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. Chapman (1973) describes literature as "the art that uses language," and Shklovsky (1965) notes that in poetry in particular we consistently find "material obviously created to remove the automatism of perception." In Shklovsky'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 Eliot (1920)'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 (Brooks, 1956; Packard, 1994; Turco, 2000). Indeed, scholars in the critical school known as "reader-response criticism" 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 (Holland, 1989).

Such claims have been tested by experimental studies from researchers in psychology and literary criticism. However, of yet we lack empirical studies in the computational linguistics community identifying 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

Previous literature has largely focused on classification Tizhoosh et al. (2008) compared five different strategies for computational classification of documents as poetry or non-poetry. To do so, they focus on three categories of features: shape, meter, and rhyme. These they define as topographical features of poetry, or features that can be examined simply by looking at the poem. They also mention other elements of poetry that are not used for classification in their work: sonic features (e.g. rhythm and meter), sensory (i.e. elements appealing to emotion), and ideational (e.g. grammar and syntax).

Additionally, Tizhoosh et al. attempted to characterize the meaning of poetry. One can get a sense of the imagery of the poem by counting the number of nouns, verbs, adjectives, and adverbs in each phrase. Also, phrase repetition is more common to poetry than prose.

Ultimately, when combining word frequencies and poetic features, they were able to identify poems with 97% accuracy. The authors point out that the word frequencies and poetic features tended to make different mistakes, so the combination of the two sets produced a robust algorithm. In particular, shape was the best indicator. Rhyme and meter added little, although it is hypothesized that it would be more useful in cases of short communication (e.g. emails, chat room comments) in which the shape is not strongly indicative.

Kao and Jurafsky (2012) 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, don't 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.

Brooke et al. (2012) identified stylistic shifts within one poem, T.S. Eliots *The Waste Land*, a poem which includes multiple interweaving voices, styles, and perspectives. Their goal was to use computational methods to identify points of a likely change in voice. To do this, they identified points at which there was a high degree of difference between the window of words before and after the proposed point of transition.

Surface features included the number of syllables, punctuation and line break frequency, frequency of certain parts of speech (in particular nouns, verbs, adjectives, and adverbs) and pronouns, verb tense, type-token ratio, and a contextuality measurea measure based on part-of-speech tags created by Hey-

lighen and Dewaele (2002).

Additionally, they used extrinsic features like sentiment polarity and extremity, formality, and proportion of words included in the Dale-Chall (1995) Readability list and the General Inquirer dictionary. In general, the extrinsic features performed slightly better than surface features, but both complemented each other in their evaluation against a gold-standard segmented version of the text. Combining these results with that of Tizhoosh et al. shows that, though form features best distinguish poetry from non-poetry, both form and semantic features best capture the variance in style.

3 Dataset

For this study, we propose a text-to-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. The poem is then associated with its set of free-text responses from which we extract data to act as the poem's *response label*. Various extraction methods of these response labels are described later in the paper.

3.1 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 similar in 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. Because 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 unclear, 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.

3.2 Corpus Composition

We filtered out poems that have less than 10 comments, which reduced our dataset size to 302 poems. We extract a total number of 15,900 comments, composed of 27,431 unique words, from these poems; on the average, there are 52 comments per poem. Each comment contains an average of 3.54 Affective words, as defined by the Linguistic Inquiry and Word Count (LIWC) dictionary (Pennebaker et al., 2001).

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 the latter to capture the human response, and we use the former to predict the latter.

4.1 Poem Features

Poetic features are in part taken from Kao and Jurafsky (2012), where they were used to analyze the aesthetics of poetry, and Tizhoosh et al. (2008) 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. We used the following features: number of lines, number of stanzas, number of lines per stanzas, number of words per line. When calculating the number of stanzas and number of lines, we use the log of the total value. 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.

Sound Device Features. Uniquely prevalent in poetry, sound device is a strong descriptive feature for both identifying poetry (Tizhoosh et al., 2008), as well as classifying profession versus amatuer poetry (Kao and Jurafsky, 2012). 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.

It should be noted here that we have used a rather strict definition of *rhyme*. One might say, for example, that "lamentable" and "preventable" are neither slant rhyme (because the vowel sounds are identical) nor perfect rhyme (because the sounds preceding the vowel are identical). Future work could combine other varieties of rhyming as features.

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 occurring within a word.

Additionally and unlike Kao and Jurafsky (2012) and Tizhoosh et al. (2008), we added features to quantify the proportion of nasals, fricatives, stops, and liquids.

Diction. Similar to Kao and Jurafsky (2012), we try to capture the use of *positive/negative* words and *abstract/concrete* words, which we derive from the Harvard General Inquirer wordlists (Stone et al., 1962). The positive (`positiv`), negative (`negativ`), and abstract (`abs@`, `abs`) categories can be used as-is, however there is no category for concrete words. Thus, we construct them as the union of pre-existing categories that describe concrete items ("space", "object", "color", "place").

We also attempt to describe affect words and their intensities. To this end, we use the NRC dictionary (Mohammad and Turney, 2010), coupling with sentiment scores collected from Amazon Mechanical Turk[1] to determine the intensity of an affect word in a given category. For the categories *joy* and *trust*, we directly substitute the sentiment score of the word for its intensity. Conversely, the intensity for the categories *anger*, *disgust*, *fear*, and *sadness* is the inverse of the sentiment score, that is

$$intensity = 1 - score_{\texttt{pos/neg}}.$$

For the categories *anticipation* and *surprise*, we obtain the intensity using:

$$intensity = 2 \times (score_{\texttt{pos/neg}} - 0.5).$$

4.2 Comment Features

We focus on three ways of categorizing comment responses: popularity, orthography, and sentiment.

Popularity. Attempt to capture the popularity of the poems by counting the total number of comments and using the poem score, which is on a 1-10 scale.

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.

We use the log of the average word length for similar reasons as mentioned in the poem feature section. This decision was further 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 word-emotion association lexicon (Mohammad and Turney, 2010) to associate given words with categories. For example, the word "hug" is associated with the *trust* category. In all, there are 10 categories: *anger, anticipation,*

[1]These sentiment scores are taken from the Stanford Sentiment Treebank.

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 (Pennebaker et al., 2001) to define affect words. To compare the number of affect words between two different sets of comments, we define the *affect ratio* as the ratio of affect words divided by the total number of words.

5 Experiments

The goal of our experiments was to predict the comment features using the poetic features. During all experiments but the emotional distribution prediction, we use a linear regression model. Due to our relatively sparse data, we use 10 fold cross validation. A summary of our findings are provided in Table 5.

Comment Feature	Improvement
Affect ratio	34.6%
Type-token ratio	17.0%
Average comment length	15.7%
Poem score	0.3%
Number of responses	-0.3%

Table 1: Comment feature and our model's improvement over guessing the average. Performance is measured as the RSS of the error.

5.1 Predicting Emotional Distributions

Our first pass at this task was to predict the distribution of emotional associations in a set of responses for a given poem. Therefore we extract poem-side features as described above, and define the label for a set of comments as the distribution of word-associations for each category in the NRC lexicon. We then use a MaxEnt classifier for classification, with the objective function set to minimize KL divergence between the classifier output and the training labels.

At test time, then, performance would be based on the KL divergence of the predicted emotional 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 while the emotional word-association *distributions* are quite similar across all poem responses, as shown in Figure 1, the primary source of variance was emotional *magnitude*. Therefore, we decided to change our experimental task, as described in the following sections.

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 Table 5.2. 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 features, we are able to predict the affect ratio with 34.6% reduction in error over simply predicting the average.

5.3 Predicting 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. As observable in Table 5.2, 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 Predicting 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, as shown in Table 5.2. That is, for example, poem type-token was positively correlated with both average response length and comment type-token ratio. We were able

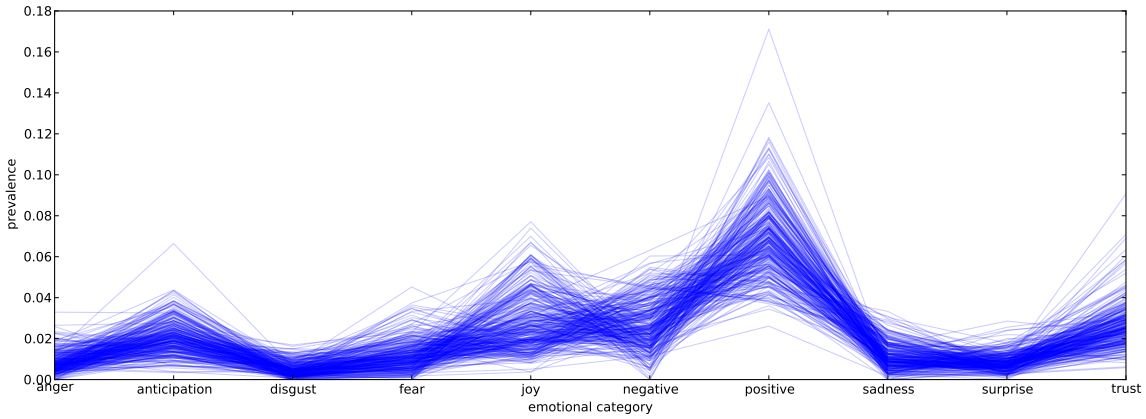

Figure 1: A histogram of word distributions of the response for each poem. Here, each poem is represented by one line. The heavy overlap suggests that there is much similarity in the word distribution over comments.

Feature	Affect ratio		Type-token ratio		Avg. comment len.		Rating		Num. comments	
	correl.	p	correl.	p	correl.	p	correl.	p	correl.	p
NRC-joy	0.51	< 0.0001	-0.39	< 0.0001	-0.34	< 0.0001	0.06	0.1883	-0.07	0.2583
HGI-positiv	0.51	< 0.0001	-0.35	< 0.0001	-0.31	< 0.0001	0.07	0.1531	-0.10	0.1013
NRC-trust	0.42	< 0.0001	-0.25	< 0.0001	-0.29	< 0.0001	-0.02	0.6105	-0.11	0.0578
type-token	-0.32	< 0.0001	0.24	< 0.0001	0.23	0.0001	-0.14	0.0019	0.02	0.7834
HGI-concrete	-0.30	< 0.0001	0.12	0.0395	0.14	0.0189	-0.07	0.1130	0.01	0.8386
NRC-anticipation	0.22	0.0002	-0.09	0.1223	-0.15	0.0109	-0.01	0.8357	-0.08	0.1523
perfect rhyme score	-0.18	0.0026	0.16	0.0075	0.16	0.0080	-0.05	0.3116	-0.09	0.1171
proportion of stops	-0.16	0.0050	0.18	0.0026	0.18	0.0021	0.03	0.4866	0.11	0.0594

Table 2: Features most strongly correlated with affect ratio.

to predict type-token ratio with 17.0% reduction in error over guessing the average.

5.5 Predicting Average Rating

Kao and Jurafsky (2012) suggest that poem aesthetics can be captured using poetic features. Using public approval as a mark of aesthetic value, one should be able to predict the average poem score using our features. However, we were unable to predict this feature well. This is also expected from Table 5.2, which shows that features for this metric did not have both high correlation and low p-value, which holds even for the omitted features.

5.6 Predicting Number of Comments

Similarly, one might use the number of comments as a mark of aesthetic value. Increased number of comments could be seen as a mark of increased popularity, or of engendering more discussion. However, we were unable to predict this either.

6 Discussion and Error Analysis

In this paper, we drew upon earlier computational work in the analysis of poetry to develop a set of linguistic and poetic features which we then used in a series of classification tasks aimed at predicting the *emotional* content of responses for a given poem. While predicting emotional word-association distributions proved to be a dead end, we achieved interesting results in predicting the magnitude of *affect* words, and type-token ratio.

In particular, considering the features that proved to have a statistically significant impact in our prediction tasks, we found that more highly emotive poems (i.e. poems eliciting a higher affect ratio) tend to have strongly distinguished sentiment diction (more positive words, more words associated

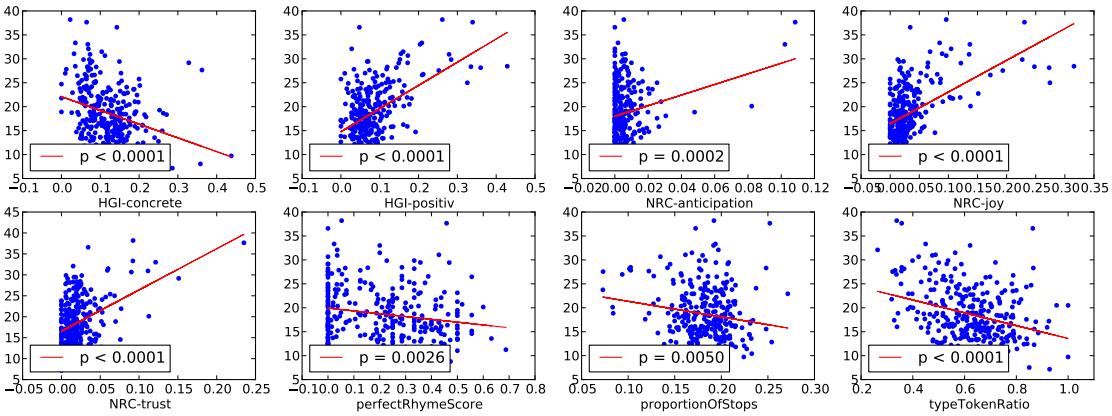

Figure 2: A collection of scatter plots showing the correlation of the poem features and the comment affect ratio.

with "joy,"trust, and "anticipation), fewer concrete words, a lower type-token ratio.

These results were surprising to a certain extent, not in the least because at first glance they appear to contradict **?**). In their work, Kao and Jurafsky found that concrete words and a higher type-token ratio were indicators of poetry written by professionals, as compared with poetry written by amateurs. One might assume that professional poetry is far more likely to be effective at producing an emotive response, so to find that within our dataset of professional poetry opposite trends produced more *emotive*' poetry per our metrics is a finding that requires a further explanation.

Looking by hand at the sets of responses which we found to have a high proportion of affect words, we notice a large contrast between commentary having high affect ratio and commentary having high affect ratio:

- *what beautiful use of words. lovely poem.*
- *you two are a couple of losers.*

contrasted with comments having low affect ratio:

- *christ's life is plausible. however, consider the theme of ambition: what it is; whether it is neutral or with the power to possess good or evil; and its source. then think to how 'we people on the pavement' took to a person whose skeletons were not on public display.*
- *this is poetic therapy at its very best. life-changing and liberating. i can't stop reading it...*

Indeed, the high aspect ratio comments tended to be very short ones, as shown in Figure 3. Further examination shows that affect ratio, log of average comment length, and comment type-token ratio are all strongly correlated with each other. We can offer an interpretation: some comments are very short and offer praise (or critique, as shown above) without explanations, and other comments are lengthy and detailed.

justification / explanation of phonetic features: Holland quote to use — "in poetic language, . . . particular sounds involve muscular actions that somehow match the sense." (Holland, 1989).

7 Future work

We expand our dataset, we scraped 40,000 additional poems by the top 400 poets on the Top 500 Poets list from the same website. Since the distribution of comments is very sparse and scraping is a time-consuming process, we scraped and only use poems with at least 10 comments in response; this requirement cut down this additional data source to only 521 new poems, each with 10 comments or more.

Acknowledgments

Many thanks to our professors Chris and Bill for a wonderful course, and to our TAs Natalia, Matt, and

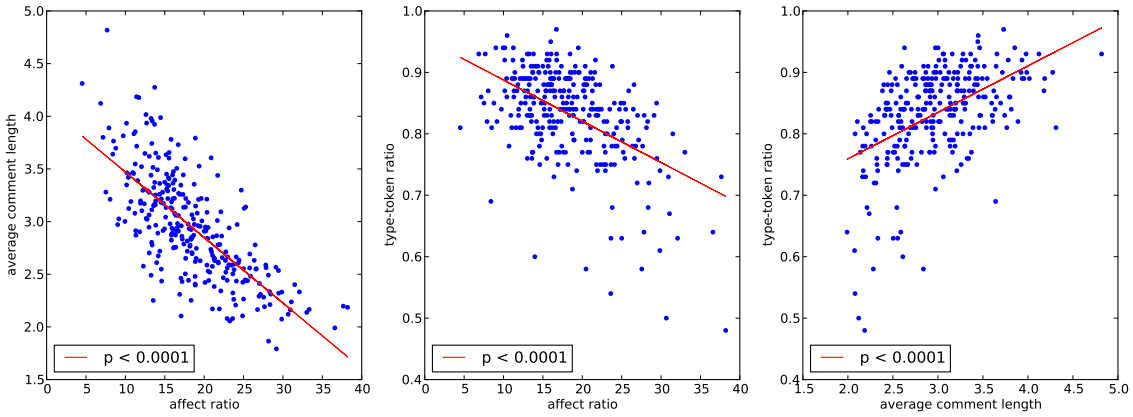

Figure 3: A collection of scatter plots for features of the comments, showing the correlations of affect ratio, log average comment length, and comment type-token ratio.

Kat for their help throughout. Thanks in particular to Chris for his specific suggestions on this project, and for providing us with access to the PragLab account for our preliminary study on Mechanical Turk.

References

Julian Brooke, Adam Hammond, and Graeme Hirst. 2012. Unsupervised stylistic segmentation of poetry with change curves and extrinsic features. *NAACL-HLT 2012*, page 26.

Cleanth Brooks. 1956. *The well wrought urn: studies in the structure of poetry*, volume 11. Harvest Books.

Raymond Chapman. 1973. *Linguistics and literature: an introduction to literary stylistics*. Edward Arnold London.

TS Eliot. 1920. Hamlets and his problems.

Norman Norwood Holland. 1989. *The dynamics of literary response*. Columbia University Press New York.

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

Saif M Mohammad and Peter D Turney. 2010. Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 26–34.

William Packard. 1994. *The Poet's Dictionary: A Handbook of Prosody and Poetic Devices*. Collins Reference.

James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*.

Victor Shklovsky. 1965. Art as technique. *Russian formalist criticism: Four essays*, 3.

Philip J Stone, Robert F Bales, J Zvi Namenwirth, and Daniel M Ogilvie. 1962. The general inquirer: a computer system for content analysis and retrieval based on the sentence as a unit of information. *Behavioral Science*, 7(4):484–498.

Hamid R Tizhoosh, Farhang Sahba, and Rozita Dara. 2008. Poetic features for poem recognition: A comparative study. *Journal of Pattern Recognition Research*, 3(1):24–39.

Lewis Turco. 2000. *The book of forms: A handbook of poetics*. Upne.