Incorporating novelty, meaning, reaction and craft into computational poetry: a negative experimental result

Carolyn Lamb, Daniel G. Brown, and Charles L.A. Clarke
David R. Cheriton School of Computer Science
University of Waterloo
Waterloo, Ontario, Canada

Abstract

We present TwitSonnet, a Twitter found poetry system. TwitSonnet attempts to build meaningful poems based on criteria we previously identified as separating good computer-generated poems from bad ones: namely, novelty, meaning, reaction and craft. We show the results of an experiment with human raters that shows that TwitSonnet poems focusing on these criteria are not artistically superior to poems that do not. We discuss the implications of this negative result for TwitSonnet's development, and the general implication of negative experimental results on computational creativity as a field.

Introduction

Computational poetry is a popular area of computational creativity in which computers are programmed construct poems. A variety of approaches have been used for this construction; a summary of the different approaches and their similarities and differences can be found in our previous paper (Lamb, Brown, and Clarke 2016b). These approaches range from simple word substitutions to very sophisticated systems using neural networks to replicate patterns in human poetic language.

One of the many existing approaches to computational poetry is found poetry, in which a computer selects appropriate excerpts of human-generated texts and remixes them into a new poetic work. A few systems, including Ranjit Bhatnagar's Pentametron (Bhatnagar 2012) and Andrei Gheorge's The Longest Poem In the World (Gheorge 2009), generate found poetry by taking text from Twitter. These systems are simplistic, choosing tweets based only on rhyme and number of syllables. Hartlová's Mobtwit (Hartlová and Nack 2013) performs a more sophisticated analysis, creating limericks out of tweets chosen for emotional contrast. However, systematic or falsifiable analysis of what makes a tweet suitable for use in found poetry has not yet been done.

TwitSonnet

TwitSonnet (Lamb, Brown, and Clarke 2015b) is a found poetry system similar to Pentametron (Bhatnagar 2012). Both systems assemble pairs of rhyming, 10 or 11-syllable tweets into a sonnet. Where TwitSonnet differs from Pentametron is that we try to select tweets that are, in ways we will define below, more poetic than others. The hope is that a Twitter sonnet containing more poetic lines will be more meaningful, more entertaining, and potentially more creative than a sonnet containing only arbitrary tweets.

The ability to evaluate unfinished work - including the suitability or unsuitability of potential components of the work - is a vital part of the creative process (Galanter 2012). Throughout TwitSonnet's development, our goal has been to focus on automating this relatively high-level judgment while abstracting away from the low-level generation of language. A system that can be shown to intelligently select lines from a corpus can then be trusted to use intelligent selection on lines of its own.

How TwitSonnet works

TwitSonnet creates topical poems out of tweets using the following stages:

1. **Data gathering.** We use the Tweet Archivist service (Tweet Archivist 2016) to pick tweets containing a topical keyword during an appropriate time interval.
2. **Filtering.** TwitSonnet counts the syllables of the tweets in the gathered data and groups them by end rhyme, using a modified version of the code from Hirjee and Brown's Rhyme Analyzer (Hirjee and Brown 2010). This code, built on top of the CMU Pronouncing Dictionary (Weide 1998), allows for imperfect rhymes with an adjustable threshold. Any tweet which has fewer than the appropriate number of syllables for a sonnet or which does not fit into a rhyme grouping with at least one other tweet is discarded. Tweets with more than the appropriate number of syllables are split into their constituent sentences, if possible. Excessive hashtags and other unpronounceable features of tweets are also removed. Because of these methods, some lines in the rhyme groupings do not contain the original keyword, but are potentially related due to their proximity to the keyword in their original context. Our modified Rhyme Analyzer code can appropriately handle common forms of Twitter slang and misspellings, but discards tweets that contain obvious non-words or are not in English.
3. **Ranking.** Tweets are given scores for desired poetic criteria as described below. Scores are normalized by range

and then the different scores for each tweet are added together.

4. **Selection.** The seven rhyming pairs of tweets with the highest scores (judged based on the second-highest tweet in the rhyming set) are selected to be placed in a sonnet.
5. **Reordering.** Optionally, the selected lines can be reranked and placed in a meaningful order. For example, they could be ordered from the most abstract introductory statements (least imagery) to the strongest concluding image (most imagery). Otherwise, the tweets are ordered according to score, with the highest scoring couplet at the end.

TwitSonnet is a fully functional system which can create a sonnet out of any sufficiently large collection of tweets. From July through the end of October 2016, we posted several of TwitSonnet's poems per week at `http://twitsonnet.tumblr.com/`.

Poetic criteria

There are various ways to evaluate the success of a creative computer system. For this project, we are focusing on the Product perspective (Jordanous 2015) in which the system is primarily judged on the quality of its output. But how, specifically, do we define quality? In previous work (Lamb, Brown, and Clarke 2016a), we developed a set of domain-specific product-based criteria for computer-generated poems by studying the responses of Experimental Digital Media graduate students to a varied and inclusive set of such poems. We grouped the desired traits expressed in these students' responses into four categories:

- **Reaction** - the reader sees the poem as interesting, or has an emotional response, based on their prior experience of poetry.
- **Meaning** - the poem coherently expresses an idea.
- **Novelty** - the poem is new, different, or subversive.
- **Craft** - the poem is written skillfully, with good use of form (if any), imagery, and poetic devices.

These categories bear parallels to, but are distinct from, other existing formalizations for evaluating digital poetry. Criteria similar to Craft, for example, appear in van der Velde's creativity criteria (van der Velde et al. 2015), in Manurung et al.'s domain-specific criteria for computational poetry (Manurung, Ritchie, and Thompson 2012), and in the Creative Tripod model (Colton 2008). A more detailed comparison of our categories to other models appears in our previous study (Lamb, Brown, and Clarke 2016a).

We developed TwitSonnet's algorithm specifically taking these categories into account, as follows:

For **reaction**, we gave a higher score to tweets containing words pertinent to a desired emotion, as measured by the NRC Hashtag Emotion Lexicon, which was created specifically for Twitter (Mohammad and Kiritchenko 2015). The Emotion Lexicon contains eight different emotions. We chose a desired emotion for each poem by measuring which emotions were most prevalent in the gathered data, and then normalizing by the rate of emotion words in non-topical data.

For **meaning**, we chose tweets relevant to a specific topic through a two-step process. First, the data gathering process using Tweet Archivist narrows in on a topic by selecting tweets by time range and keyword. Second, at the ranking stage, we created a trigram frequency data set for the tweet corpus and gave higher scores to tweets consisting of trigrams with high frequency scores.

For **craft**, we did two things. First, as mentioned, we selected tweets for rhyme and meter and arranged them into a sonnet, which is a recognized poetic form. Second, we gave a higher score to tweets containing stronger primary process imagery, as measured using the Regressive Imagery Dictionary (Provalis 1990). The Regressive Imagery Dictionary gives higher scores to "primary process" words relating to physical senses, experiences, drives, and the body, and lower scores to more abstract, "secondary process" words. Selection for such concrete physical imagery in poetry is supported by Simonton (Simonton 1990), who used the Regressive Imagery Dictionary to show a greater presence primary process imagery in more successful sonnets, and by Kao and Jurafsky (2012), who used related measures to show that professional contemporary poetry uses more concrete imagery than the poetry of amateurs.

For the purposes of this study, we did not find a satisfactory method of measuring novelty. Some obvious attempts, such as selecting for unusual trigrams, seemed to only increase the number of off-topic, "random", and nonsensical tweets. In context of our previous study, the category of Novelty refers to interesting juxtapositions, new thoughts, and subversions of existing concepts, not to this type of "mere novelty". We did reduce repetitiveness by placing a limit on the number of times TwitSonnet was allowed to repeat a poem's keywords, replacing repetitive tweets with the highest ranked alternatives that did not contain the topic keywords.

These specific operationalizations of our criteria are made for the specific domain of found poetry, and the criteria would be operationalized differently in a poetry generator which was creating its text from scratch or through a template.

In summary, our system is explicitly built to satisfy our domain-specific product-based criteria. However, like any system, its success at satisfying them in practice needs to be tested empirically. We will now describe how we have tested previous and current versions of TwitSonnet.

Previous evaluation

A proof-of-concept version of TwitSonnet, then called TwitSong, was evaluated using a pair preference study. Non-expert participants compared TwitSong's poems to a control group in which the ranking stage assigned every tweet the same score (Lamb, Brown, and Clarke 2015b). Participants significantly preferred sonnets in which the ranking was based on certain criteria, especially topicality (the equivalent of our current category Meaning), to control sonnets. However, the scoring in this early study was not done by TwitSonnet, but by workers on a crowdsourcing website.

Review coming tomorrow/this afternoon.
doctor strange was amazing. cant wait for Thor.
closer look at the evolved hero costume
Visually stunning, left me wanting more
What is a Doctor Strange collector corps box?
Check out the latest new movie details!
So excited to see Marvel in the parks!
what was your first Doctor Strange comic? #Strange-Tales
I have 10 more tickets to give away
Doctor Strange 8:45 Ill be there
Doctor Strange is pretty, and pretty OK:
gonna lowkey fall asleep in this chair
It better be worth slacking on my dreams!
Doctor Strange (with Christy at Platinum Screens

Figure 1: A sample of TwitSonnet's output, regarding the movie "Doctor Strange". (The keyphrase used was "Doctor Strange", and the time range used was the movie's opening weekend.)

The purpose of the study was to show that line selection based on criteria does, in fact, produce a better poem than arbitrary line selection. We then moved on to the current step of having TwitSonnet do its own, automated line selection.

Evaluating TwitSonnet

We had two goals in evaluating the current version TwitSonnet. First, we wanted to confirm that the effect of the automated scoring was similar to the effect of the crowdsourced scoring. Second, we wanted to improve on the methodology of the previous study by including expert raters, who are more consistent when rating creative artifacts than non-experts (Kaufman et al. 2008). Indeed, in the domain of poetry, judges with little to no poetry experience can have the opposite of the preferences of an expert (Lamb, Brown, and Clarke 2015a).

Method

Experts in poetry can be difficult to recruit for studies. We recruited participants using snowball sampling on the social networks of all three of this paper's authors, particularly the first author, who is a published poet under a pen name.

Participants were asked demographic questions and classified as experts or non-experts. In keeping with the recommendations of Kaufman et al (2008), we based our definition of expertise not in the study of poetry but in experience actively generating successful poetry. Participants who had published poetry in a magazine or collection, read their own poetry at a reading or slam, and/or published digital poetry were considered poetry experts.

The poetry experts consisted of 13 women, 12 men, and 11 non-binary gendered poets. (While this is a serious over-representation of non-binary poets - likely an artifact of the snowball sampling method - we do not expect it to skew our results, as none of the poems in the study pertain to gender or queer/trans* issues.) The median age was 32, ranging from 17 to 56. All but two of the experts were native speakers of English.

The non-experts consisted of 12 women, 19 men, three non-binary, and one non-expert who did not disclose their gender. The median age was 36, ranging from to 21 to 70. 29 of the 35 non-experts were native speakers of English.

As a result of our snowball sampling, most of our "non-expert" participants could actually be considered quasi-experts: they reported that they were regular readers of poetry, had written unpublished poetry for pleasure, taken classes in poetry, listened to poetry podcasts, attended poetry readings, or taught poetry to K-12 students. (An additional form of experience, being a poetry editor for a magazine or other publication, did not appear among non-experts. Seven of our 36 expert participants reported having been a poetry editor.) Only three participants had no significant experience with poetry, and one of these was a graduate of a prose creative writing program. Thus, we would expect less difference between the experts and non-experts in this study then we would see if the non-experts were completely inexperienced.

Each participant was shown 8 poems in a random order, from the same selection of 8 current events topics and 8 emotions. The topics included three topics from recent movies and television, two astronomy topics, a ban on the "burkini" in France, and two topics relevant to the recent 2016 Summer Olympics. Each topic was associated with an emotion from the NRC Hashtag Emotion Lexicon: anger, anticipation, disgust, fear, joy, sadness, surprise, or trust.

Each of these 8 poems was in turn drawn at random from one of three groups. In Group A, poems were generated using steps 1 and 2 from the TwitSonnet process, but not the remaining steps. In other words, these were our control poems, in which no filtering or reordering based on our four criteria was performed. Poems in Group B were generated using steps 1 through 4 (so they were generated and filtered using our four criteria, but not reordered), and poems in Group C used all five steps including reordering. For each of the 8 poems, participants were then asked the following questions, each on a 5-point Likert scale:

1. "How much do you like this poem?" *(Reaction)*
2. "How creative is this poem?"
3. "How well does this poem express the emotion of [emotion]?" *(Reaction)*
4. "How meaningfully does this poem summarize its topic?" *(Meaning)*
5. "How new and different is this poem?" *(Novelty)*
6. "How successful is the imagery in this poem?" *(Craft)*
7. "How cohesive is the narrative of this poem?" *(Meaning)*

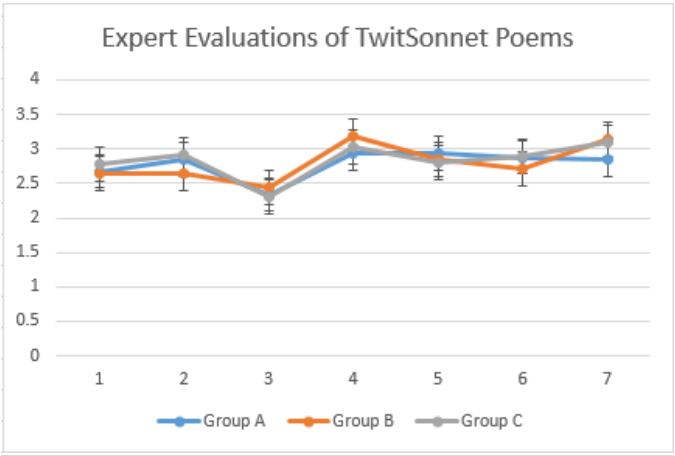
The answers provided at each point of the Likert scale were

- Not at all
- Not much

- A little
- Somewhat
- Very much

Apart from "How creative is this poem?"—an irresistible option in a computational creativity project—each of the questions is designed specifically to assess TwitSonnet's success at one of our four domain-specific categories. Our hypothesis was that the poems from Groups B and C would score higher than Group A on at least some questions, and that Group C would score higher than Group B specifically for narrative cohesion. Participants were also given a freeform text box in which to write any other comments they had about the poems.

Results



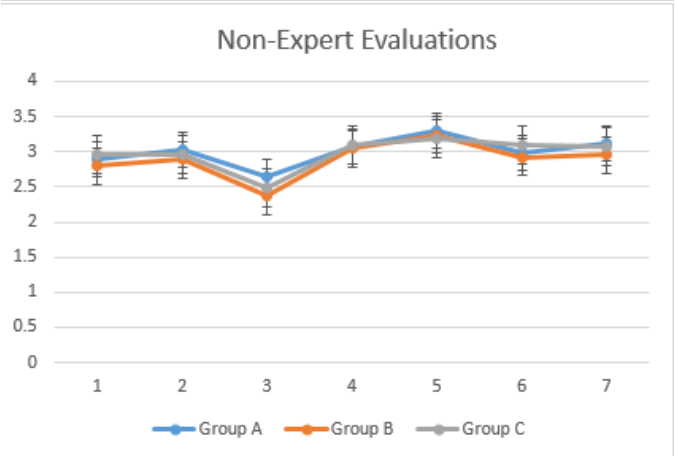


Figure 2: Experts' and non-experts' evaluations of TwitSonnet's poems. The X-axis shows the seven evaluation questions in the same order as they are listed in our Method section. The Y-axis shows answers on a 5-point Likert scale, with 5 being the most positive response and 1 the least positive. Error bars show 95% confidence intervals.

Unfortunately, our hypotheses were not confirmed. As can be seen in Figure 2, there was little difference in either experts' or non-experts' reactions to poems from the different groups. Standard deviations within groups far exceeded the difference in mean between groups, and for most criteria, the size of the 95% confidence interval also exceeded this difference. The largest apparent difference was in narrative cohesion as judged by experts, in which groups B and C ($\bar{x} = 3.13$ and 3.01, respectively) outperformed group A ($\bar{x} = 2.84$)—but the standard deviations of these groups on this question were 1.29, 1.27, and 1.28, more than five times the size of this difference. None of the differences were statistically significant.

We compiled the most common freeform comments by experts and nonexperts. Experts stated that the poems seemed random and choppy; often there would be small sections with a satisfying juxtaposition but they would be mixed with other lines that didn't fit. There was too much focus on rhyme and meter at the expense of content, with several experts stating they would have preferred if the poems did not rhyme. There were also too many lines that trailed off in the middle of a sentence or even a word. However, several experts said that they found the idea behind the project very interesting in spite of any criticism they might have of the poems. Nonexperts had fewer comments and responded more to surface features of the poem: for example, several nonexperts said they would have preferred not to see hashtags in the poems, as well as typos, bad punctuation, and other errors. Nonexperts also agreed with experts that the poems lacked coherence.

Discussion

The negative result here is surprising because, in our previous study, the difference between the equivalents of Group A and Group B was statistically significant (Lamb, Brown, and Clarke 2015b). There are three possible explanations for this.

First, perhaps the difference is due to a difference in how we performed the evaluation this time (for example, Likert scales vs pairwise preferences). While this is important to consider, we believe that, with the possible exception of narrative cohesion measured by experts, the current study shows a striking lack of difference between groups, which is not attributable merely to the use of a less sensitive statistical method.

A related suggestion is that perhaps the current events topics chosen in this study were not the correct choices. For instance, raters might have had stronger opinions about the emotions expressed in a poem if the poem was on a more polarizing topic. Such polarizing topics are plentiful in current events, especially as the study was run during the lead-up to the divisive 2016 U.S. presidential election. TwitSonnet's online incarnation did indeed create poems on divisive political topics: an example is shown in Figure 3. We chose not to include these poems in the study so as not to conflate a rater's political opinion with their artistic opinion of this poem. This may or may not have been the correct choice.

Second, perhaps the automated judgments we are using contain too much error when compared to human judgment and are thus not suitable for this purpose. We have deliberately used computationally simple methods in order to pro-

Final Presidential Debate (10/19)
Donald Trump is master of the head fake
This East Texas pole shows a leftward lean
but goodnight this all debate gave me headache
Much smarter than his brother Crooked John
An interesting debate is taking place
While America tuned in to watch Don
Trump doing the deniro mobster face
started by her very sleazy campaign
Debate Watch Party SAC 305
Donald Trump is LITERALLY insane
watching guy fieris diners drive-ins and dives
That was the sound of women everywhere
Its a humanitarian nightmare

Figure 3: A TwitSonnet poem posted online, using the keyword "debate", immediately after the 2016 U.S. presidential election debates.

cess large numbers of tweets on a large number of topics. It is possible that these methods are simply not up to the tasks assigned them.

Third, while the focus on this study was on the ranking and ordering steps, the filtering step has also improved since the previous study. Humans are unlikely to judge nonsensical tweets as being very topical or as having a clear emotion. Automated judgment is less sensitive to nonsense, and in addition, our filtering step has improved at automatically removing nonsense from both ranked and unranked poems. Thus, it is possible that some of the effect in the previous study was due the ranking step reducing nonsensical tweets, and that this reduction is no longer noticeable in the current study.

Filtering for rhyme and meter (craft) and the use of keywords in data selection (topicality) was already in place in very close to its current form in the previous TwitSong study, so these steps alone cannot be used to account for our current results, but it should be noted that due to these techniques, even the poems in Group A are not "raw" control poems in the sense of having no attention paid to the four criteria. Neither would, for example, Pentametron's poetry, since it too is selected for rhyme and meter (Bhatnagar 2012). The use of a *pure* control group - for example, a completely random selection of English-language tweets - would likely produce something closer to a significant result. However, it would not tell us if our filtering techniques, specifically, were working as intended.

Pearce *et al.* (2002) and, more recently, Bown (2014) call attention to the need for falsifiability in computational creativity evaluation. Unfortunately, the use of falsifiable techniques will sometimes produce a negative result. A negative result does not necessarily invalidate the worth of the project, but it is a sign that the creative system in its current form is not performing as intended.

There are several possible responses to this specific negative result. First, we could try performing a different evaluation. Second, we could modify our line selection techniques and engage in further analysis of existing poems to see which techniques might be most promising.

Third, we could step back and ask ourselves what goals we are working towards with TwitSonnet. A different methodology might serve those goals better. For example, if our goal is to teach a computer to identify poetic lines, we might consider using source text richer in poetic style and technique than Twitter. If our goal is to entertain with amusing poetic summaries of news events, we might ask if the present project is the best way to do that. In particular, it is notable that in both this and the previous study, Twitter's informality and conventions such as hashtags were offputting to many participants. These may be aspects of Twitter which make it inherently more difficult as a repository for poetic speech. To verify this interpretation, one option would be to "clean" gathered tweets of hashtags, typos, and other traits that bothered the non-expert raters, before running the study again.

In all cases, a negative result like this one points to a need to reassess and change some aspects of our project, to a greater or lesser degree, so that it fits more precisely with our actual research goals.

Conclusion

While negative results can be discouraging, this result gives us information which is useful for the further development of TwitSonnet and related projects, and which we might not have obtained if we had not performed a falsifiable evaluation. We learned in the previous study that selection of tweets based on specific criteria can indeed produce superior poetry to arbitrary selection. However, we could not show using falsifiable methods that the current method of tweet selection achieved this. We have therefore learned that we should be more careful in the future about exactly how lines for a found poem are selected and what, if anything, this selection contributes to the output. As always, empirical testing is needed so as to ensure that tweet selection, or any other component of a creative system's process, works as intended.

References

[2012] Bhatnagar, R. 2012. Pentametron. `http://pentametron.com/`.

[2014] Bown, O. 2014. Empirically grounding the evaluation of creative systems: incorporating interaction design. In *Proceedings of the Fifth International Conference on Computational Creativity*, 112–119.

[2008] Colton, S. 2008. Creativity versus the perception of creativity in computational systems. In *AAAI spring symposium: creative intelligent systems*, volume 8.

[2012] Galanter, P. 2012. Computational aesthetic evaluation: past and future. In *Computers and Creativity*. Berlin: Springer. 255–293.

[2009] Gheorge, A. 2009. The longest poem in the world. `http://www.longestpoemintheworld.com/`.

[2013] Hartlová, E., and Nack, F. 2013. Mobile social poetry with tweets.

[2010] Hirjee, H., and Brown, D. G. 2010. Rhyme analyzer: An analysis tool for rap lyrics. In *Proceedings of the 11th International Society for Music Information Retrieval Conference*.

[2015] Jordanous, A. 2015. Four pppperspectives on computational creativity. In *The AISB15's 2nd International Symposium on Computational Creativity (CC2015)*, 16.

[2012] Kao, J., and Jurafsky, D. 2012. A computational analysis of style, affect, and imagery in contemporary poetry. In *NAACL Workshop on Computational Linguistics for Literature*, 8–17.

[2008] Kaufman, J. C.; Baer, J.; Cole, J. C.; and Sexton, J. D. 2008. A comparison of expert and nonexpert raters using the consensual assessment technique. *Creativity Research Journal* 20(2):171–178.

[2015a] Lamb, C.; Brown, D. G.; and Clarke, C. L. 2015a. Human competence in creativity evaluation. In *Proceedings of the Sixth International Conference on Computational Creativity June*, 102.

[2015b] Lamb, C. E.; Brown, D. G.; and Clarke, C. L. 2015b. Can human assistance improve a computational poet? *Proceedings of Bridges 2015: Mathematics, Music, Art, Architecture, Culture* 37–44.

[2016a] Lamb, C.; Brown, D. G.; and Clarke, C. L. 2016a. Evaluating digital poetry: Insights from the cat. In *Seventh International Conference on Computational Creativity*.

[2016b] Lamb, C.; Brown, D. G.; and Clarke, C. L. 2016b. A taxonomy of generative poetry techniques. In *Proceedings of Bridges 2016: Mathematics, Music, Art, Architecture, Culture*.

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

[2015] Mohammad, S. M., and Kiritchenko, S. 2015. Using hashtags to capture fine emotion categories from tweets. *Computational Intelligence* 31(2):301–326. Full lexicon available at `http://saifmohammad.com/WebPages/lexicons.html`.

[2002] Pearce, M.; Meredith, D.; and Wiggins, G. 2002. Motivations and methodologies for automation of the compositional process. *Musicae Scientiae* 6(2):119–147.

[1990] Provalis. 1990. Regressive imagery dictionary. `https://provalisresearch.com/products/content-analysis-software/wordstat-dictionary/regressive-imagery-dictionary/`.

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

[2016] Tweet Archivist. 2016. Tweet archivist. `http://tweetarchivist.com/`.

[2015] van der Velde, F.; Wolf, R. A.; Schmettow, M.; and Nazareth, D. S. 2015. A semantic map for evaluating creativity. In *Proceedings of the Sixth International Conference on Computational Creativity June*, 94.

[1998] Weide, R. L. 1998. The cmu pronouncing dictionary. `http://www.speech.cs.cmu.edu/cgibin/cmudict`.