# Can Human Assistance Improve a Computational Poet?

Carolyn E. Lamb, Daniel G. Brown, Charles L. A. Clarke
Cheriton School of Computer Science, University of Waterloo
200 University Avenue West, Waterloo, ON N2L 3G1
c2lamb@uwaterloo.ca, dan.brown@uwaterloo.ca, claclark@plg.uwaterloo.ca

## Abstract

Good computational poetry requires sufficiently interesting poetic phrases to be generated or chosen. Different metrics for determining what makes a sufficiently interesting phrase have rarely been directly compared. We directly compare of a number of metrics—topicality, sentiment, and concrete imagery—by collecting human judgments on each metric for the same data set of human-generated phrases, then having humans judge computationally generated poems chosen to include high-scoring phrases against each other. We find through a quantitative analysis that the output of at least some of these metrics is perceived as better than output using none of these metrics.

## Introduction

Many different methods are used to generate, select, and evaluate the words and phrases of a computationally-generated poem (*e.g.* [2], [5], [9], [10]). The variety of methods available is exciting, with each method potentially shedding different light on the properties of good poetry or on computational creativity in general. However, this variety of methods also presents a problem: different methods are rarely directly compared to allow the assessment of single variables. Instead, each researcher gets an idea for a method to use, implements it, and builds on the implementation as their system evolves. Therefore, it is very difficult to compare different systems to each other. What is more, it is difficult to separate a concept from the details of its implementation. A system that performs badly might be using an inappropriate concept, or it might be using an unsuccessful implementation of a very appropriate concept.

We investigate this problem by building a system that does compare different methods directly, and can abstract a concept away from its computational implementation. Specifically, we are building a system which generates poetry based on the following steps:

1. Mine for rhyming phrases of the appropriate length from Twitter.
2. Use crowdsourcing to obtain ratings for these phrases on various metrics that interest us, such as the emotions in the phrase.
3. Create short poems using a generative system which puts phrases together based on their ratings.
4. Evaluate the appropriateness of the metrics, and the effectiveness of the system as a whole, by crowdsourcing comparative ratings of the different generative poems.

Using this methodology, we can experiment with many different means of selecting lines, while holding everything else constant. We can then perform experimental work in the relative effectiveness of different line selection methods, without confounding them with any of the many other decisions involved in constructing a poem.

Crucially, our approach avoids being constrained by the limits of current natural language processing. Even if a metric, such as a tweet's relevance to a topic, is difficult to calculate computationally, we can obtain good ratings simply by asking a human. It would be nice in future work to be able to automate all these judgments, but our current method allows all metrics to be considered fairly. This also allows our data,

or other data collected in this manner, to be used as training data for a machine learning or regression-based approach in the future.

The three metrics we test in this study are topicality (the extent to which the line reflects the poem's topic), sentiment polarity (the intensity of positive or negative emotions in the line), and imagery (the extent to which the line contains concrete, sensory information). We also test a combined metric generated by adding the three standard metrics together. We computationally create short poems, using phrases with high scores for the metrics, and compare the resulting poems using crowdsourced pairwise comparisons. Poems chosen from phrases with high scores for all these metrics, except for negative sentiment polarity, outperform a baseline approach which does not use ratings or other intelligent methods of phrase selection. Topicality and the combined approach perform especially well. This suggests that computational poetry systems with an emphasis on these metrics will outperform systems unaware of these features.

In the following sections, we describe some existing computational poetry systems and their approaches to selecting lines. Next, we describe the computational poetry system, TwitSong, which we are creating, and how it generates poems based on a Twitter corpus. We also describe our method of obtaining ratings on each metric and of evaluating the resulting poems. Finally we discuss lessons learned and limitations of our approach.

## Related Work

The use of computers to generate poetic text goes back to at least the 1950s (Roque [11] gives a partial review). Previous work in "Found" poetry, in which existing human writing is rearranged by a computer into a poetic form, includes Colton *et al.*'s system [2] which generates poetry based on the news. Hartlová *et al.*'s system [5] generates found poetry using Twitter, but the tweets are selected through a partly manual process and without attention to meter.

Each of these computational poetry systems uses a different metric for selecting lines. Hartlová *et al.* [5] select lines for emotional contrast, using a Support Vector Machine to classify the emotional polarity of source tweets. Colton *et al.* [2] compare four different methods of selecting lines: emotional polarity, topic relevance, "lyricism" (use of constraints such as rhyme) and "flamboyance" (use of a variety of new/different words throughout the poem) as well as combinations of two or more of these. Netzer *et al.* [10] produce haiku using word association norms from a database, while Harrel's GRIOT system [4] makes use of complex semiotic theory. Other systems (*e.g.* Diaz *et al.* [3], Manurung *et al.* [9]) optimize their poems based on rhyme, meter, and grammaticality.

The only researchers we have come across who directly compare different metrics for selecting words or lines are Colton *et al.*. However, Colton *et al.* did not do a formal comparison of poems resulting from their four methods. Instead they found that their program made less use of those methods than intended and they were subjectively disappointed with the resulting output. Colton *et al.*claimed that combining more than two of their metrics did not work well because multiple metrics diluted the effect of others and resulted in poems with no discernible style. However, no evidence for this was provided other than the researchers' opinion.

Other researchers have analyzed the lexical properties of successful human poetry, without trying to generate their own. Kao and Jurafsky [8] found that the most significant difference between professional and amateur contemporary poetry was the increased presence of concrete imagery in the professional poems. Other signs of lexical complexity (*e.g.* larger type-token ratio - the ratio of total number of words to number of different words) were also significant. Simonton [12] found similar results when comparing Shakespeare's most famous sonnets to his less successful ones. Greater lexical complexity and more "primary process" imagery (meaning concrete, visceral, and sensory as opposed to analytical) appeared in the successful sonnets.

There was also a characteristic progression of these linguistic traits through the 14 lines of the sonnet. Hirjee and Brown [6] found that the lexical properties associated with critical acclaim in rap lyrics are not the same as the lexical properties associated with commercial success, but all these effect sizes were small.

While strictly quantitative lexical properties such as type-token ratio are worth looking into, our present work focuses on properties which may be difficult to automatically calculate.

## Method

**The TwitSong System.** We are in the process of building a system, TwitSong, which mines lines and sentences from Twitter, analyzes them for meter and rhyme, and puts them together into a metrical poem. TwitSong uses the Twitter API to gather tweets from a specific time period and filters them based on a keyword or other regular expression. (Another option would be to filter based on hashtags. We chose not to use this method because we estimated based on our own Twitter experience that there could be a large number of tweets about a specific topic which did not use the hashtag for that topic.) It then uses Hirjee and Brown's [6] algorithms, modified to take into account Twitter slang and other setting-specific requirements, to identify the meter and rhyme of each tweet. Tweets are grouped into RhymeSets of two or more rhyming lines, also stratified by meter and length of the phrase, allowing for close but potentially inexact matches ("slant" rhymes) between lines in the set. Either full tweets or single sentences within a long tweet can be used in a RhymeSet. To create a Twitter poem, TwitSong then looks for RhymeSets with the desired number of syllables and selects the best lines based on some metric, given the rhyme scheme and meter supplied. The question we would like to answer in this research is how to choose an appropriate metric.

| Metric | Highest Rated | | Lowest Rated | |
|---|---|---|---|---|
| Topicality | 5 | 2 teams to go #Sochi2014 | 1 | One day he gone say you crowding my space |
| | 5 | Way to go USA Men's Hockey team | 1 | 73205 |
| | 5 | Sochi Winter Olympics day six live | 1 | i have done SOOOO much work this afternoon!!! |
| Sentiment | 5 | I smile when you smile...I love when you care. :) | 1 | i hate how people judge me on my size. |
| | 5 | Love this sport #Olympics2014 | 1 | WERE STUCK IN A SHITTY ANIME DEAN |
| | 5 | The Olympic free skating is so cool!! | 1 | hey fuck Anthony , everyone hates him |
| Imagery | 5 | Food. Food. Food. Food. Food. Food. Food. I love food | 1 | ! ! !!!!!! !!!! #2014 #sochi2014 |
| | 4.67 | 15 Pictures That Will Make Your Heart Stop | 1 | something George Costanza would think about. |
| | 4.67 | Sochi Olympic Park As Seen From Space | 1 | You mess one section up and you pay f |
| Combined | 12.67 | Love this sport #Olympics2014 | 4 | hey fuck Anthony , everyone hates him |
| | 12 | The Olympic free skating is so cool!! | 4 | Nobody owes anyone anything |
| | 12 | Figure Skating judges give it a 9 | 4 | but when I do it, I'm being a dick |

**Table 1** : *Examples of some of the highest and lowest-rated tweets for all three scoring metrics from the Olympics dataset. Theoretically possible Combined scores range from 3 to 15; other scores range from 1 to 5.*

We then used Crowdflower[1], a crowdsourced microtasking service, to gather human judgments. Each tweet for each topic was scored by three Crowdflower workers (the number that Crowdflower's documentation recommends for most tasks) on a five-point Likert scale on three metrics: sentiment polarity (very positive to very negative), topicality (very relevant to very irrelevant), and concreteness (very concrete to very abstract). The exact questions given were as follows:

- Topicality: "How relevant is this tweet to the topic of [topic]?" (Very Irrelevant to Very Relevant)
- Sentiment: "How positive or negative are the sentiments in this tweet?" (Very Negative to Very Positive)
- Imagery: "How abstract or concrete is this tweet?" (Very Abstract to Very Concrete)

The 3 scores given to each tweet on each metric were then averaged. Table 1 shows examples of high and low-rated tweets on each of these metrics.

Topicality might seem like an odd metric to use given that our tweets were already selected by keywords to be relevant to a given topic. For example, the original tweets in our New Year's data set were all posted on December 31, 2013 or January 1, 2014 and all contained the string "2014". However, not all tweets containing the string "2014" were actually tweets about New Year's Eve celebrations. Also, sentences within a tweet, rather than the entire tweet, can be used. Therefore, not every sentence processed by TwitSong has the string "2014" in it. While the New Year's data set contained many tweets about New Year's Eve celebrations, it also contained sentences about other topics, spam, and even a few meaningless strings of numbers. We felt it was plausible that selecting tweets based on a human judgment of topicality might improve poem quality

After collecting ratings, we formatted the data for use with TwitSong, using the Likert scale scores as ratings for each tweet. TwitSong generated a sonnet for each pairing of a topic and a rating metric. Since we tried both positive-sentiment and negative-sentiment sonnets, we considered this four rating metrics even though only three scoring metrics were used. For each RhymeSet in the data and each rating metric, TwitSong selected the highest-rated pair of lines in the set (ignoring pairs that rhymed because they both ended with the same word) and gave the RhymeSet an overall rating equal to the rating of the second-highest-rated line in the pair. This minimum rating method ensured that lines with bad human ratings would not be used in the poem just because they happened to rhyme with a good line; instead, all pairs of lines used would be reasonably good. TwitSong then selected the seven RhymeSets with the highest ratings and arranged them into the format of a Shakespearean sonnet (three quatrains and a couplet, with the rhyme scheme ABAB CDCD EFEF GG). In the absence of more sophisticated processing for order, pairs of lines were placed into the poem in order of ratings, with the highest-rated RhymeSet appearing last.

In addition to the four poems made directly using rating metrics, we created a fifth poem for each topic model by adding all of the scores for each tweet together. Since a tweet cannot have both a positive and negative sentiment score, we selected negative sentiments when making a combined poem about climate change, and positive sentiments for the other three topics. Examples of tweets with high and low scores on this combined metric are also shown in Table 1. The other two components of the combined score were used in the same way for all topics. Based on Colton *et al.*'s results [2], we expected that these combined poems might be less successful than others, but it still seemed intuitively plausible and worth testing that a good poem could be topical, strongly emotional, and concrete all at once.

The output of this process was a set of $4 \times 5 = 20$ computationally generated sonnets. To this data, we then added two control poems for each topic. The first control poem, intended to serve as a lower bound, was constructed by TwitSong without using ratings. When ratings are not available, all potential lines are implicitly rated 0. Thus, TwitSong assembles the first seven valid pairs of rhyming lines it encounters without any attention to their content or meaning. The second control poem, intended to serve as an upper bound, was

[1] http://www.crowdflower.com

made by one of the authors who has published poems in paid journals, and who manually chose appropriate lines from the RhymeSets that were available. This brought the total number of poems in the experiment to $4 \times 7 = 28$.

Examples of TwitSong's output, and of control poems, are given in Table 2.

| **Human** | **Control** |
|---|---|
| In 2014 I'll talk less and listen more<br>live a little more and stress a little less.<br>Never give up,Do it better than before<br>Oh and try to lose some weight in the process | PLEASE FOLLOW ME ? ILY GUYSS 25<br>Skies the limit #NewYears #2014 #BelAir<br>I LOVE YOU VERY MUCH MY ANGEL ¡3 5<br>Oh hey, it's 2014. #Ireallydontcare |
| **Combined** | **Negative** |
| Hey Nashville...2014 is pretty awesome!<br>Happy 2014 friends! Be safe out there!!<br>Had a great New Years Eve at Magic Kingdom<br>We started off 2014 with a prayer | Nothing's changing except people I fuck with<br>2014 already took Uncle Phil<br>2014 already startin off with death<br>started my 2014 off vomiting brill. |

**Table 2**: *Excerpts from poetry used in our study. The Human poems were put together by a human from the tweets available, using TwitSong only to create sets of possible rhyming lines to choose from. The Control poems were put together by TwitSong through arbitrary selection of lines from these sets. Also shown are poems made by TwitSong using the Combined metric (the sum of the topicality, positive sentiment, and imagery scores) and the negative sentiment metric, which performed very poorly. The poems given are from the 2014 dataset. For space reasons, we include only a single stanza from each poem; the full poems are 14 lines long and in sonnet form.*

**Evaluation.** The question of precisely how to evaluate computationally creative systems is currently a topic of great contention (see Jordanous [7] for a detailed review). For the purposes of this study, we used a simple pairwise comparison metric. Human raters were recruited using Crowdflower and each rater was given four pairs of poems. For each pair of poem, the rater was instructed to indicate which poem they preferred, and to justify their choice in one sentence or less.

A second filtering mechanism was then used. Of each rater's four pairs of poems, two were control pairs—comparisons of a human-constructed Twitter poem with a control poem (made arbitrarily by TwitSong without using any ratings) on the same topic. If a rater did not prefer the human-constructed poem to the control poem in both of their control pairs, they were then removed from the data.

We checked the validity of the control pairs as a data cleaning mechanism by getting the two non-poet authors to blindly judge each possible control pair. Both distinguished human-constructed poems from control poems with perfect accuracy. In theory, raters who were answering at random would still have a 25% chance of passing this test. In practice, our filtering mechanism removed about half of the data, which means that approximately two-thirds of the data left can be trusted to be non-random.

There are other conceivable reasons, besides answering at random, why a particular rater might prefer the arbitrary control poem to a human-constructed poem. However, for the puposes of this study, we were interested in constructing computational poems which shared the traits that made human-constructed poems more meaningful and entertaining than arbitrary assortments of tweets. Thus, we were interested only in the opinions of raters who showed a clear preference for human-constructed poems.

Following the removal of unsuitable raters, we were left with a set of pairwise comparisons in which each computationally generated type of poem (those chosen for topicality, imagery, positive and negative sentiment, the combined metric, and the control poems) appeared between 100 and 120 times. For each individual poem, we counted the number of times that the poem was selected in preference to the one next to it, and divided this by the total number of times that the poem appeared. We then ran a one-way ANOVA

to test for significant differences between the six varieties of poem.

## Results

**Scoring.** One concern for us was whether or not the tweet ratings we received on Crowdflower were accurate. We therefore ran these ratings through a few informal tests.

First, we informally looked at the tweets sorted from smallest to largest on each metric. The distribution of which tweets were scored as most topical, off-topic, happy, sad, or neutral accorded very closely with our own intuitions. For example, the tweets which consisted of arbitrary numbers were consistently rated as very off-topic, and tweets expressing joy or strong negative emotions were found at the appropriate ends of the sentiment spectrum. The scores for abstract vs. concrete imagery also accorded somewhat with our intuitions, but we noticed some irregularities in the data. The data was very biased towards rating tweets as abstract. The average rating for a tweet was below 2 on a scale of 1 to 5 (1 being most abstract, 5 being most concrete), with more than half of the tweets rated as 1. Very few tweets were scored as being highly concrete. Table 1 shows examples of the highest and lowest-scored tweets on each metric, using the Sochi Olympics dataset.

We also tested each scoring metric by taking a subset of tweets and calculating the correlation between average Crowdflower rating and the manual rating of one of our authors. Again, crowdsourced workers gave results very similar to our own ratings for topicality and sentiment ($0.77 < R < 0.81$), but less so for imagery ($R = 0.37$).

**Evaluation.** Figure 1 shows the results of our pairwise evaluations for each poem. Human raters preferred the topical, positive sentiment, and combined poems to control poems. Imagery-based poems were rated slightly better than controls. To our surprise, negative sentiment poems performed worst of all, being selected in only 27% of pairs on average (compared with 41% for control poems). Raters' written comments indicated that human raters often reacted very negatively to poems they saw as negative, depressing, angry, sarcastic, or crude. Comments like "Poem A is very negative and makes me angry reading it," and "poem A has too many offensive words" were common in cases when raters chose a different poem over a negative sentiment poem. A single-factor ANOVA demonstrated that these differences were statistically significant ($F = 5.79$, $p < 0.01$) and remained significant with the highest and lowest performing categories removed.

## Discussion

Our work demonstrates quantitatively and with reasonable controls in place that selecting lines based on simple metrics like the ones we have chosen can significantly improve the appeal of the poem to a general audience, and that some metrics perform better than others. Contrary to Colton *et al.*'s reported results, we found that combining more than one filter did not dilute the style of a poem in any problematic way; poems using a combined metric performed as well as poems using the highest performing single metric.

While poems selected by imagery were rated slightly better than controls, the difference was not substantial. This may due to the poor performance of crowdsourced workers at correctly identifying tweets with concrete imagery. Workers on Crowdflower reported, to a much higher degree with this than the other tasks, that it was too difficult. (The imagery task was rated an average of 3.125 and 3.175 out of 5 on Instructions Clear and Ease of Job, respectively; compared to 4.03 and 3.98 for topicality and 4.06 and 4.26 for sentiment.) Given that writing with concrete imagery is a task many beginning poets struggle with, it is perhaps not surprising that people without a poetic background could not be quickly taught to identify such imagery. Such results run counter to our original assumptions, that crowdsourced workers would do better at scoring tweets based on their meaning than a computer. While the workers were good at identifying topicality and

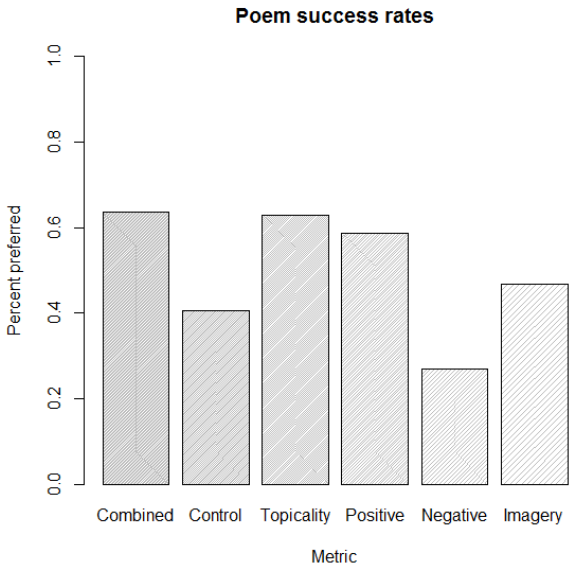


**Figure 1**: *Success rates for types of computationally generated poems in pairwise comparisons with other poems. The height of a given bar represents the number of times a poem from that category was selected in preference to any other poem, divided by the number of times a poem from that category appeared in a comparison. Compared to Control poems (those generated arbitrarily without a line-rating metric) almost every other type of poem did significantly better, except for Negative poems (those generated based on lines with strong negative sentiments), which do significantly worse.*

sentiment, it is plausible that a specialized resource, such as the dictionary of primary process imagery used by Simonton [12], might do better at identifying concrete imagery than most humans.

An alternative explanation for the poor performance of imagery-based poems might be the relative dearth of tweets with good imagery in them. For instance, in the Olympic data, only 96 out of 333 lines were rated more than 3 out of 5 by Crowdflower workers, and 61 rated more than 3 out of 5 by us. With few or no good tweets to choose from, it might simply be more difficult to put a good poem together.

The extremely poor performance of negative sentiment poems was a surprise. It does not take much expertise in poetry to know of poets, such as Sylvia Plath, who are admired for their eloquence in describing negative sentiments. However, there are many possible explanations for why raters in this task would strongly dislike negative sentiment poems. Describing negative sentiments in an engaging manner may be more difficult than describing positive sentiments engagingly, and TwitSong may not be up to the task. Strong negative emotions may not have been a good fit for the subject matter or the casual tone of the poems. Or the raters on Crowdflower, likely to be ordinary people without much poetic background, may have feelings about negative or depressing poetry which differ from those of literary scholars.

There are of course many other metrics which could be tested in this manner, including humor, beauty, presence of poetic devices (such as metaphor and allusion), or how pleasing the line is to say aloud; we have not yet addressed any of these. Our study also leaves unaddressed the question of higher-level constraints for poem generation, such as coherence. This, along with the kind of progression of imagery through a poem discussed by Simonton [12], is a topic for future research. One promising resource for research in dependencies between lines of a poem is Burns's EVE' model [1], which calculates the impact of a poem or joke on the reader based on an information theoretic account of surprise and meaning, in which later lines provide a surprising new interpretation of the lines before them.

## Conclusions

We have established a methodology for directly comparing different line selection metrics independently of poetic form or computational implementation. This methodology has flaws, including imperfect human performance at rating lines in certain kinds of metric. However, for at least some commonly used metrics we have established that this methodology is useful. We have also demonstrated that using a properly implemented, meaning-based metric for line selection produces better results: the poems are preferred by human evaluators. However, using a metric which does not match an evaluator's desires actually makes things worse. Even though the task of making poems from Twitter posts is somewhat frivolous, and despite the subjective and difficult nature of poetry evaluation, the task proved to be meaningful enough that methods of line selection made a significant difference in how the poems were read and evaluated.

## References

[1] K. Burns. EVE's energy in aesthetic experience: A bayesian basis for haiku humor. *Journal of Mathematics and the Arts*, 6:77–87, 2012.

[2] Simon Colton, Jacob Goodwin, and Tony Veale. Full face poetry generation. In *Proceedings of the Third International Conference on Computational Creativity*, pages 95–102, 2012.

[3] Belén Díaz-Agudo, Pablo Gervás, and Pedro A González-Calero. Poetry generation in COLIBRI. In *Advances in Case-Based Reasoning*, pages 73–87. Springer, 2002.

[4] D Fox Harrell. Shades of computational evocation and meaning: The GRIOT system and improvisational poetry generation. In *Proceedings, Sixth Digital Arts and Culture Conference*, pages 133–143, 2005.

[5] Eliška Hartlová and FM Nack. *Mobile social poetry with Tweets*. Bachelor thesis, University of Amsterdam. 2013.

[6] Hussein Hirjee and Daniel G Brown. Using automated rhyme detection to characterize rhyming style in rap music. *Empirical Musicology Review*, 2010.

[7] Anna Katerina Jordanous. *Evaluating computational creativity: a standardised procedure for evaluating creative systems and its application*. PhD thesis, University of Sussex, 2013.

[8] Justine Kao and Dan Jurafsky. A computational analysis of style, affect, and imagery in contemporary poetry. In *In Proceedings of the NAACL-HLT 2012 Workshop on Computational Linguistics for Literature. Montreal, Canada*, pages 8–17, 2012.

[9] Ruli Manurung, Graeme Ritchie, and Henry Thompson. Using genetic algorithms to create meaningful poetic text. *Journal of Experimental & Theoretical Artificial Intelligence*, 24(1):43–64, 2012.

[10] Yael Netzer, David Gabay, Yoav Goldberg, and Michael Elhadad. Gaiku: Generating haiku with word associations norms. In *Proceedings of the Workshop on Computational Approaches to Linguistic Creativity*, pages 32–39. Association for Computational Linguistics, 2009.

[11] Antonio Roque. Language technology enables a poetics of interactive generation. *Journal of Electronic Publishing*, 14(2), 2011.

[12] Dean Keith Simonton. Lexical choices and aesthetic success: A computer content analysis of 154 shakespeare sonnets. *Computers and the Humanities*, 24(4):251–264, 1990.