Progress Report

Affect in the Poetry of Emily Dickinson

Trevor Kearns, Ian Torres

CS 585, Fall 2017

Dataset

We are working with the entire set of 1,775 poems written by Emily Dickinson as published in the 1955 variorium edition of her poetry, edited by Thomas H. Johnson [1]. Dickinson's work is difficult for many reasons, not least of which is the variations with which it has been published in the nearly century and a half after her death. Most of her poems were discovered by her sister after she died, tied together with ribbons in little packets that Dickinson referred to as "fascicles." While she published only a handful of poems during her lifetime, and shared many more in private letters with friends and family, the complete set of her poems only became generally available with the Johnson edition in the mid-1950s. Several shorter editions of several hundred poems apiece were published before then, allowing Dickinson's work to begin influencing other writers, but these editions presented a distorted picture of her work; their editors added titles (with very rare exceptions, Dickinson did not title her poems), normalized her eccentric lineation and punctuation (Dickinson is almost unique among English writers for her reliance on dashes in lieu of regular punctuation and idiosyncratic capitalization of common nouns), and arranged her poems in thematic groups that Dickinson clearly had not intended for the presentation of her work.

Johnson's variorium edition was the first to restore her original stylistic choices, strip away the added titles, and number the poems according to Johnson's best judgment of their chronological composition. This ordering will forever be unknown with absolute precision since Dickinson did not track the date of writing for all poems, and scholars, including those who disagree with Johnson's numbering, must all base their work on data such as handwriting style, correlation with life events, and dated revisions, along with a certain degree of guesswork. Rather than venture into this controversy, we are simply assuming that Johnson's numbering is broadly accurate over the span of Dickinson's life.

As it is all in the public domain, Dickinson's work was easy to obtain, but more difficult to prepare for NLP. The digital version of the Johnson edition is riddled with inaccuracies, including words normalized incorrectly by a spell-checker and poems run together without intervening numbers indicating a new poem has started (a group of ten such poems being the most egregious example). We compared the digital version of each poem with a correctly formatted PDF in order to correct these problems. Also, in the cases where more than one version of a poem is included in the Johnson edition, we have opted to only include the first version in our dataset for simplicity's sake. Finally, for baseline testing, we removed all punctuation aside from apostrophes (used mainly in the usual contractions such

as *you're* and *I'm* as well as in now-archaic expressions like *'twas* and *e'en*) and converted poems to lowercase.

Along with the poetic corpus, we used the SentiWordNet 3.0 lexicon [2], which provides negative and positive values for the WordNet corpus (117,000 synsets), to create affect scores for Dickinson's poems. The poetic corpus of 1,775 poems contains 10,933 unique tokens and 94,424 total tokens. We designated 300 poems drawn from a random uniform distribution as a test set, and we read and annotated all of these with primary and secondary thematic tags. The tag set at this point consists of the following: {'insects', 'connection', 'death', 'mother', 'madness', 'art', 'mind', 'falsehood', 'wonder', 'family', 'scorn', 'privacy', 'determination', 'spring', 'love', 'frustration', 'hesitation', 'time', 'sadness', 'loss', 'fame', 'seasons', 'life', 'sleep', 'disconnection', 'freedom', 'happiness', 'mystery', 'woman', 'salvation', 'imagination', 'fear', 'beauty', 'trees', 'peace', 'marriage', 'weather', 'sickness', 'friendship', 'suffering', 'work', 'winter', 'smallness', 'children', 'surprise', 'summer', 'flowers', 'birds', 'nature', 'science', 'memory', 'grace', 'faith', 'autumn'}. This tagset will be refined (possibly condensed), and we plan to review the tags for the test set.

Baseline Affect Scores

We originally planned to assign affect scores (AS) to each line in a poem as well as to each poem overall, calculating the latter by combining the affect score for a sentence in some way with the scores of the lines over which it broke. This would be an attempt to quantify the affective effect of the tension that exists naturally between individual verses and the larger syntactic units in which they participate, an effect that poets control and modulate as part of their essential techniques. However, two difficulties made this plan unworkable: often, Dickinson writes in a way that leaves the relation between line and syntactic unit ambiguous (so that a reader must interpret arbitrarily where sentences might begin or end), and the mathematical relation between line affect and sentence affect is inherently undefined, so that any combination of the two would again be founded on arbitrary assumptions. We have also abandoned our original ternary approach to affect classification since a "neutral" category could indiscriminately represent both low-affect poems and poems with high affect balanced equally between negative and positive scores, and opening a third "neutral" category would introduce the problem of drawing threshold values from arbitrary choices.

Furthermore, our original intent was to make use of a dimensional representation of affect in order to capture more of the breadth of emotional nuance inherent to poetry, but we were unable to acquire the code necessary for a dimensional representation. Doing so would allow for a much finer-grained analysis of affect, one unfortunately beyond the scope of our project this semester.

To overcome all these difficulties, we opted instead to use the positive and negative sentiment scores in SentiWordNet 3.0 for overall poem scores. For the baseline calculations, we did not filter tokens in the poetry by synset, but did a direct lookup, subtracting the negative from the positive score for each matched token, then averaging over two values to calculate two different affect scores: a “stripped” score (averaged over the total tokens matched) and a “pure” score (averaged over total word count for the poem). These can be thought of as affect densities for each poem; the stripped score is more useful for comparing the affective cores of poems (within this corpus or against another poet’s corpus), while the pure score offers more information about the typical style of an individual poet (for example, Dickinson’s style is notoriously compressed, packing a lot of meaning into a small number of words and often omitting elementary syntactic words, thus forcing readers to digest her work slowly and with difficulty). We hope that in the future, this pure score will help us build a classifier to predict whether a poem was written by Dickinson or not. These scores do not currently generate different predictions for any given poem, but they might once we incorporate several refinements to the calculation process (see below).

POEM #	HUMAN 1	HUMAN 2	AS STRIPPED	AS PURE
120	-	+	-	-
151	-	-	-	-
199	-	-	-	-
227	+	+	+	+
237	-	+	+	+
245	-	-	+	+
269	-	-	+	+
287	-	-	-	-
294	+	+	-	-
314	-	-	-	-
340	+	+	-	-
352	+	+	+	+
354	-	-	-	-
382	-	-	+	+
393	-	-	-	-
401	-	-	-	-
411	-	-	-	-
455	+	+	-	-
467	-	+	-	-
471	-	+	-	-
505	-	+	-	-
506	+	+	-	-
566	-	-	-	-
570	-	-	+	+
581	-	-	+	+
669	-	-	+	+
691	-	+	-	-
702	-	+	-	-
841	-	+	+	+
919	+	+	-	-
1016	+	+	-	-
1028	+	+	-	-
1041	+	+	+	+
1122	+	+	+	+
1145	+	+	-	-
1299	-	-	-	-
1328	-	-	-	-
1458	-	-	-	-
1516	+	+	-	-
1773	-	-	-	-

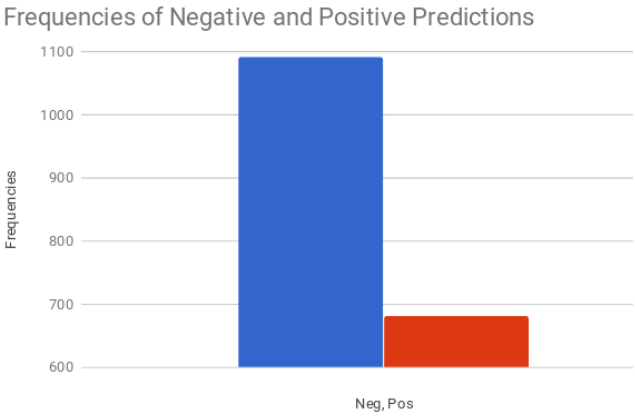
Table 1: Comparison of Human Affect and AS Baseline Predictions. Green indicates complete agreement, yellow mixed results between human and algorithm, and red disagreement between human and algorithm. Blue poems are particularly ambiguous in terms of affect.

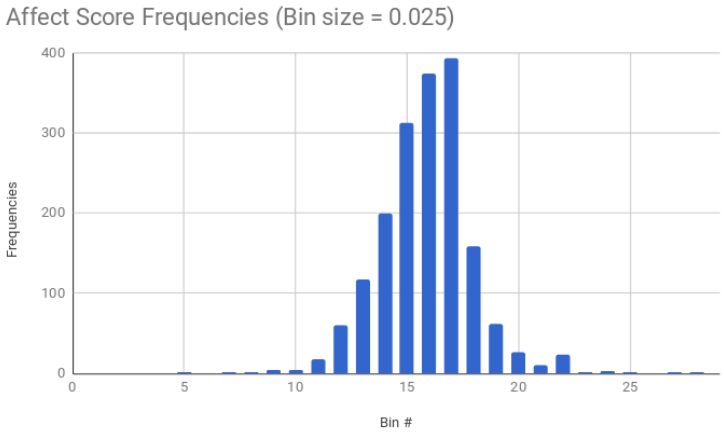
To test the validity of this baseline, we chose 40 poems at random from our test set of 300 (also randomly chosen uniformly from across the corpus). We both read and predicted affect for these poems independently, then compared our results with the baseline counting. We agreed 80% of the time (32/40 poems), a figure that rises to 90% if we remove divergent readings for particularly ambiguous poems (120, 467, 471, and 702). Given the subjectivity with which poetry must be interpreted in general, and the difficulties particular to Dickinson's work, we consider these to be acceptable agreement rates, and worth aiming for as we refine our representation of affect in these poems.

As seen in Table 1, the baseline representation is on par with random prediction (53.1% agreement rate with poems on which we agreed, 17/32; discounting ambiguous poems 269, 294 and 340, this rises to 17/29, or 58.6% agreement), so clearly there is much room for improvement in how we calculate affect scores. First, however, we used these baseline scores to analyze the corpus.

Baseline Analysis of Corpus

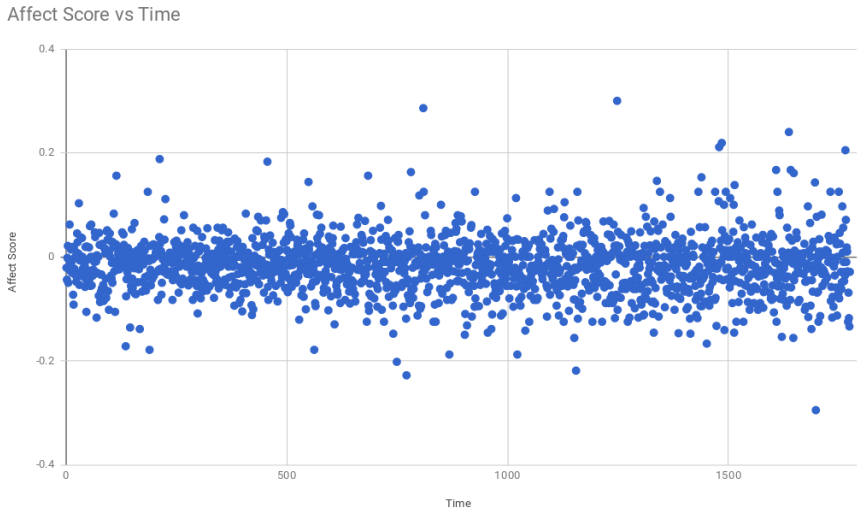
To get an overview of Dickinson's affective predilections, we generated histograms of simple positive/negative scores (pos ≥ 0, neg < 0) and bins with size 0.025. Stripped AS scores were used.





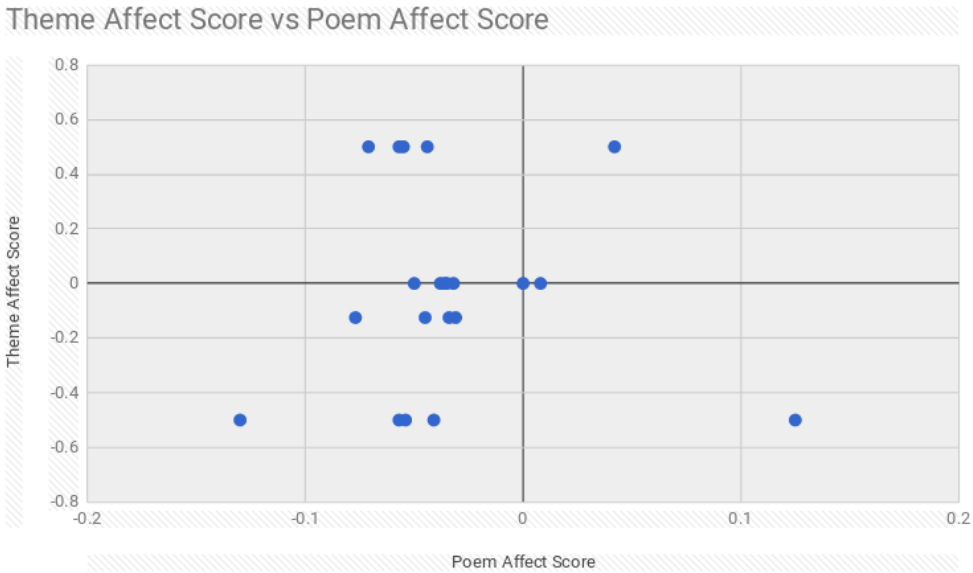
The three highest-scoring bins (15, 16, and 17) have the following respective ranges: [-0.05, -0.025), [-0.025, 0), [0, 0.025]. The average stripped AS for the corpus is -0.01675, while the average pure AS is -0.00891. This slight preference for negative affect correlates broadly with favored themes sampled from the 300-poem test set: the top five themes by summing frequencies of primary and secondary instances are *wonder* (54 total), *death* (52), *loss* (41), *suffering* (34), and *nature* (34). Note that *wonder* here incorporates both ecstatic and overawed responses to phenomena, so it is an ambivalent category from an affective standpoint, while *death*, *loss*, and *suffering* are exclusively negative.

Also, since these poems have been numbered by hypothesized chronology, we are able to check if Dickinson's affective preferences underwent significant shifts over time:



At this point, it does not appear so, but improved AS calculations might reveal more on this question.

Finally, we are able to examine the relationship between baseline AS and various themes. Since we are not filtering by synset yet, we are only able to compare the AS of tags that appear in SentiWordNet as entries unto themselves, so this type of analysis is not fully available to us at present, although it will be once we implement improvements (see below). Here we have plotted several poems with these tags: *determination* (AS 0), *grace* (0.5), *death* (-0.5), *salvation* (-0.125), *family* (0).



Eventually, we will produce cluster graphs of this kind for the entire corpus, or for poems with tags that fall into natural synsets (for example, *nature*, *weather, insects*, *birds*, and *trees* can all be subsumed under *nature*). This will allow us to make observations about any interesting relationships between themes and the AS of poems that participate in them. We might also combine the affect scores of the

two tags in order to produce the Theme Affect Score; we are still working out how best to represent this axis.

Planned Improvements and Timeline

We propose two approaches to improving AS calculations: one is a refinement of simple adding of SentiWordNet scores, while the other is more involved and, we think, more promising in terms of accuracy.

We will first improve the baseline method by implementing synset filtering for tokens when looking up SentiWordNet scores. This itself will incorporate bigram lookup and POS parsing since the lexicon supports particles and includes separate synsets for different word senses. We will correctly parse contractions like *'twas* during this process as well. This "baseline plus" method should generate more AS values for each poem, since more tokens will be identified in the lexicon through their synsets, yielding a more accurate representation of overall affect based on SentiWordNet. We will implement "baseline plus" before the end of November.

The more intriguing method we will also implement reduces our reliance on SentiWordNet, whose limitations as a tool for analyzing older forms of English should be obvious (it is based on WordNet, itself heavily biased toward modern usage, from which its sentiment scores have also been generated, thus failing to account for shifts in affect for many words since Dickinson's writing lifespan in the mid-to-late 19th century). We plan to calculate PMI scores for all words in Dickinson's corpus against seed lists generated from the intersection of her most frequently used words and SentiWordNet, sorted by most positive/negative. This way, the basis of the PMI score will be affective in nature, and the PMI scores will represent associations among tokens within her corpus (rather than the much more tenuous relations to the modern SentiWordNet lexicon). AS calculations will be made using these PMI scores. We will implement this "affective PMI" method by the second week of December.

We will compare the accuracies of these two new methods to our baseline and fine-tune them. Finally, we plan to build a classifier that can use these three methods to make predictions about affect in the rest of Dickinson's corpus for further confirmation of their comparative accuracy. This will require additional human annotation of a number of poems for affect. We plan to complete this portion of the project by the second week of December as well.

If we have time, we will try to use the affective PMI scores to make predictions about poems published during Dickinson's lifetime, working off the assumption that she would have shared a general sense of the affective valence of her vocabulary with her poetic peers. This will be done no sooner than the third week of December, but it may not take much time since we already have a dataset of 19th-century poems.

Works Cited

[1] E. Dickinson, *Poems of Emily Dickinson; Including Variant Readings Critically Compared with All Known Manuscripts,* ed. Thomas H. Johnson, Harvard UP, Cambridge, 1955, Vols. I-III.

[2] S. Baccianella, A. Esuli, and F. Sebastiani, "SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining," In *LREC*, vol. 10, pp. 2200-2204. 2010.

Github repository: E-Dickinson_NLP