Ranking contemporary American poems

Michael Dalvean
Australian National University

Correspondence:
Michael Dalvean,
School of Politics and International Relations, Australian National University, Canberra ACT 0200, Australia
Email:
michael.dalvean@anu.edu.au

Abstract

The purpose of this article is to examine the linguistic differences between poems written by 'amateurs' and those written by 'professionals' and then to use these characteristics to rank a number of contemporary American poems. The corpus of poems used consist of 100 poems randomly selected from a recent anthology of professional poets and a control group of 100 poems written by amateurs. The poems were reduced to ninety-eight linguistic and psycholinguistic variables, and these were used in a machine learning algorithm to build an ensemble classifier. The accuracy of the classifier was 84.5%. The probability scores of the individual poems was then used to rank the professional poems on a continuum representing amateur at one extreme and professional at the other, thereby providing an objective means of ranking contemporary poems.

1 Introduction

The purpose of this article is to examine what distinguishes a 'professional' poem from an 'amateur' poem. The central idea here is that professional poets are more likely than amateur poets to have grasped the basic skills associated with writing poetry and have therefore been able to produce poems of lasting quality. Amateurs, on the other hand, are less likely to have mastered the basic required skills and are therefore less likely to have produced work of lasting quality. Intuitively, we know that there are differences between the skills of amateurs and professionals in various fields, and we are quick to make aesthetic judgments based on our raw subjective responses. However, the objective quantification of the factors that lead to such responses is rarely considered. By using computational linguistics, it is possible to objectively identify the characteristics of professional poems and amateur poems. This way an objective basis for our subjective responses can be identified.

The upshot of identifying the characteristics of high-quality poems is that we can then come up with a means of placing poems on a continuum according to how much a poem exemplifies the characteristics of an amateur poem or, at the other extreme, a professional poem. We can then use this continuum to rank professional poems and, in doing so, we can make some objective statements about which poems are 'better'. There is a tradition of considering some poets as 'minor' and others as 'major' (Eliot, 1946). Placing poems on a continuum that is based on the extent to which poems possess the craftsmanship of a professional may be a step towards explaining why some poets are 'greater' than others. However, it should be stated that this article specifically examines the differences between amateur and professional poems rather than the examination of what makes a 'minor' or 'major' poet or what constitutes 'greatness' in a poem. Thus, an important element of this article is the creation of a continuum using a corpus of contemporary American poets and contemporary 'amateur' poems.

doi:10.1093/llc/fqt036 Advance Access published on 28 June 2013

1.1 Related work in computational linguistics

Several computational linguistic approaches to the analysis of poetry have been made. Rhyme and meter have been quantified (Green *et al.*, 2010), and methods to classify poems according to individual authors and styles have been used (Kaplan and Blei, 2007). However, only two attempts have been made to isolate the variables associated with poetic talent. The first study to use computational linguistics to identify high quality poetry is Forsyth (Forsyth, 2000), which looked at the characteristics of English poems over the last 400 years. The analysis here was based on a study group of poems that consistently appeared in recent anthologies. A control group was selected by matching each 'successful', frequently anthologized, poem with an 'obscure', seldom-anthologized, poem written by an author born within 10 years of the 'successful' poem's author. This gave rise to a sample consisting of eighty-five 'successful' and eighty-five 'obscure' poems approximately matched by date of composition. The study found that the successful poems had fewer syllables per word in their first lines and were more likely to have an initial line consisting of monosyllables. It was also found that successful poems had a lower number of letters per word, used more common words, and had simpler syntax. Thus, contrary to what we might expect, the more successful poems used simpler language. In essence, poems that use language that is simple and direct are more likely to be reproduced in anthologies. The second study is that of Kao and Jurafsky (2012). This study used a study group of 100 'successful' American poems, where success was defined as having been reproduced in the anthology *Contemporary American Poetry* (Poulin and Waters, 2006). They used a control group of 100 amateur poems selected from an amateur poetry website (www.amateurwriting.com). In terms of effect size and statistical significance, the biggest difference was that the professional poets used words that were more concrete than the amateur poets. Furthermore, the amateur poets were more likely to use perfect rhymes rather than approximate rhymes, more alliteration and more emotional words, both negative and positive. Finally, professional poets tend to use a greater variety of words than amateur poets. That is, the number of different words in the 100 professional poets is greater than the number of different words in the amateur corpus. This is not to say that they use more complex words, merely that they use a greater variety of simple words.

1.2 An alternative approach

In this article, I attempt to extend the kind of analysis undertaken in Forsyth (2000) and Kao and Jurafsky (2012). That is, I wish to determine what distinguishes a well-crafted poem from a less well-crafted poem. I use the same data as that used by Kao and Jurafsky (2012). However I extend the analysis in two ways. Firstly, I examine a broader range of linguistic variables than Kao and Jurafsky. The significant insight from Kao and Jurafsy's (2012) analysis is that the concreteness of words is far more important an indicator of poetic quality than any of the characteristics we might usually associated with poetic craft such as perfect end rhyme frequency or the type/token ratio. Therefore, if a search is made for linguistic characteristics using the types of variables that have been investigated in relation to language processing then there is the possibility that the insights gained by Kao and Jurafsky (2012) can be further extended. For this purpose, I use sixty-eight linguistic variables derived from Linguistic Inquiry and Word Count (Pennebaker *et al.*, 2001) and thirty-two psycholinguistic variables from the Paivio *et al.* (1968) word norms. It will become apparent that this approach provides a further insight into the types of linguistic characteristics that distinguish professional from amateur poems.

A second way in which I extend the analysis of Kao and Jurafsky (2012) is to use machine learning to develop a classifier. The idea here is that if there are characteristics that distinguish amateur from professional poems then it should be possible to classify a given poem as being more towards the amateur end of the spectrum or more towards the professional end. This being the case, it is also possible to *rank* individual poems according to their position on the spectrum. Thus, given Kao and Jurafsky's (2012) selection of 100 professional poems, it should be possible to rank them according to where they are on the spectrum. In this sense, it is possible to state that, even among professional poets, some are better than others.

2 Method

2.1 The data

The data consist of the 200 poems used by Kao and Jurafsky (2012).[1] Of these 200 poems, 100 are professional poems drawn from *Contemporary American Poetry* (Poulin and Waters, 2006) and 100 are amateur poems drawn from www.amateurwriting.com. The professional poems were written in the later half 20th century by poets who have been members of the Academy of American Poets. In the 100 poem corpus there are sixty-eight individual poets.[2] The number of poems chosen from the anthology was in direct proportion to the number of poems the poet had in the anthology. Where a poem was >500 words, it was removed and replaced by another poem by the same poet. The final selection of 100 poems had an average of 175 words (min = 33; max = 371) (Kao and Jurafsky, 2012, p. 4).

The 100 control poems were selected from www.amateurwriting.com, which is a free website on which anyone is able to post their writing. Of the 2,500 available at the time of selection, 100 were randomly selected and corrected for grammar and spelling. The average length of poems was 136 words (min = 21; max = 348) (Kao and Jurafsky, 2012, p. 4). There is no reliable information as to the authorship of the amateur poems in Kao and Jurafsky's article as the poems are submitted anonymously.[3] Thus, the relative proportions of poems from individual authors may not mirror those of the study sample. That is, there may be more than one poem from one or more authors, or there may be 100 from 100 different authors. The question as to whether this would affect the results can be addressed by considering two observations. The first is that the unit of analysis is the individual poem rather than poets. There is certainly going to be a correlation between the writing styles used by one author over several poems. However, the results below show that there can be a wide variation in the styles used by the same poet: there are professional poets in the sample that have poems that are classified in both the amateur and the professional range. The important point here is that we would expect there to be some variation in the quality of the amateur poems so that even if there were significantly fewer poets in the amateur sample or the proportions were very different from those of the professional group, there would be sufficient variation to provide a viable control group. The second observation is that the out of sample classification accuracy is very good for both the professional and amateur poets. If there were a problem with the control, we might find that the control group classifies well but the study group does not. We will see that there is a high sensitivity and specificity which indicates that the generalization ability of the models is high. This is unlikely to occur if there is a fundamental problem with the control group.

2.2 The variables

The dependent variable in the analysis is a binary taking the value of 1 if the poem is by a professional poet and 0 if it is not. The independent variables are linguistic variables derived from two sources—Linguistic Inquiry and Word Count (LIWC) and the Paivio, Yuille, and Madigan (1968) word norms and their extension by Clarke and Paivio (2004).

Sixty-eight linguistic variables were derived from LIWC. This program breaks text down into sixty-eight linguistic categories according to a specifically designed dictionary (Pennebaker *et al.*, 2001). The categories used are based on common behavioural and cognitive processes and include Negative Emotion, Affect, Leisure, Work, Family, Social Activities, and Psychological Processes. The categories were derived from lists of words empirically associated with each category. Thus, the Psychological Processes category was derived from words developed from the Positive Affect Negative Affect Scale (Watson *et al.*, 1988, cited in Pennebaker *et al.*, 2007), Roget's Thesaurus, and standard English dictionaries. Thus, with sixty-eight linguistic categories, LIWC captures a great deal of the linguistic content of a given text. Of the sixty-eight variables, two were excluded: Word

Count (number of words) and Words per Sentence. The word count was dropped because the intention was to isolate the linguistic characteristics of the words used rather than the quantity of words. The Words per Sentence variable was dropped because there are idiosyncratic uses of punctuation and line length in both groups of poems, which affect the raw count of words per 'sentence'.

An additional thirty-two psycholinguistic variables were derived from Paivio *et al.* (1968) word norms and the extension of these by Clarke and Paivio (2004). The Paivio *et al.* (1968) and Clarke and Paivio (2004) (PYMCP) word norms are derived from a sample of 925 nouns. For each word, thirty-two linguistic and psycholinguistic variables were derived. Some of these are structural such as the number of letters and number of syllables. Another set of variables were derived from subjects' responses to the words by getting to answer questions on a number of psycholinguistic dimensions. The variable 'meaningfulness' was derived by asking subjects, for each word, how many associated words they could think of in 30 seconds while the variable 'age of acquisition' (AOA) was derived by asking subjects at what age they estimate they learnt each of the 925 words. The result is that there are thirty-two variables for each of the 925 words that measure their structural and psycholinguistic properties. To illustrate how the poems were scored on each of these thirty-two variables I shall use the 'ease of definition' (Def) variable. This variable was derived by asking how easy it was to define each of the 925 words on a scale of 1 (very hard) to 7 (very easy). Thus, for each of the 925 words we have a Def score. Out of the 925 word sample the word that was easiest to define was 'baby' (score = 6.79), and the word that was the hardest to define was 'gadfly' (score = 1.92). The average score for the 925 words was 5.14. Words with in this range were 'vessel' (5.13), 'warmth' (5.13), 'alimony' (5.17), and 'caravan' (5.17).

To use the raw Def scores to score poems, the first stage was to determine, for each poem, which of the 925 words in the PYMCP sample were present. The average Def score for each poem could then be calculated. Consider for example the sentence

'The baby ridiculed the gadfly's caravan',

In this sentence the words 'the' and 'ridiculed' are not in the 925 word sample so they are not part of the calculation. The remaining words, 'baby', 'gadfly', and 'caravan', are in the sample and have scores of 6.79, 1.92, and 5.17, respectively. The sentence contains three words from the sample so the 'Def' score for the sentence is calculated as follows:

$$(6.79 + 1.92 + 5.17)/3 = 4.6.$$

Using this methodology, we get a proxy for the average Def (ease of definitions) of words used in each poem. It is only a proxy because it is based on a 925 word sample. The poems were scored on all thirty-two psycholinguistic variables in the same way as described above for Def.

It should be stated that, as there are 925 words in the sample, it is possible for a given text to not contain any of the words. However, this is unlikely as the selection is very broad, covering a wide variety of common and obscure words. There was no case across the sample of 200 poems in which at least one of the words in the 925 word sample did not occur.

Thus, the data consist of a corpus of 200 poems with the 100 professional poems scored as 1 and the amateur poems scored as 0. For each of these poems, there are sixty-six linguistic variables derived from LIWC and thirty-two derived from the PYMCP norms.

2.3 Machine learning

It is apparent that the number of variables under consideration is half the sample size. In traditional hypothesis testing this would be a problem. However, recent advances in machine learning have pointed the way towards making sense of situations in which there is a great number of independent variables. Much of this approach has been developed in the context of gene sequencing in which it is not unusual to have a sample size of <200 and yet the number of independent variables that need to be considered is several thousand. Ultsch and Kämpf (2004) give an example of a data set consisting of 72 leukemia patients and 7192 variables. Clearly there needs to be some way of selecting the variables that are likely to provide the best signal. The solution used in this article is to use logistic regression with forward stepwise selection. Under this procedure, variables are selected according to an algorithm that surveys all the independent variables and selects the independent variable that provides the best logistic fit for the dependent variable. This procedure continues until no additional variables can be found that add to the model's ability to fit the data. Clearly, this can lead to problems because it is possible that variables are selected due to their ability to learn the 'noise' in the dataset rather than generalize. This is known as 'overfitting' (Hawkins, 2004). To prevent overfitting, an independent holdout sample can be used to check the generalization ability of the model at each of the steps in the stepwise procedure. The idea here is that several testing samples will be 'held-out' from the model building procedure and will only be used to test the generalization ability of the model at each stage of its development. Typically, the generalization ability of a model rises with the first few independent variables added and then falls away as more independent variables are added. As independent variables are added, the *internal* measures of model fit such as R^2 tend to rise consistently but the *external* generalization ability (that is, the ability to classify cases that were not used in the creation of the model—the 'held-out' cases) falls considerably after the first few variables are selected. The idea is to choose the model that maximizes the external generalization ability.

It is important to specify the holdout sample correctly, as it must at all times be separate from the sample of the data used to create the model. The idea here is that a certain proportion of the data p should be used to create the model and the remaining proportion $1-p$ should be used to test that the model has not been overfitted. If the model is able to generalize then it should be able to correctly classify cases that were not used in creating it. This 'hold-out' sample is one way of doing this and is a standard method of testing models in machine learning.

Another technique derived from machine learning is the use of an ensemble of models to increase the classification accuracy. The idea here is that averaging the outputs of several different models will likely increase the overall accuracy. This assumes that the errors of each constituent model in the ensemble are not correlated. One way to do this is to train different models on different subsets of the data. Another way is to use different variables in each constituent model. In this article, both approaches are used.

Before discussing the modelling process in detail, it is worthwhile to consider a question that arises in relation to the studies that have been done with this data previously: Why not simply use the logistic equation from Kao and Jurafsky's (2012) analysis? The answer is that there is a problem with overfitting in any modelling and, although it is possible that their equation is not overfitted, in the absence of an independent test using a holdout sample or some similar method, it is always possible that the equation is overfitted to the data. In such cases, the model does not truly generalize but instead 'learns' the noise in the sample and is therefore not useful for actually classifying poems into professional and amateur. This is despite the fact that certain variables may have been identified as being important in such a classification scheme. There is a distinction between traditional hypothesis testing and machine learning. Traditional hypothesis testing is based on the idea that the identification of statistically significant variables is the essential aim, as it is required to develop theoretical explanations. The problem with such an approach is that it can lead to the identification of variables that have statistical significance but little discriminant power. The central aim of machine learning, on the other hand, is classification, so the variables selected must be strongly associated with the dependent variable to the extent that the variables can be used to discriminate between the two classes. The statistical significance of variables is not as important as whether they are able to increase the classification accuracy of the model.

3 Modelling and Results

The modelling was undertaken twice: once with the LIWC variables and then again with the PYMCP variables. The reason for this is that ensembles of models, in which the output of several models is averaged, work best when they use different variables. The reason for this is that the efficiency of ensembles increases where the errors of the constituent models are less correlated. Creating models with different sets of variables is one way to reduce the extent to which the errors are correlated.

The first stage of the modelling procedure is to divide the sample ($n = 200$) into ten ‘folds’. This involves randomly dividing the sample into ten folds of $n = 20$. The procedure is to build a model using nine of the folds and then testing on the ‘held-out’ fold. The training sample is used to create models using the stepwise procedure while the testing sample is ‘held-out’ from the model-building procedure and used only to test each model created at each step of the stepwise procedure. At each step, a the forward selection process selects the variable that most efficiently increases the model fit in terms of R^2 until there are no more variables that can do so. The held-out fold is then used to test the external generalization ability of the model. The best model is that which best classifies the held-out fold. The above procedure is repeated ten times using each of the held-out folds once as the validation set. Because there are different subsets of the data, there will be some variations in the variables selected. Table 1 shows the variables, their coefficients, and the accuracy statistics for all ten folds for the modelling using the LIWC variables.

Two of the variables, affect (words indicating affect such as ‘like’and ‘tense’and ‘grieve’) and article (articles such as ‘the’ and ‘a’), occur in nine of the models and the variable present (use of the present tense) occurs in eight, indicating that these variables seem to be closely linked to the classification of poems into amateur and professional classes.

The average overall accuracy across the ten folds is 84% while the sensitivity and specificity are 85% and 83%, respectively. Importantly, the classification accuracy of each fold is >50%, with the lowest accuracy of 65% occurring with Fold 9 (Sens = 67%, Spec = 64%).

The procedure was repeated using the PYMCP variables. Furthermore, a different random selection of cases was used to create a second tenfold division of the data. The results are presented in Table 2.

The variable Con (concreteness) occurs in all models. The variable Rhy (rhyming similarity—a measure of the extent to which words are perceived to rhyme with other English words) occurs in four models while Len (word length in letters) occurs in three. Emogd (emotional goodness—a measure of the extent to which words express ‘good’ emotions)

Table 1 Variables, coefficients and accuracy metrics for LIWC modelling

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Fold 6	Fold 7	Fold 8	Fold 9	Fold 10
Variables										
Affect	−0.32	−0.41	−0.34	−0.36	−0.30	−0.32	-0.27	−0.37	−0.43	
Article	0.31	0.15	0.21	0.24	0.20	0.20	0.24	0.21		0.28
Filler		1.14								
Insight	−0.60	−0.68	−0.46					−0.53		
Number	0.58	0.99		0.47				0.49	0.66	
Past	0.18									
Ppron									−0.19	
Present		−0.18	−0.13	−0.20	−0.16	−0.16	−0.15		−0.19	−0.21
Sixltr	0.11									
Time	−0.28	−0.46	−0.23	−0.33	−0.25			−0.33	−0.32	
Work	0.83									
Constant	−0.79	4.92	3.26	2.97	2.66	1.22	0.67	3.01	7.27	−0.65
Accuracy										
Acc%	90	90	80	80	90	90	85	90	65	80
Sens%	80	100	80	73	100	83	92	100	67	78
Spec%	100	78	80	89	75	100	75	88	64	82

Table 2 Variables, coefficients and accuracy metrics for PYMCP modelling

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Fold 6	Fold 7	Fold 8	Fold 9	Fold 10
Variables										
CON	0.55	1.06	0.90	0.86	0.91	0.74	0.92	0.91	0.98	0.64
EMO	−0.87									
EMOGD						−1.06				−1.43
LEN							0.85	0.67	0.74	
RHY		−1.86		−1.67	−1.74	−1.55				−1.39
Constant	1.25	2.56	−4.20	2.82	2.80	4.11	−8.57	−7.49	−8.25	4.45
Accuracy										
Acc%	80	75	85	90	85	75	75	80	75	80
Sens%	100	80	80	100	90	77	92	88	67	82
Spec%	67	70	90	83	80	71	43	75	82	78

occurs in two models and Emo (emotional content) occurs in one.

The average accuracy across the tenfolds is 80%, with sensitivity of 86% and a specificity of 74%. All classification accuracies are >50%, with the lowest at 75%. The specificity of 43% for fold 7 indicates that this model is not as efficient as other models with the same overall accuracy of 75%, such as fold 2, 6, and 9. However, even with this low specificity, the overall confidence intervals for this model (95% CI: 50.9–91.3%) indicate that the classification accuracy is >50%. As such, including it in the ensemble can potentially increase the accuracy of the ensemble.

Across the twenty models, each case is represented twice: once in each of the holdout samples of the LIWC modelling and the PYMCP modelling. The ensemble score is derived by averaging the logistic score for each case across the two sets of models. If the average is above 0.5 the case is scored as a 1 while if the score is below .5 the case is scored as a 0. This procedure provides an overall accuracy of 84.5% (95% CI: 78.7–89.2%), sensitivity of 90% (95% CI: 82.3–95.1%), and specificity of 79% (95% CI: 69.7–86.5%). The kappa value is 0.69 (Test of Ho: Kappa = 0, z = 9.82, $P = 0.0000$, two tailed test). Thus, the ensemble provides an efficient classification system.

Before I leave the discussion of the modelling procedure, I should mention how the ensemble could be used to assess a poem that is not in the sample. That is, how would we classify an 'unseen' poem that was not part of the initial sample of 200? The answer is that we would reduce the poem to the LIWC and PYMCP variables and then run the data through all twenty models and then average the results. This method, known as a cross-validated committee (Chali *et al.*, 2009), has been found to be better than attempting to assess which is the best model out of a selection of potential models: '[c]ombining outputs of multiple classifiers into an ensemble (committee) output is one of the most important techniques for improving classification accuracy' (Verikas *et al.*, 2010, p. 69–57).

3.1 Ranking the poems

The upshot of the preceding section is that we have an algorithm that is able to correctly classify poems as professional/amateur with an accuracy of 84.5% using linguistic variables. There are several applications for such an algorithm. For example, a publisher who needs a quick way of sorting through the voluminous submissions received on a weekly basis could first select a filtered list by running poems though such an algorithm. This is not to suggest that this should be the only means by which a publisher should sort poems. The idea is that given the large number of submissions a large publisher receives, many of which may not even be read, a preliminary method of sorting the poems into those that have professional characteristics and those that do not would be one way of ensuring that a potential gem is not relegated to the depths of the 'slush pile'. The method would thereby provide a ranking which could provide the publisher with a starting point for the reading task. Depending on the resources of the publisher, a proportion of the poems could be read using the ranking as the means of determining which poems were selected.

However, I wish to discuss a different application—the ranking of contemporary established poems. There is a tradition of regarding poets as 'great' or 'minor'. We tend to ignore the fact that some poets are not great or minor but are simply forgotten, as Forsyth's (2000) study emphasizes. T.S. Eliot points out that there is a distinction between major and minor poets but that most people would disagree about which poets should be on which lists (Eliot, 1946). The point of ranking poems using a classification scheme such as the one advocated in this article is that such a method provides an objective measure of the likely subjective judgments of many individuals.

The procedure is to use the ensemble classifier to give each of the established poems a score which can then be used to place them on a continuum from most professional to least professional. The score is simply the score derived by the ensemble classifier. That is, the score is the average logit score derived from the two logit scores of the two holdout samples in which the individual poem occurs.

The amateur poets are excluded from this comparison for the simple reason that their status is not in contention. However, it should be noted that there is no reason that we could not provide a score for the purposes of identifying amateur poems who are producing work of a professional standard. In this regard it is worthwhile noting that in the control group of 100 amateur poets, there are twenty-one with logit scores in the 'professional' range of >.5. Of these 21, six score in the very high range of >.8, suggesting that these poems may be indicative of future poetic success.

Table 3 lists the poems and authors in descending order of logit scores. The highest score is .97 for the poem *The Image* by Robert Hass. The lowest score is .04 for *The Language* by Robert Creely.

As indicated by the sensitivity of 90%, the vast majority of the poems, 90 out of 100, have scores in the 'professional' range of >.5. In other words, ten of the professional poems score in the amateur range of <.5. That is, there are ten professional poems that are more like amateur poems than professional poems. An interesting observation can be made about the poets who have more than one poem in the corpus in that there is a great deal of consistency in the classifications of their poems. Of those poets who have more than one poem in the corpus, most show consistently high or low quality. For example, Robert Hass has two poems in the corpus, *The Image* and *Our Lady of the Snows*, which score in the high to very high range of.72 and .94, respectively. At the other extreme are Galway Kinnell and Robert Creely who also have two poems each in the corpus but whose poems both score in the amateur range of <.5. Finally, there are poets who have poems in each of the high and low scoring categories. Louise Gluck, for example, scores .27 for *Celestial Music* and .93 for *Nostos*. Similarly, John Berryman scores .77 for *Dream Song 172 Your face broods* and .43 for *Dream Song 26 The Glories of the World Struck Me.* In all, there are three poets who straddle the two categories. Given that there are thirty poets with more than one poem in the corpus, the majority (twenty-seven) have poems in one category or another. Thus, the three that straddle two categories represent the exceptions rather than the norm. Furthermore, where a single poet has more than one poem in the 'amateur' range, this is not merely a result of the 10% sensitivity error of the classifier but may indicate that the poems are in fact more like amateur poems than professional poems.

4 Discussion

The analysis indicates that an objective means of ranking poems is certainly possible. The question remains, however, as to whether the findings can be used to extend our understanding of poetics in general.

One way to isolate the important variables in the ensemble is to focus on variables that occurred in at least five of the ten models generated in each of the LIWC and PYMCP models. The general

Table 3 Professional poems ranked by logit scores

Poem	Author	Logit
The Image	Robert Hass	0.974
The Room of My Life	Anne Sexton	0.970
Wingfoot Lake	Rita Dove	0.948
Working Late	Louis Simpson	0.936
Lying in a Hammock at William Duffys Farm in Pine Island Minnesota	James Wright	0.933
The Prediction	Mark Strand	0.932
Nostos	Louise Gluck	0.929
The Choir	Olga Broumas	0.929
How Simile Works	Albert Goldbarth	0.923
Notice What This Poem Is Not Doing	William Stafford	0.922
Hello	Naomi Shihab Nye	0.913
My Indigo	LiYoung Lee	0.901
Gin	David St John	0.894
Facing It	Yusef Komunyakaa	0.890
Writing in the Afterlife	Billy Collins	0.889
Power	Adrienne Rich	0.888
More Blues and the Abstract Truth	CD Wright	0.888
Absences	Donald Justice	0.882
To Kill a Deer	Carol Frost	0.876
Variations On A Text by Vallejo	Donald Justice	0.873
The Small Vases from Hebron	Naomi Shihab Nye	0.870
Clear Night	Charles Wright	0.861
The Fish	Elizabeth Bishop	0.861
May 1968	Sharon Olds	0.856
Traveling through the Dark	William Stafford	0.853
Years End	Ellen Bryant Voigt	0.849
Letter	Jean Valentine	0.841
Crossing The Water	Sylvia Plath	0.839
Heaven as Anus	Maxine Kumin	0.834
The Porcelain Couple	Donald Hall	0.827
In Trackless Woods	Richard Wilbur	0.825
Oranges	Gary Soto	0.817
This Night	William Heyen	0.817
Dearest Reader	Michael Palmer	0.817
The Dancing	Gerald Stern	0.812
b o d y	James Merrill	0.808
Japan	Billy Collins	0.807
Tomatoes	Stephen Dobyns	0.806
The Stairway	Stephen Dunn	0.806
Cleaning a Fish	Dave Smith	0.800
Warning to the Reader	Robert Bly	0.798
New Vows	Louise Erdrich	0.795
Nurture	Maxine Kumin	0.794
Onions	William Matthews	0.792
Why I Am Not A Painter	Frank OHara	0.789
The Undressing	Carol Frost	0.788
GlassBottom Boat	Elizabeth Spires	0.788
The Older Child	Kimiko Hahn	0.785
Minor Miracle	Marilyn Nelson	0.778
Eating Alone	LiYoung Lee	0.772

(continued)

Table 3 Continued

Poem	Author	Logit
They Feed They Lion	Philip Levine	0.772
Dream Song 172 Your face broods	John Berryman	0.766
The Summer Day	Mary Oliver	0.760
The Mutes	Denise Levertov	0.754
The Intruder	Carolyn Kizer	0.750
Pacemaker	WD Snodgrass	0.748
Her Kind	Anne Sexton	0.742
Our Lady of the Snows	Robert Hass	0.721
Twentyyear Marriage	Ai	0.716
Aubade Some Peaches After Storm	Carl Phillips	0.716
Charles on Fire	James Merrill	0.715
Fork	Charles Simic	0.692
The Russian	Robert Bly	0.687
Root Cellar	Theodore Roethke	0.684
Audacity of the Lower Gods	Yusef Komunyakaa	0.683
Animals Are Passing From Our Lives	Philip Levine	0.683
at the cemetery walnut grove plantation south carolina 1989	Lucille Clifton	0.673
Hay for the Horses	Gary Synder	0.669
When You Go Away	WS Merwin	0.668
The Abduction	Stanley Kunitz	0.654
The Singing	C K Williams	0.654
The Strange People	Louise Erdrich	0.645
University Hospital Boston	Mary Oliver	0.645
To Speak of Woe That Is in Marriage	Robert Lowell	0.639
To an Adolescent Weeping Willow	Marvin Bell	0.617
Those Winter Sundays	Robert Hayden	0.610
The Night The Porch	Mark Strand	0.606
Approximately Forever	CD Wright	0.598
My Noiseless Entourage	Charles Simic	0.598
Sexual Jealousy	Carol Frost	0.590
To Dorothy	Marvin Bell	0.589
Reuben Reuben	Michael S Harper	0.586
Degrees Of Gray In Philipsburg	Richard Hugo	0.553
For the Anniversary of My Death	WS Merwin	0.549
A Blessing	James Wright	0.547
Fragments	Stephen Dobyns	0.546
Thrall	Carolyn Kizer	0.539
scar	Lucille Clifton	0.517
Riot Act April 29 1992	Ai	0.505
Personal Poem	Frank O. Hara	0.500
After Making Love we Hear Footsteps	Galway Kinnell	0.476
Dream Song 26 The glories of the world struck me	John Berryman	0.433
Celestial Music	Louise Gluck	0.286
A Lovely Love	Gwendolyn Brooks	0.236
Adultery	James Dickey	0.224
WeddingRing	Denise Levertov	0.213
Blackberry Eating	Galway Kinnell	0.204
Playing Dead	Andrew Hudgins	0.089
The Warning	Robert Creeley	0.059
The Language	Robert Creeley	0.041

finding of this article is that professional poems tend to use more concrete language. They use demonstrative language, as indicated by the use of articles. The negative association with the PYMCP variable 'Rhy'—a proxy for the extent to which words elicit other words that rhyme with the stimulus word—indicates that professional poets use words that are somewhat unusual but not necessarily complex. Professional poems have fewer words denoting affect but more words denoting number. Professional poems also refer less to the present and to time in general than amateur poems.

These findings support some of the findings by Forsyth (2000) and Kao and Jurafsky (2012). Forsyth, for example, found that the language use was relatively simple (Forsyth, 2000, p. 54). In this study the relatively high level of the use of articles, numbers, and concrete words by professional poets indicates that the language has a basically demonstrative orientation. Articles and numbers tend to be used in directly demonstrative language, and concrete language is in many senses more fundamental than abstract language.

Kao and Jurafsky link concreteness not with language simplicity but with imagery. Concrete words are concrete because they enable us to generate a tangible sensory image associated with the word. Thus, the word 'baby' is easider to imagine than the word 'unreality'. The idea here is that concrete words are better at creating sensory-based images:

> '. . .[P]oems written by professional poets contain significantly more words that reference objects and significantly less words about abstract concepts and generalizations. This result suggests that professional poets follow the sacred rule of 'show, don't tell' and let images instead of words convey emotions, concepts, and experiences that stick to readers' minds' (Kao and Jurafsky, 2012, p. 15).

This idea has a long tradition in poetry. Keats, for example, knew how important sensory imagery was in his poetry. He points out that his poem *Lamia* has a 'sort of fire in it which must take hold of people in some way—give them either pleasant or unpleasant sensation. What they want is a sensation of some sort' (Keats, 1958, p. 189).

Thus, the finding in this study and Kao and Jurafsky's that concreteness is a marker of professionalism has some backing in literary circles.

The absence of negative affect in professional poems was noted by Kao and Jurafsky (Kao and Jurafsky, 2012, p. 15). They did not specifically also look for positive emotion. The present analysis finds that professional poets use less overall affect in general. The LIWC variable affect occurs in nine of the ten models created using LIWC and has a negative coefficient. The variable captures both positive and negative affect. Thus, we could say that the current analysis extends Kao and Jurafsky's analysis in that we now have evidence that it is affect overall rather than merely negative affect that is used less in professional poems.

The finding that affect is not a prominent component of professional poems may seem counter-intuitive. The idea that poetry and emotion are somehow connected is ingrained in the discussion of poetry. Wordsworth, for example, stated that [a]ll good poetry is the spontaneous overflow of powerful feelings. (Wordsworth, 2009, p. 22 [1802]). Writing over 100 years later, Collingwood (1938) argued that art is the expression of emotion in a particular medium. Thus, on this account, poetry is the expression of emotion using poetic language and form. A recent account of poetic theory from an existentialist viewpoint holds that '[p]oetry may be thought of as the emotional microchip, in that it may serve as a compact repository for emotionally charged experiences' (Furman, 2007, p. 1).

The problem with these accounts of the link between poetry and emotion is that they do not explain why professional poets would use fewer emotion terms. All of the above accounts are consistent with the idea that using a lot of emotionally charged language is compatible with the writing of poetry. From the three accounts, there is an intimation that the more emotion, the better, but this is not what the linguistic analysis reveals. What seems to take place is that the language used is concrete and objective and the circumstances depicted by the language are what evoke the emotion in the reader rather than the language itself. A literary theorist who comes close to this idea is Eliot who said that '[t]he only way of expressing emotion in the form of art is by finding an ''objective correlative''. . . a set of objects, a situation, a chain of events which shall be the formula for that particular emotion; such that, when the external facts, which must terminate in sensory experience, are given, the emotion is immediately evoked' (Eliot, 1998, p. 68). There is evidence that this occurs in professional poems, not only in that there is less affect in professional poems but more use of articles and number terms, both of which tend to be used in relatively objective depictions of events.

Before concluding this discussion of the findings in the context of literary theory, it is worthwhile considering whether there is any evidence that some of the ideological schools of literary criticism are supported. Given the broad expanse of ideological literary criticism and the vagueness with which concepts are defined in these schools, it is not possible to make more than a few general remarks on this issue in this article. Furthermore, the vagueness with which the concepts are defined means that it is difficult to determine how the central ideas of an ideology could be tested. For any attempt to operationalize an ideological theory in order to test it, there is an ideologue who can argue that the means of operationalizing a supposedly salient variable is for some reason not a valid way of capturing the phenomenon in question. Despite these problems, the opportunity exists for some general observations about what ideological interpretations the data allow us to make.

Marxist interpretations of literature are based on considering the material conditions in which they were formed (Eagleton, 1976). Thus, we might expect that those who had not achieved eminence (the amateur control group) would use work terms differently than the established poets as their material and cultural outlooks may be coloured by their experiences. In fact, there is no such difference. The LIWC variable work (use of words such as 'wage', 'tax', and 'hour') was not significant in the analysis. The Marxist acolyte could respond by saying that the amateur poets have already been subverted by the dominant paradigm and have therefore adopted the mode of expression of the dominant class. If this is the case then testing Marxist literary theory with any dataset may be difficult, as it would be difficult to get controls who had not been subverted in a similar way.

Feminist literary criticism, which, according to one of its manifestations, holds that there is a uniquely male perspective and that this is what needs to be emulated by men and women if they wish to succeed in an androcentric world (Paul, 2012). This androcentrism is not supported. There is a specific gender variable in the PYMCP variables. This variable measures the extent to which words are 'gendered' on a scale of 1 for masculine and 7 for feminine with gender neutral at 4. If there were some systematic difference between the amateur and professional poems in terms of the gender ladenness of the words they use then this variable would pick this up. No such difference was found. Feminist critics could maintain that this may indicate that even the amateur poets have already assimilated the androcentric way of looking at the world. Whether or not this is the case we can still maintain that the difference between professional and amateur poets has nothing to do with the the use of gendered language. One other observation that should be made in relation to gender is that the average score for poems by women is .74 ($n = 35$) while the average score for poems by men is .69. This difference is not significant ($P = .24$, two tailed test).

Several of the various postmodern ideologies share an idea that all texts are equal in terms of merit because there is no 'correct' meaning of a text (Grenz, 1996, p. 110). The upshot of this is that the postmodernist holds that there is no difference in terms of literary merit between the cannon of great literature and less lauded text such as advertising copy. This idea had an ironic manifestation in 1996 when Alan Sokal submitted a hoax article to a then leading postmodernist journal *Social Text* and had the article accepted (Sokal, 1996). It seems that the editors, true to their ideological position, did not distinguish an article from a hoaxer from an article from an acolyte. There is little evidence of such confusion in the analysis of the difference between professional and amateur poems. As demonstrated, there is an objective way of classifying these two categories with an accuracy of 84.5%. It should be noted that despite this evidence, postmodernists might say that such an analysis is based on the grand narrative of machine learning and to ‘privilege’ the machine learning grand narrative over theirs is a mistake.

4.1 Future work

The results of this analysis, as with any similar analysis, are dependent on the selection of the cases and the controls. As such, future work in this field would benefit from extending the selection of the cases and controls. The use of the Poulin and Waters (2006) anthology as a source for the professional poems may be a limiting factor in the analysis, and there may have been some systematic bias in the selection of the poems in the anthology. There are potential gains from using a method of case selection similar to that of Forsyth (2000) which used several recent anthologies. Using such a method would reduce any systematic bias from one particular anthology and thus result in a more representative sample. Similarly, the selection of the control sample from only one website of amateur poems introduces the possibility of some selection bias, which could be mitigated by using several such sources.

However, an interesting extension of the current analysis would be to pool the cases from both the current study ($n = 200$) and the Forsyth (2000) study ($n = 170$). This would provide a cases and controls selected using different criteria. Furthermore, the poets are selected over different times and two different countries—the US and the UK. The selection biases of each of the individual studies would be mitigated by such pooling.

5 Conclusion

In this article, I have extended the work of Kao and Jurafsky (2012) in three ways: First, I have examined a greater number of linguistic variables and in the process I have identified a number of variables that have not previously been linked with poetic skill. Secondly, I have created an ensemble classifier consisting of an ensemble of several models. The classifier has a holdout sample accuracy of 84.5%. I have then used the classifier to rank a corpus of contemporary American poems. This ranking is a relatively objective means of determining which poems are more like amateur poems and which are more like professional poems. I then discussed these findings in relation to several traditional accounts of poetics as well as several ideological schools of literary criticism. Finally, I discussed how this study could be improved by extending the selection of poems.

Note added in proof

Since the submission of this article a working application which instantiates the algorithm discussed has been created and is available online at www.poetryassessor.com.

Please note that this is a beta test of the application and is not a commercial site.

References

Chali, Y., Hasan, S., and Joty, S. (2009). *A SVM-Based Approach to Multi-Document Summarization, Proceedings of the 22nd Canadian Conference on Artificial Intelligence (CAI 2009)*. Kelowna: Springer-Verlag, pp. 199–202.

Clarke, J. and Paivio, A. (2004). Extensions of the Paivio, Yuille and Madigan (1968) norms. *Behavioral Research Methods*, **36**(3): 371–83.

Collingwood, R. (1938). *The Principles of Art*. Oxford: Oxford University Press.

Eagleton, T. (1976). *Marxist Literary Criticism*. Berkley: University of California Press.

Eliot, T. (1946). What is Minor Poetry? *The Sewanee Review*, **54**(1): 1–18.

Eliot, T. (1998). *The Sacred Wood and Early Major Essays*. Mineola, NY: Dover.

Forsyth, R. (2000). Pops & flops: some properties of famous English poems. *Empirical Studies of the Arts*, **18**(1): 49–67.

Furman, R. (2007). Poetry and narrative as qualitative data: explorations into existential theory. *Indo-Pacific Journal of Phenomenology*, **7**(1): 1–9.

Green, E., Bodrumlu, T., and Knight, K. (2010). *Automatic Analysis of Rhythmic Poetry with Applications to Generation and Translation, Proceedings of the 2010 Conference on Empirical Methods in Natural Language*

Processing (EMNLP) 2010. MIT, Massachusetts, USA, October 2010, pp. 524–533.

Grenz, S. (1996). *A Primer on Postmodernism. Grand Rapids*. Michigan: Wm. B. Eerdmans Publishing Company.

Hawkins, D. (2004). The problem of overfitting. *Journal of Chemical Information and Computer Science*, **44**(1): 1–12.

Kao, J. and Jurafsky, D. (2012). A Computational Analysis of Style, Affect, and Imagery in Contemporary Poetry. *NAACL Workshop on Computational Linguistics for Literature*. http://www.stanford.edu/~jurafsky/kaojurafsky12.pdf (accessed 4 January 2013).

Kaplan, D. and Blei, D. (2007). *A Computational Approach to Style in American Poetry, Proceedings of the Seventh IEEE International Conference on Data Mining*. Omaha, Nebraska: IEEE Computer Society. 28–31 October, pp. 553–8.

Keats, J. (1958). Letter 199, To George and Georgiana Keats. In Rollins, H. (ed.), *The Letters of John Keats*, vol. 2. Cambridge: Harvard University Press.

Paivio, A., Yuille, J., and Madigan, S. (1968). Concreteness, imagery, and meaningfulness values for 925 nouns. *Journal of Experimental Psychology*, **76**(1 Pt 2): 1–25.

Paul, J. (2012). Feminist Philosophy on Language. In Zalta, E. N. (ed.), *Metaphysics Research Lab, Center for the Study of Language and Information*Stanford Universityhttp://plato.stanford.edu/entries/feminism-language/#1.1 (accessed 29 March 2013).

Pennebaker, J., Chung, C., Ireland, M., Gonzales, A., and Booth, R. (2007). *The Development and Psychometric Properties of LIWC 2007*. Austin, TXwww.LIWC.net (accessed 18 March 2008).

Pennebaker, J., Francis, M., and Booth, R. (2001). *Linguistic Inquiry and Word Count (LIWC)*. Mahwah, NJ: Erlbaum.

Poulin, A. and Waters, M. (eds), (2006). *Contemporary American Poetry*. 8th edn. Boston: Houghton Mifflin Company.

Sokal, A. (1996). A physicist experiments with cultural studies. *Lingua Franca* (May). http://www.physics.nyu.edu/faculty/sokal/lingua_franca_v4/lingua_franca_v4.html (accessed 24 March 2013).

Ultsch, A. and Kämpf, D. (2004). *Knowledge discovery in DNA microarray data of cancer patients with emergent self organizing maps. ESANN 2004 Proceedings - European Symposium on Artificial Neural Networks*. Bruges, pp. 28–30. April. pp. 501–6.

Verikas, A., Gelzinis, M., Kovalenk, M., and Bacauskienea, M. (2010). Selecting features from multiple feature sets for SVM committee-based screening of human Larynx. *Expert Systems with Applications*, **37**(10): 6957–62.

Watson, D., Clark, L., and Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology*, **54**(6): 1063–70.

Wordsworth, W. (2009). Preface to a second edition of Lyrical Ballads (1802). In Damon, M. and Livingston, I. (eds), *Poetry and Cultural Studies: A Reader*. Champaign: University of Illinois Press, pp. 21–4.

Notes

[1] I would like to thank Justine Kao for supplying me with the data used in the analysis.

[2] An oversight in Kao and Jurafsky's original paper led to the number of professional poets being recorded as 67. The actual number is 68 (J. Kao, personal communication, April 2 2013).

[3] Personal communication J. Kao, April 2 2013.