In a Sentimental Mood: A Computational Analysis of Emotional Responses to Poetry

P. Thomas Barthelemy
Computer Science
Stanford University
bartho@stanford.edu

Rob Voigt
East Asian Studies
Stanford University
robvoigt@stanford.edu

Jean Y. Wu
Symbolic Systems
Stanford University
jeaneis@stanford.edu

Abstract

What makes a poem more sentimental than the others? What in a poem makes its reader feel more connected? We use computational methods to analyze the correlation between the features of poetry and the types of emotive responses it elicits. We attempt to predict emotional response distributions for new poems, and discuss which features contribute most significantly to these predictions.

1 Introduction

Poetry is when an emotion has found its thought and the thought has found words.

— Robert Frost

Literature in general and poetry in particular present unique challenges for natural language understanding systems. Literary scholars often articulate the manner in which the primary purpose of literature is deviance, in some sense, from the common expectations we hold of human language. Raymond Chapman describes literature as “the art that uses language,” and Viktor Shklovskij notes that in poetry in particular we consistently find “material obviously created to remove the automatism of perception.” In Shklovskij’s terms, literature effects a “defamiliarization” that surprises, delights, and moves to emotion in a way normal language does not.

It is fascinating that poetry is often able to concisely deliver a high emotional impact to its readership, and indeed this is an aspect of poetry that sets it apart from other genres of text. Literary interpretations of this characteristic of poetry have focused on highly contextual, semantic, and topical factors; for example, consider T.S. Eliot’s concept of the “objective correlative . Eliot proposes that emotion in poetry is generated by means of ”a set of objects, a situation, [or] a chain of events which shall be the formula of that particular emotion,” that is, by placing the reader at the center of an artfully described - and therefore richly imagined - context that would naturally generate such an emotion.

There is also a common, and perhaps somewhat contradictory, emphasis in the literary community on “formal devices, including aspects of the structure of poetry as well as literary devices that poetry often employs, such as alliteration, rhyme, meter, and other forms of wordplay and creative use of language. Indeed, some scholars have suggested that our emotional response to works of art comes much more from form than from the act of meaning, an interpretation aimed at explaining why we respond far more emotionally to music and poetry than prose or other genres of text . [pg 158] Researchers in psychology and

However, of yet we lack computational empirical studies that computationally identify formal and linguistic features of poetry that contribute to its capacity to produce an emotional response in its readers. Therefore we propose to collect a dataset of poems and a set of associated responses to discover the features of poetic texts that correlate highly with a strong emotional response. In so doing, we aim to further our understanding of what makes poetry *poetic*, and identify the extent to which such linguistic features might contrast with broader-scale semantic and contextual considerations.

2 Related Work

?) used computational methods to classify aesthetics of contemporary poetry. In particular, they analyzed diction, sound devices (rhyme, alliteration, and assonance), and imagery. As a proxy for a labeling poems as aesthetic or non-aesthetic, Kao and Jurafsky simply used the classes of professional versus amateur, which is expected to closely represent the former two categories with the advantage of having obvious labels.

Kao and Jurafsky hypothesized that diction would be important for classifying poetry. Poetic language is often “intentionally ambiguous, attempting to capture multiple meanings simultaneously. Additionally, it is more likely to include uncommon “strange words for the purpose of being distinguished. For this latter point, it is hypothesized that poetry would include more words with lower word frequencies. It was also hypothesized that poetry would utilize more varied vocabulary–“varied meaning including more word types, and avoiding the repeat of words. However, results showed that professional poets did not use more “strange words; words used by professional poets were not significantly more unusual from words used by amateur poets. On the other hand, poets did use more distinct word types.

Additionally, it was observed that professional poets use more concrete words. Essentially, this could be viewed as a measure of imagery–imagery is conveyed through concrete details, and concrete details require non-abstract language. Similarly, professional poems were less likely to include psychological terms or positive/negative emotional terms, which further suggests that poets prefer to explain emotions via scenario description. In short, a professional poets follow the adage “show, dont tell.

Finally, with regard to form, professional poetry employs far fewer overt sound devices than does amateur poetry. So, though the findings of Tizhoosh et al. suggest that poetry is easily recognized by form, Kao and Jurafsky suggest that good poetry uses these cues far more sparingly. Further, though the perspective that the goal of poetry is to be distinguished (i.e. the poetry-as-deviance perspective) is weakened by the absence of strange words in poetry, it is revived in the observation that good contemporary poetry defies conventional poetic form.

3 Dataset

For this study, we propose a free-text to free-text framework for analysis. That is, we frame the problem as a classification task where each poem acts as a training example from which we extract a set of linguistic and poetic features. Each poem is then associated with the set of free-text responses to that poem from which we extract data to act as the poem’s “response label.” Various implementations [wrong word? fix] of these response labels are described later in the paper.

Data Collection

For a first pass at data collection, we ran a trial experiment on Amazon’s Mechanical Turk service with five ten-line poems. We showed each poem to 20 Turkers and asked them to read and digest its contents. We then asked for them to provide a free-text response describing their personal, subjective emotional reaction to the poem.

However, upon manual review of the results, we found them to be of roughly similar quality to comments provided on the website `PoemHunter.com`, and therefore changed our approach for data collection to web-scraping, to allow both for a larger dataset and a more varied selection of poems. We treat the comments for a given poem as our free-text responses for the purposes of this study.

We began by scraping the ”Top 500 Poems” section of the website, comprised of poems by professional poets. Since there are many thousands of poems on the site and comments are relatively sparse, these poems offered a high density of comments for collection. It is not clear, however, how these ”Top” poems were selected (it is evidently not based upon rating or number of comments), and so this selection has the potential to skew our results.

Therefore, we scraped an additional 25,000 poems from the site by following links to poets in the ”Top 500 Poets” list. Since the distribution of comments is very sparse and scraping is a time-consuming process, we scraped and only use poems from this set with over 10 comments in response; this requirement cut down this additional data source to only approximately [do exact number?] 500 new poems, each with more than 10 comments.

Corpus Composition

4 Methodology

Our task is quite generally defined as the prediction of a human response to poetry. To this end, we use features to describe the poems and their resulting comments. We use latter to capture the human response, and we use the former to predict the latter.

Poem Features

Poetic features are in part taken from **?**), where they were used to analyze the aesthetics of poetry, and Tizoosh et al, where they were used to classify poetry. We can divide them into three groups: orthography, sound device, and diction.

Orthographic Features Orthographic features capture the *shape* of the poems. To decribe this, we used the following features: number of lines, number of stanzas, number of lines per stanzas, number of words per line, and type token ratio.

Sound Device Features Uniquely prevalent to poetry, sound device is a strong descriptive feature for both identifying poetry (Tizhoosh et al.) and classifying profession versus amatuer poetry. We used the CMU pronunciation dictionary, which maps words to phonemes.

In particular, we quantified perfect rhyme, slant rhyme, and alliteration. Perfect rhyme is defined as rhyme having the same ending vowel sound but differing consonant sound preceding it. Slant rhyme is defined as having the same ending consonant sound, but a different vowel sound preceding it. To simplify feature extraction, we avoid searching for particular rhyme schemes (e.g. aabb, ABAB), and simply check whether the ending word of a given sentence rhymes with the ending word of one of the previous two sentences. This ignores cases in which rhyme skips two lines, and would grossly underestimate the rhyming of, say, a Petrarchan Sonnet (abbaabbacdecde), though these rhyme schemes appear rarely.

We capture alliteration by counting the words within a fixed window having the same starting consonant sound. This is an approximation, as alliteration is rather defined as the repetition of stressed consonant sounds, which may include sounds occuring within a word.

Additionally and unlike (**?**) and Tizhoosh et al, we added features to quantify the proportion of nasals, fricatives, stops, and liquids.

Diction

Comment Features

We focus on two main ways of categorizing comment response: orthography and sentiment.

Orthographic Features At first glance, there is a stark contrast between those comments that are thoughtful and lengthy and those comments that simply praise the poet in few words. Thus, we attempt to distinguish between these types of responses by counting the average response length (in words) and the type-token ratio.

As a technical note, we use the log of the average word length. Intuitively, this captures the notion that a difference in comment length of 10 and 15 is much larger than a difference in 50 and 55. This decision was justified in that it was easier to predict than the actual average word length. As an additional technical note, we extract the type-token ratio for a particular poem's response by treating all comments as one document. Then, we sample 100 words from each response and calculate the type-token ratio for this. This offers a more principled way of comparing type-token ratios across differently sized documents.

Sentiment Features It is predicted that human response will vary depending on the poems; some might elicit joy, others sadness. To this end, we attempt to measure the distribution and the magnitude of sentiment words. To measure the distribution of sentiment words, we use the NRC lexicon to associate given words with categories. For example, the word "abacus" is associated with the "trust" category. In all, there are 10 categories: anger, anticipation, disgust, fear, joy, negative, positive, sadness, surprise, and trust.

Additionally, we try to categorize the magnitude of the affect response by counting the frequency of affect words in a particular comment. We use the Linguistic Inquiry and Word Count (LIWC) dictionary to define affect words. We define the *affect ratio* as the ratio of words contained in this affect lexicon divided by the total number of words.

5 Experiments

[maybe description of overall experimental setup here? like, we use what classifier, what experimental setup. note that poem-side features are exactly the same in every case but wer ejust changing what comment-side labels were trying to predict]

5.1 Predicting Emotional Distribution

Our first pass at this task was to predict the responses. Performance would thus be based on the KL divergence of the predicted sentiment distribution and the actual distribution. This proved to be a difficult task, however, as our predicted distribution was no better than simply average distribution over all poems. Further analysis revealed that the word distributions are quite similar across all poem responses, as shown in Figure **??**. Notable from this figure, however, is that there is a seemingly large difference in the proportion of affect words, which examine in the next task.

5.2 Predicting Proportion of Affect Words

Next, we aim to characterize the magnitude of affect words in the user response. We describe this concept with the relative proportion of affect words. Multiple poetic features are correlated with the affect ratio, as shown in Figure **??**. Notably, the strongest correlated features with $p \leq 0.005$ are: HGI-positiv, HGI-concrete, NRC-joy, NRC-trust, NRC-anticipation, type-token ratio, perfect rhyme score, and proportion of stops. When using such words, we are able to predict the affect ratio with 34.6% reduction in error over simply predicting the average.

5.3 Prediction of Average Response Length

Clearly, there is an identifiable difference in the text of the responses. With the goal of further clarifying this difference, we attempt to predict the average response length. Using the same features as above, we are able to predict this with 15.7% reduction in error over guessing the average. The features correlations with average response length was the opposite of their correlation with average response length. That is, for example, though increased type-token ratio in the poem was correlated with decreased affect ratio, it was also correlated with *increased* comment response length.

5.4 Prediction of Type-Token Ratio

Additionally, we attempted to predict the type-token ratio of the responses given the poetic features identified as most useful in predicting affect ratio. Feature correlations with type-token ratio were similar to those with average response length. That is, for example, poem type-token was positively correlated with both average response length and comment type-token ratio. We were able to predict type-token ratio with 17.0% reduction in error over guessing the average.

A summary of our findings are provided in Table **??**.

Table 1: This table shows the comment feature and our model's improvement over guessing the average

Comment Feature	Improvement
Affect ratio	34.6%
Average comment length	15.7%
Type-token ratio	17.0%

6 Discussion

In this paper,

justification / explanation of phonetic features: Holland quote to use in poetic language [...] particular sounds involve muscular actions that somehow match the sense. pg 136

7 Future work

Acknowledgments

Do not number the acknowledgment section. Do not include this section when submitting your paper for review.

References

Justine Kao and Dan Jurafsky. 2012. A computational analysis of style, affect, and imagery in contemporary poetry. *NAACL-HLT 2012*, page 8.

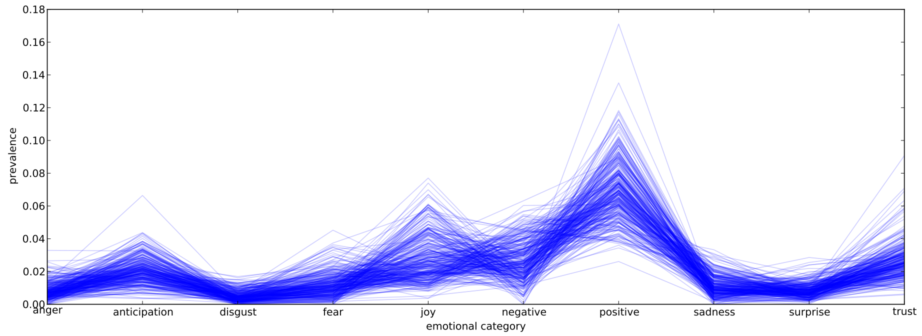

Figure 1: A histogram of word distributions of the response for each poem. Here, each poem is represented by one line. The heavy overlap suggests that there is much similarity in the word distribution over comments.

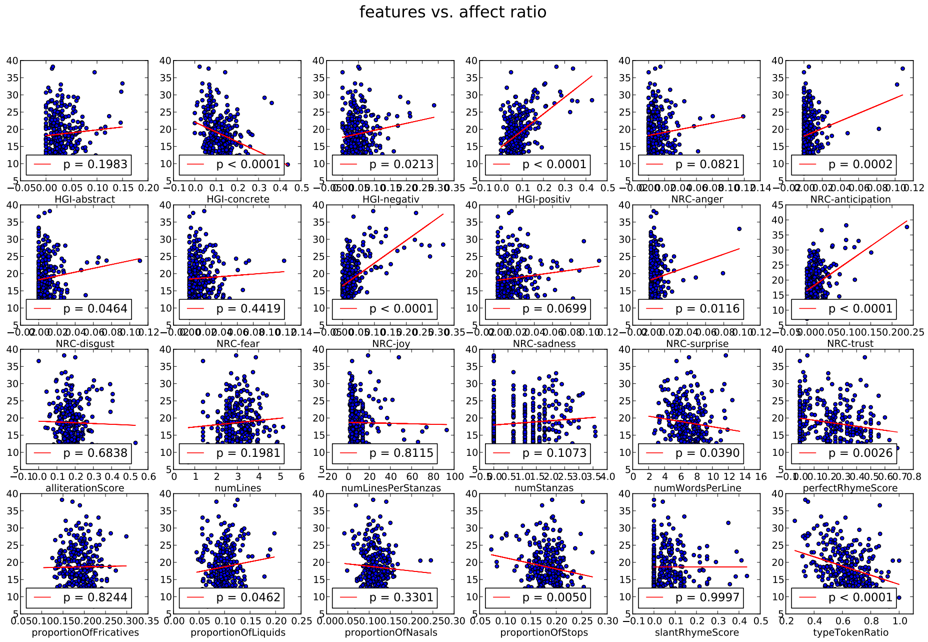

Figure 2: A histogram of word distributions of the response for each poem. Here, each poem is represented by one line. The heavy overlap suggests that there is much similarity in the word distribution over comments.