley, Washington	Edith Wharton, Jack London, Henry Adams, Frank Norris, Theodore Dreiser,	Ken Kesey, F. Scott Fitzgerald, Ernest Hemingway, Gertrude Stein,

	Henry Fielding, Laurence Sterne, Tobias Smollett, Sarah Fielding	Irving, James Fenimore Cooper, Harriet Beecher Stowe, Herman Melville, Louisa May Alcott, Sir Walter Scott	Upton Sinclair, Hamlin Garland	William Faulkner, Virginia Woolf, DH Lawrence, Joseph Conrad, John Dos Passos, Sherwood Anderson, H.D. /Hilda Doolittle
Content and Style Markers	Satire, emphasis on the personal, character adventures	Revolt against scientific rationalism, emphasis on aesthetics, embrace of the exotic/pre-industrial	Belief that environment and heredity shaped character, influenced by Darwin, recreation of believable human experience. Topics include sex, filth, poverty, other harsh aspects of life	Stream of consciousness, multiple narrative points of view, breakdown of norms, preservation of individual identity

III. Datasets

From Project Gutenberg, we extracted 282 English texts written by selected authors. We created the following datasets to test our classifier on various different aspects.

1. All texts are in both the training set and testing set; The result from this dataset is the best possible we could produce from all datasets. We use this the check for overfitting.
2. (*Randomized data set*) No overlapping texts between training set and test set. All texts in the test set have at least one text in the training set that is written by the same author. The dataset is generated by randomly distributing each text into either the training set or test set will mostly likely fall into this category.
3. No overlapping texts and no overlapping authors between training set and test set. Since literary period classification is not authorship attribution, it has to be able to generalize style among authors in the same period. An author classifier will do fine in dataset 2, therefore we use this dataset to test that our classifier is not merely an authorship classifier.
4. No overlapping texts and no overlapping authors between training set and test set. Numbers of texts in the training set for each class are equal. We have very few texts from the Modern period, so we deleted training data from other classes to prevent overtraining them. (not include all 282 texts)
5. (*Most general data set*) No overlapping texts. One or two authors in the test set are also in the train set and the same number of texts in test set are written with some authors in the train set. This is our base dataset, which is the most general.

IV. Features

Our largest group of features was based on a limited unigram model. We used the words with the 150 largest counts as features and also used unigram counts of the top 20 punctuation marks occurring in the text.

We then included a range of shallow stylistic features, including:

- Vocabulary size
- Number of sentences
- Number of unigrams
- Ratio of words to punctuation
- Ratio of non-punctuation characters to punctuation chars
- Average word length
- Average sentence length
- Number of unique words
- Number of words found in dialogue passages
- Average length of dialogue passages

All of the measures that are listed as pure counts were scaled linearly with the document length to account for the significant differences in length between documents.

Note that we use word unigrams instead of bigrams because we believe that bigrams will be too sparse and too content specific. Although literature movement is content specific to some degree, bigrams will rely on the content too much and would end up selecting for books similar in content (e.g. similar settings) over books with similar styles.

Finally, we included part-of-speech tag trigram counts as a feature. We used all of the trigrams as features, however we used a limited tag set including only the 8 traditional parts of speech (verb, adjective, noun, adverb, conjunction, interjection, proposition) as well as an unknown tag and a sentence beginning tag. Part-of-speech trigrams were computed and saved for all test and training files prior to running the classifier.

For all of our features, rather than using real-valued features, we chose to binarize the features. Based on the range of each particular feature, we divided each feature up linearly into 20 buckets and used the bucket to which the real value was mapped to the binary feature.

V. Results

Dataset 2 (Randomized)

	Precision	Recall	F1

Augustan	0.863	0.905	0.864
Romantic	0.922	0.94	0.931
Naturalist	0.889	0.923	0.906
Modernist	1.0	0.75	0.875

These represent our best overall performance on any of the data sets, with the combination of features listed above. However, we believe that this is not the most realistic representation, because ideally one would want to be able to identify authors who were not included in the training set to determine how closely they represent the literary movement. As such, we also show our results on a more general training/test set combination, in which some of the authors appear in both, but several only appear in the test data.

Dataset 5 (General)

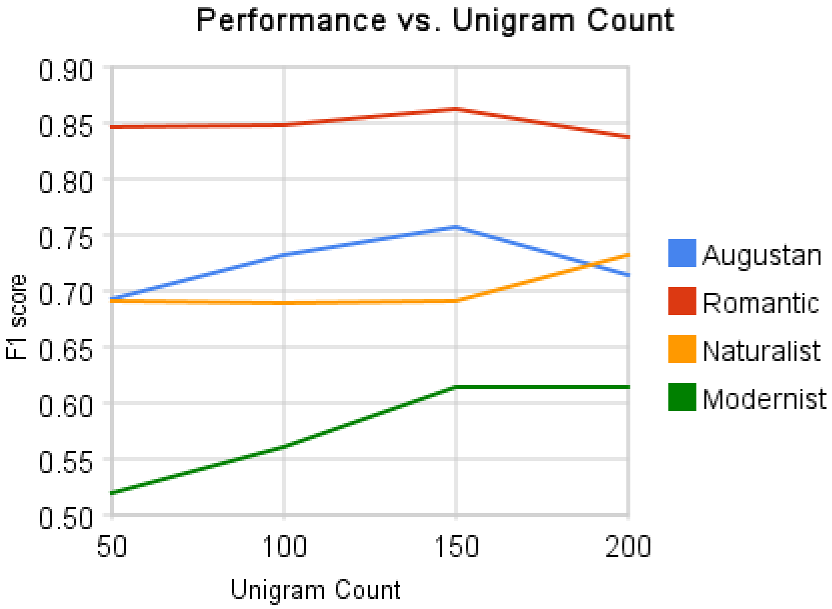
	Precision	Recall	F1
Augustan	1.0	0.706	0.828
Romantic	0.846	0.957	0.898
Naturalist	0.727	0.96	0.828
Modernist	1.0	0.5	0.667

As expected, the performance suffers somewhat from not having seen some of the authors before. However, the statistics are still reasonably high, showing that our classifier is identifying some unifying characteristics of the authors for given periods.

VI. Analysis

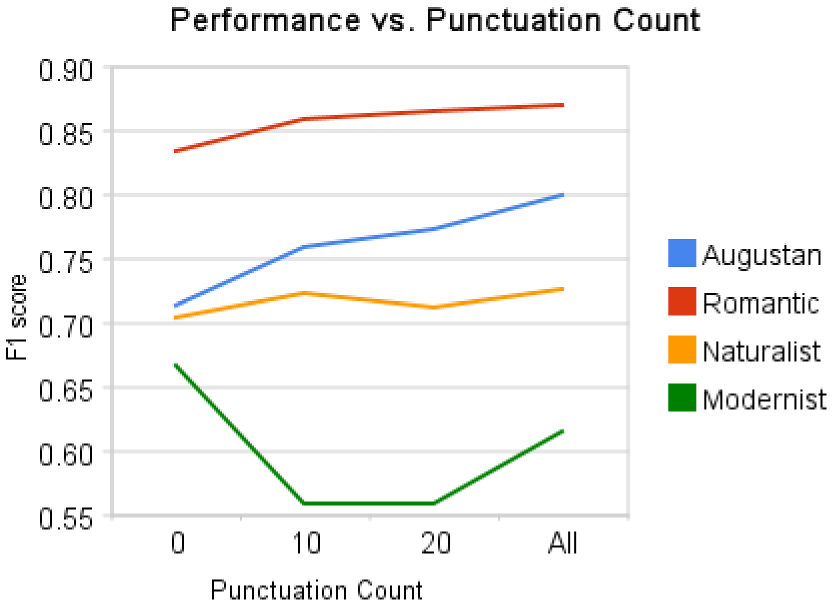
Feature Selection

To determine the number of unigrams to include, we tested on the general dataset (dataset 5) with 15 punctuation unigrams, 20 bins, no POS tagging, and no other features.

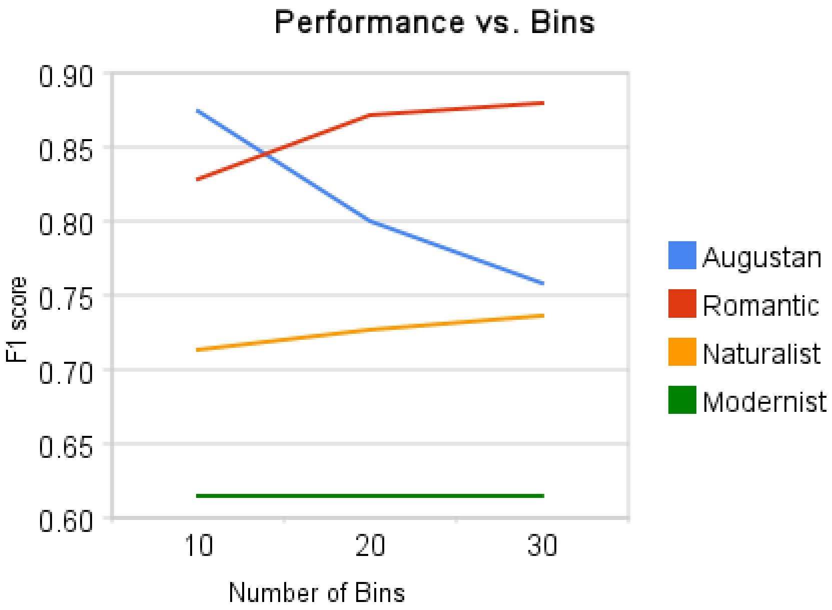


From these tests, we conclude that in the beginning, the more word unigrams, the better the performance. However, too many word unigrams not only reduces efficiency (by increasing the number of features) but also begins to overfit the data and consequently also worsens performance. Therefore, we chose to use only 150 most frequent word unigrams as features.

To determine the number of punctuation unigrams to use, we ran a series of tests with 100 word unigrams, 20 bins, and no other features, testing on the general dataset (dataset 5).



Because we binarize the features, we needed to decide how many bins to use for each real-valued feature range. To do this, we tested on the general dataset (dataset 5), with 100 word unigrams, 15 punctuation unigrams, and no other features.



Raising the number of bins to a certain point helps, but after that it improves some classes at the expense of others, so we decided to go with 20 bins.

For the above features, choosing the correct number means finding a balance between getting as much useful information about the test set as possible, while not overfitting it. This why we can not just increase all the measures indefinitely and expect linear improvements, but also why including a certain number of them improves the scores in the first place. The extreme case for word unigrams, for example, can be described by picturing two classifiers, one that counts zero unigrams, the other that counts all of them. The first would classify all documents equally because we have no information about how similar the documents actually are, while the second would only score highly documents that are very similar in word content to the initial documents, which is also not useful because it selects for a good deal of subject matter as well as style.

We also had to determine which set of POS tags to use for POS n-grams. The part of speech tagger provides the full Penn Treebank tagset, but some of the literature suggested using only 8 part of speech tags, so we created a tagset that includes only the following the 8 traditional POS tags (noun, verb, adjective, adverb, conjunction, preposition, interjection, pronoun) along with an "other" tag (for cardinal numbers, articles, etc) and a sentence beginning tag.

Test 1: Using Penn Treebank POS tags

	Precision	Recall	F1

Augustan	1.0	0.529	0.692
Romantic	0.789	0.978	0.874
Naturalist	0.676	0.92	0.779
Modern	1.0	0.333	0.5

Test 2: Using reduced POS tag
See table for dataset 5 in results section.

Test 3: Using both tagsets

	Precision	Recall	F1
Augustan	1.0	0.647	0.786
Romantic	0.833	0.978	0.9
Naturalist	0.686	0.96	0.799
Modern	1.0	0.333	0.5

We decided, based on these results, to only use the reduced part of speech tags. The likely reason why the full tagset does not work as well as the reduced tagset comes back to the problem of sparsity of data and overfitting. With the full tagset, there are a very large number of trigrams, a large fraction of which are only seen once and many which are not seen at all. When we reduced the number of tags, it ended up combining similar n-grams into one, making the counts for that more significant compared to n-grams which were truly rare or non-existent.

We also tried limiting the number of POS n-gram features as we did with word unigrams, however, we found that using all of the n-grams gave a better result, probably because use of rare and moderately rare grammatical constructions is a better marker of style than use of rare words.

Some other small considerations to be made were the size of the n-grams for part of speech tagging, for which trigrams were the best, combining the most information with only acceptable amounts of overfitting. We also chose a range of smaller features, which when included improved the scores by a small amount.

For the final test, we tested on all the features and chose different limits on number of word unigrams and punctuation unigrams we used. We found that feature combination of 150 word and all punctuation unigrams does not actually produce the best result, possibly due to overfitting, so we reduced the number of features and found that the combination of 100 word unigrams and 15 punctuation unigrams does a better job when combined with POS and other style features. This is the final feature set we used.

Dataset Testing

Dataset 3

	Precision	Recall	F1
Augustan	1.0	0.273	0.429
Romantic	0.636	0.933	0.757
Naturalist	0.571	1.0	0.727
Modernist	No modernist texts were reported	-	-

In this dataset, there was no author overlap between the training and the test sets. The main purpose of this dataset was to see if our classifier works beyond authorship attribution - we wanted to see if it could accurately detect literary period even if it had never seen the authors before.

Clearly, these results are not as impressive as our other ones. Our classifier, while still performing fairly well on Naturalist and Romantic texts, performs poorly on Augustan and cannot even correctly identify a single modernist text. This drastic change indicates that it is likely that we still have a substantial dependence on author. In some ways, this is unavoidable because we are attempting to determine literary style, which is author-dependent. We also believe we have an overfitting problem, which is discussed below.

However, we are still performing well on two datasets, but we believed that may have been from unequal amounts of training data: we have much more Romantic data than we do Modernist, and we thought that might have been skewing the results slightly. So we created another dataset in which no authors overlapped with the additional constraint that all sets had the same number of training documents.

Dataset 4

	Precision	Recall	F1
Augustan	0.8	0.690	0.741
Romantic	0.838	0.792	0.814
Naturalist	0.562	0.72	0.632
Modernist	No modernist texts were reported	-	-

This dataset shows worse performance in the two categories which had previously performed well and dramatically increased performance in one which had not performed well. But because Romantic and Naturalist classification remained fairly high, we conclude that it was not solely because of the amount of training data that these classes performed well. Additionally, the improvement in Augustan classification leads us to believe that that particular data set was being overfit, which is why we performed a test in which all of the data was used in both training and testing.

Dataset 1

	Precision	Recall	F1
Augustan	1.0	1.0	1.0
Romantic	1.0	1.0	1.0
Naturalist	1.0	1.0	1.0
Modernist	1.0	1.0	1.0

We believe this, along with previous results, indicates that we do have an overfitting problem with our classifier. The classifier still works to some extent, but our tests seem to indicate that it is fitting very closely to particular authors. For example, in one test run the only texts classified as modernist were those written by Joseph Conrad.

Even though this classifier does end up fitting fairly closely to particular authors, we can still however make some analysis of the various literary periods. For example, because even when training on completely different authors, Romanticism and Naturalism have fairly high F-scores, the authors in this period probably had a much more cohesive writing style as a movement than those in the Augustan or Modern periods. Historically, this makes sense: the Augustan period was the beginning of the novel as major genre of writing, so the art form had yet to fully mature and styles and conventions had yet to fully coalesce. At the other end, Modernism is the beginning of a literary rebellion, leading into Post-Modernism, which rejects conventions altogether. Indeed, though they are both pinnacles of Modernism, Ernest Hemingway and F. Scott Fitzgerald have drastically different writing styles. While the other movements were unified in style as well as theme, it may simply be difficult to classify Modernist literature stylistically because the Modernist movement was centered primarily around a set of ideas and not around a literary style.

Efficiency

There were some features that we experimented with, but did not end up including because of efficiency reasons. One feature suggested in the literature was CFG rule production counts, but this required parsing more than a million sentences. Even when restricting the size of the sentences, this proved to be too time-intensive, so we dropped this for the much faster part-of-speech tagging. We also experimented with the use of some regular expressions for parsing dialogue, but once again this proved very costly. We then decided to go with a more simplistic, but faster approach which just scanned the texts for quotation marks and counted the number of words in-between each pair.

VII. Related Work

This particular classification task falls into the category of "stylometry," or classifying a text based on stylistic features rather than content based features. Related work includes genre detection (though this often used for web-based documents which include stylistic features not present in novels, such as images) and, more closely, authorship attribution. Authorship attribution by stylistic features is not a new idea, and many computational feature sets have already been explored, including vocabulary features such as "vocabulary richness," vocabulary overlap, or use of synonyms; syntactic properties like part of speech n-grams; and even metadata features like web page layout or the number of spaces following a period (Juola 2008). Other shallow linguistic features as described by Stamatatos et. al. (2001) include pure unigram counts, sentence length, characters per word, and punctuation counts and included using chunking to find phrases and sentence boundaries and take average numbers and sizes of chunks. However, as many of these are highly dependent on the length of the text, some research has been done on using functions of count (e.g. unigram counts) for better results (Kessler et. al. 1997). Lately there has also been investigation into "deep" linguistic features of the text, including the use of context-free grammar productions (Gamon 2004).

VIII. Further Work

As there has not been extensive literature on this particular subject, there are several areas of potential interest where further work can be done. Applications of this sort of classifier run from the more immediately familiar, say by electronically organizing scanned texts for a digital library, to more interesting ones, like investigating how much of what we as humans call literary style can be captured in more superficial features that don't require extensive knowledge of the world around us.

One area worth following up on is whether or not our classifier is truly a genre classifier. That is, do the features we have here actually capture the stylistic and content markers listed in Table 1 for each literary genre? Or are these features instead tied more closely to the author or time period in which the text was written? A few preliminary experiments we ran suggested that these features captured information more about time periods (e.g. Mark Twain's *Huck Finn* was classified as a Naturalist text, because it is from the same era), but more work needs to be done. Another area of work concerns poetry, as opposed to prose. While some of the texts were poems, they comprised a very small portion of the corpora. Building a classifier for poetry would be far more challenging, given that it is a much shorter text form.

IX. Collaboration

Yiam wrote much of the structural code for training and testing the classifier, but feature selection and implementation, parsing, testing, and writing of the report were done collaboratively in group project sessions.

X. Citations

Gamon, Michael. 2004. "Linguistic correlates of style: authorship classification with deep linguistic analysis features." *International Conference on Computational Linguistics.*

Juola, Patrick. 2008. *Authorship Attribution.* Now Publications, Inc.

Kessler, Brett, Geoffrey Nunberg, and Hinrich Schutze. 1997. "Automatic Detection of Text Genre." Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics, 32-38.

Stamatatos, Efstathios, Nikos Fakotakis, and George Kokkinakis. 2001. "Automatic Text Categorization in Terms of Genre and Author." *Computational Lingusistics*, Vol. 26, No. 4.