Instructions for *ACL Proceedings

Eesha Barua
Student / University of Illinois
email@domain

Second Author
Also Eesha Barua
not sure how to only have one author
yyup

Abstract

In this paper, I am using word2vec word embeddings to map semantic relationships between different emotion words in subgenres of poetry, and comparing the results.

1 Introduction

Poetry is one of the last frontiers of human language that computers fail to understand. Previously, I have used supervised learning in an attempt to classify emotions in lines of poetry, which bring significant drawbacks because labels are susceptible to the unique bias and perception of the few annotators in annotated corpora. To remedy this problem, I then attempted an unsupervised learning approach. Poetry as an art form varies wildly and contains many registers, irregular word formations, and other factors that contribute to its unstandard language use. As a result, strong clusters were unable to model high-level topics in poetry such as emotion. In this paper, I will be performing a comparison among two embedding spaces. I will be comparing a lexicon of emotion words as used in different subgenres of poetry.

2 Methodology

2.1 Building the Corpus

I utilized the Project Gutenberg corpus to build corpora for this paper. I decided ultimately that sourcing text directly from Project Gutenberg would be the best method because I could parse large amounts of data directly. Many other online poetry corpora do not sufficiently annotate the genre of each poem, and I thought Project Gutenberg's subject tagging would be the most credible for this project as it is the most widely used. On top of that, the previous corpora I had used from, both Supervised Learning and Unsupervised Learning, were sourced from a random selection of Project Gutenberg poetry, ranging from Middle English to Children's diction to epic poetry. In this paper, by building corpora out of specific Project Gutenberg genres, I could run experiments where the style and form of the language were more consistent for each subcorpus.

2.1.1 Data Collection and Aggregation

First, I downloaded the metadata of the Project Gutenberg corpus. I extracted the 'Subject' tag from all Project Gutenberg texts and filtered out all the tags that included Poetry. From there, I filtered out all texts written in exclusive English. Because the results involve qualitative analysis and due to a lack of corpora in the languages I can understand, I exclusively analyzed English poetry. In addition, word2vec is pretrained on the Google News dataset, which is an English corpus; as a result, I decided English poetry would produce the most comprehensive results. I grouped the Poetry text metadata by subject and saved the metadata for the top 10 most populous poetry classes (see Table).

Table 1: Poetry Data

Title	**Text**	**Lines**
Children's poetry	210	63001
American poetry – 19th century	99	1205
American poetry – 20th century	90	58
American poetry	202	1187
English poetry – 19th century	120	55
English poetry – 20th century	85	85
English Poetry	209	24872
Epic poetry	109	793412
Poetry	64	63076
1914-1918 – Poetry	64	115

Then, I used the Project Gutenberg API to download the corpus .txt files for each PGID found in the .csv file of Project Gutenberg metadata. For each of the above poetry subgenres, I used file parsing to extract the poems from associated text files and stored their poetry lines into a corresponding dataframe. I experimented with stanzas of poetry instead of lines of poetry and realized that stanzas did not perform nearly as well in unsupervised learning, and therefore may not be as useful in semi-supervised learning as well.

Table 2: 20th Century American Poetry dataframe

ID	Title	Subjects
160	The Second Book of ...	20th American
184	The Song of the Stone Wall	20th American
...	...	...
849	The Path to Home	20th American

2.1.2 Cleaning the Data

The corresponding dataframes with lines of poetry for each subgenre underwent several stages of cleaning. First, I tokenized the data and removed all lines of poetry with less than 10 lines. This cleaned the lines of the text files that included poem titles, as well shorter lines that wouldn't provide enough context. Next, I removed all special characters and converted all words to lowercase. I decided that losing case information about words would not get rid of any important semantic information that pertains to emotions.

I removed duplicate rows, as many texts had been cross-listed under two tags and therefore their lines of poetry showed up in each dataframe twice. Finally, as a final pre-processing step before training the word2vec model, I ignored all English stopwords.

2.2 Word Embeddings

I chose ten prototypical emotions (Happiness, Sadness, Anger, Fear, Surprise, Disgust, Trust, Anticipation, Joy, and Love) and generated 10 seed words for each (see Appendix). Then, to further expand this list and encompass as many registers as I could, I used WordNet to find synonyms for each of the original 50 seed words and compile a large dictionary of 65 emotion words.

After iterating over each emotion word for each subgenre, I generated a word2vec vectorized representation of each emotion that was present in that subgenre's corresponding corpus. For all subgenres, I parameterized the word2vec model with a vector size of 100 dimensions, window of 5, and 4 workers.

Finally, I used TSNE to reduce the 100-dimensional vectors to 2D space so that I could map the relative distances between each emotion word present in the corpus.

3 Results and Future Work

The following diagrams show the mappings for vectorized representations of emotion words for each subgenre of poetry (Children's, Poetry, American, English, Epic, 19th Century American, 20th Century American, 19th Century English, 20th Century English, and 1914 - 1918).

Notably, I noticed that the distribution of words across space was not uniform , and that antonyms were often grouped together. For example, in American Poetry, contentment and discomfort had the very similar Component 1 coordinates.

It was difficult to compare between embedding spaces simply because the emotion lexicon I built was biased towards today's vernacular, and as a result older poetry corpora (i.e. 19th Century English Poetry, 1914 - 1918 Poetry, and English poetry in general) had marginally less word vectors mappings. This could be reflective of the fact that word2vec was trained on the Google News dataset, which is a relatively modern American English corpus.

In addition, the larger subcorpora (i.e. Children's Poetry, Poetry, and Epic Poetry) were too large for my computer and my Jupyter Notebook kernel shut down. In the future, I would want to find a way to sample from these larger subcorpora or utilize more robust embedding models such as fastText.

Lastly, I notice that 19th century American poetry seems to form a close-knit cluster as opposed to American Poetry, which feels more distributed. This makes me wonder if the relevant current-day vocabulary used to describe emotions is only a subset of emotions used in the 20th century.

Overall, in the future I would like to experiment with a more diverse set of emotion seed words, with different embedding models, and with more computing power. The results of this experiment are proof that word embedding techniques can convey a great deal about the associations between different emotions in poetry.

3.1 Citations

Abdul-Mageed, M. and Ungar, L. (2017). EmoNet: Fine-grained emotion detection with gated recurrent neural networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages

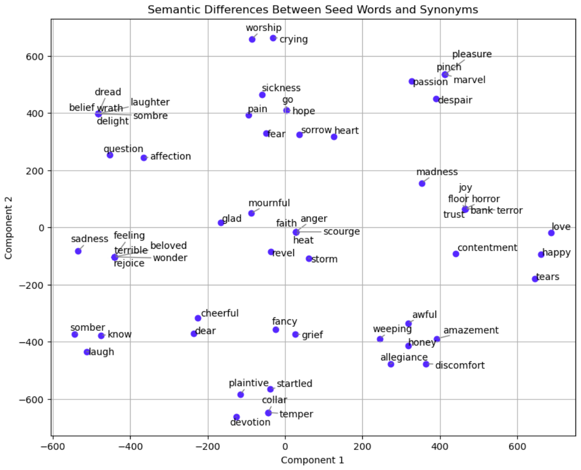

Figure 1: Mapping of Emotions in American Poetry

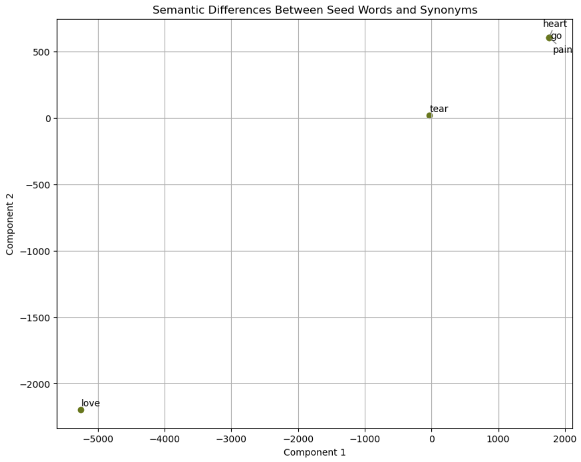

Figure 4: Mappings of Emotions in 19th Century English Poetry

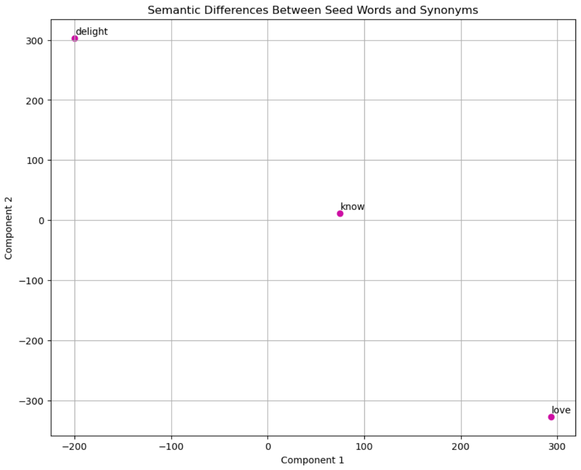

Figure 2: Mapping of Emotions in English Poetry

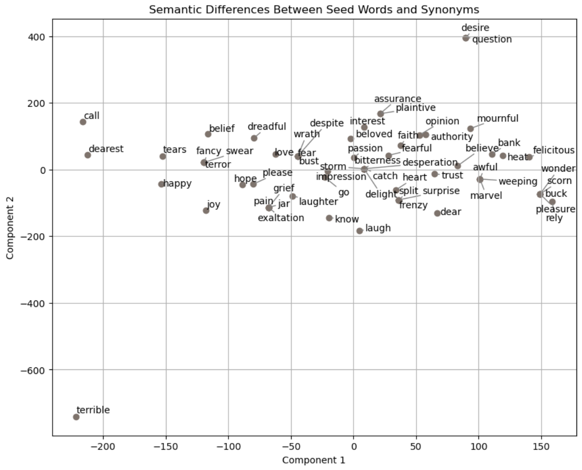

Figure 5: Mappings of Emotion in 19th Century American Poetry

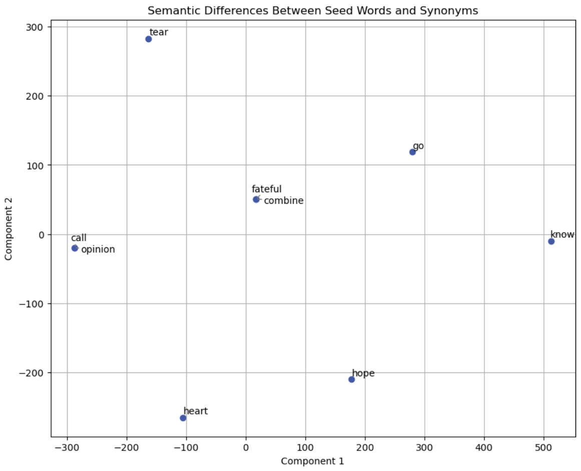

Figure 3: Mappings of Emotions in 1914 - 1918 Poetry

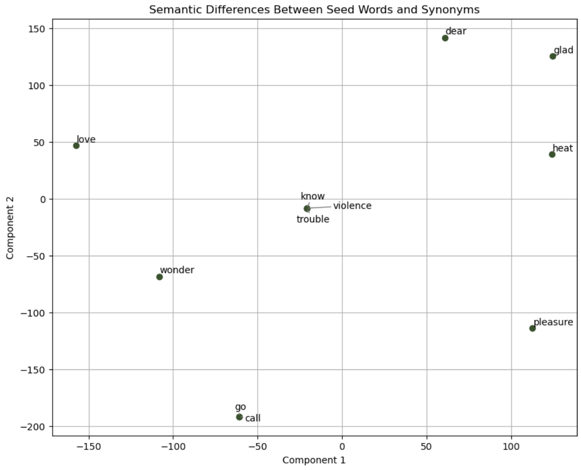

Figure 6: Mappings of Emotion in 20th Century American Poetry

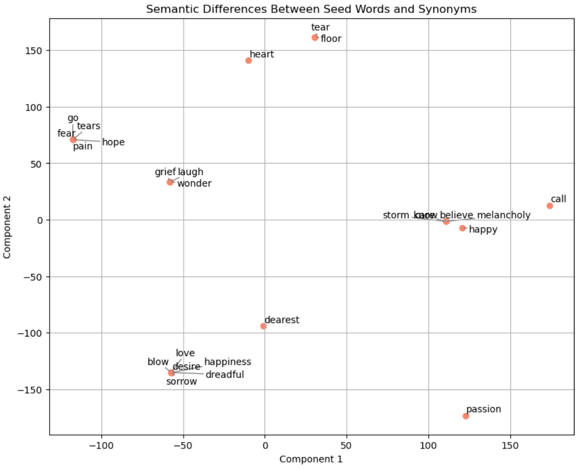

Figure 7: Mappings of Emotion in 20th Century English Poetry

718–728, Vancouver, CA.

Acknowledgements

Thank you to Riya Cherlakola and Danny Kim for their thoughts and prayers.

4 Appendices

Table 3: Seed Words for Each Emotion Category

Feeling	**Seed Words**
Happiness	joy, delight, happy, cheerful, laughter, love, contentment, bliss, euphoria, excitement
Sadness	sadness, sorrow, grief, despair, melancholy, unhappiness, tears, loneliness, heartache, mournful
Anger	anger, rage, fury, wrath, irritation, frustration, resentment, hostility, indignation, annoyance
Fear	fear, anxiety, terror, fright, panic, worry, dread, apprehension, unease, phobia
Surprise	surprise, astonishment, amazement, wonder, shock, disbelief, awe, startle, unexpected, startled
Disgust	disgust, revulsion, repulsion, nausea, aversion, loathing, contempt, abhorrence, repugnance, detest
Trust	trust, confidence, faith, reliance, belief, assurance, loyalty, credibility, dependability, honesty
Anticipation	anticipation, expectation, excitement, eagerness, hope, optimism, suspense, foreboding, expectation, expectancy
Joy	joy, happiness, pleasure, delight, bliss, ecstasy, elation, jubilation, euphoria, exuberance
Love	love, affection, adoration, passion, romance, fondness, devotion, intimacy, warmth, tenderness