Technology Usage in Poems

How Poetry is Analyzed and Generated by Technology

Aaron Chao
University of California, Santa Cruz
Santa Cruz, CA
aalichao@ucsc.edu

***Abstract*—This document will examine recent approaches to analysis and generation of poetry presented in research papers.**

***Keywords*—BiDirectional Long Short Term Memory, term frequency-inverse document frequency, Large Language Model, dynamic time warping framework, Support Vector Machine Classifier.**

I. Introduction

Poetry provides a unique challenge for computer comprehension, especially given its preference towards metaphors and dramatization of the human experience. This is especially so given how differently structured and ordered poetry is between pieces and styles. Besides the standard structured poetry based on meter or lack thereof, all poems can be seen as fitting upon a spectrum between purely lyric- or symbolic- and narrative poems [1]. While narrative poems already pose a challenge, the prevalence of metaphors and details associated with an emotional center in lyric poems would pose difficult for a computer to comprehend, especially given that many humans also have a hard time understanding poetry in general.

II. Human Generation of Poetry

The variance of poetry's structure makes it inherently difficult for computer comprehension. While a vast majority of humans can detect the nuances and effects of literary devices through reading a poem out loud and hearing their effects, computers must break up words and go through rule-based or other scheme-based comprehension methods.

Besides structured poetry, with a consistent rhyming scheme and meter, free-verse poetry has become far more commonplace than in the past. Meter can be thought of how words and phrases are stressed, where the common measurement of measure, feet, is usually known as a unit of stressed and unstressed syllables, with common combinations given names, such as the iamb, an initial unstressed syllable with one following stressed syllable, reminiscent of a heartbeat. While free-verse poetry allows more creative freedom, especially with rhythm and structure, these creative differences make poetry even more inconsistent, and therefore difficult to comprehend or create through computer programs. Rhythm, after all, is not solely created by meter, but also consists of elements such as pacing, which can be dictated by aspects such as repetition or punctuation, as well as how stanzas are broken up.

However, despite all of the nuances and forms poetry undertakes, poetry's main purpose is to dramatize an idea in unique ways compared to other art forms. Nowhere is this more present than the use of metaphors to build connections between ideas and senses that would make little sense when compared directly. The ability to craft novel connections through metaphors, is another feature of poems that is difficult to completely comprehend by humans, much less machines. Furthermore, the building out of these metaphors and connections between subjects within each line is still hard to generate with technology and computer programs.

The purpose behind poetry may also vary. While poems and other forms of literature usually attempt to make an impact on the reader, some poems may not attempt to make a clear message or theme, and instead have goals such as confusion or surrealism [1].

III. TECHNOLOGIES EXAMINED IN THIS PAPER

A. Artificial Intelligence / Neural Networks

Current approaches to evaluating poetry tend to utilize neural networks to classify and evaluate poetry on metrics such as dominant emotion. An interesting approach for this was the use of a layer that utilized BiDirectional Long Short Term Memory, which allows for input to be evaluated from beginning to end and end to beginning concurrently, which adds future context to the evaluation of lines.

B. Statistical Approaches

Another interesting application on evaluating poetry regarding text reuse between different poets, specifically with famed William Butler Yeats and 5 of his poetic influences, was a recreation of the Tesserae Project's use of TF-IDF, term frequency-inverse weighting scheme. This weighting scheme assigns the importance of words on a document based on frequency. The researchers then used these weights to assign a weighted score to common word-pairs within different documents, or poems in this case. Using these scores to common word pairs, poems were assigned a score based on the perceived level of text reuse based on these metrics.

C. Large Language Models

Researchers developed a model called CoPoet that sought to write poetry alongside a human partner to avoid the tendency of Large Language Models to suffer from long-term coherence. Researchers used a sequence-to-sequence model based on instruction-output pairs generated by rules and synthesis of creative texts respectively. The model was then evaluated through human opinion on their preference on the poetry generated by this model and poetry generated by other LLMs.

D. Machine Translation

Translations of poetry are difficult due to the necessity to preserve the form and meaning of the poem from another language. Oftentimes, translations by humans come with compromises to either meaning or form, such as rhythm or rhymes. The ability to incorporate machine translations with models that can weigh the benefits of multiple translations as they are created, is something that is very tempting to use for the purpose of poetic translations. Whereas human poets that translate poetry with an emphasis on form tend to first come up with rhyming words, then put other words into place to make up the meter of the poem, machine translations can put a score associated with translations, and decide which branch to go with when it comes to decisions on which hypothesis is better [7].

E. Machine Learning on Audio Recordings

Support Vector Machine Classifiers were used on melodic pitch files computed from MELODIA, a Sonic Visualizer Plugin, using a dynamic time warping framework [6]. Essentially, machine learning was used on visualized audio files using a framework that takes into account timing. Although the sample set that the research was performed on, 100 poems, with 10 for each Malayalam meter and female and male speakers each reading half of the poems per meter, only 50 poems were used for evaluation of the methodology [6]. Overall, the algorithm had an accuracy of around 86%, although different meters had different accuracies [6]. Despite the accuracy seeming low, this result is fairly impressive given how different speakers can meterize poetry differently, as well as noise in the dataset given to the classifier.

IV. Research

A. Emotion Evaluation

Research regarding usage of computing to analyze poetry was conducted by researchers of Zayed University and Gomal University regarding emotional classification through the use of deep learning. The researchers begin by discussing limitations in their program's analysis, listing "limited size of the benchmark poetry dataset, lack of an extended set of emotions and use of classical machine learning classifiers" as reasons that make analysis of poetry difficult [4]. These factors are fundamental ones that currently do not have good solutions. Choosing and pruning the poems within the benchmark poetry dataset is arguably the most important step in any classification task. Furthermore, picking which subsets of emotions as output categories for the model is highly important for how well the model will perform. In this case, the output labels were made up of emotions defined as the set of "alone, anger, courage, fear, hate, hope, joy, love, nature, peace, sadness, suicide and surprise" which did not appear to intersect too much, although different poems may have them overlap in their themes [4]. The researchers also took note of other relevant approaches to their topic already presented in the academic community, including approaches in machine learning, deep learning, and an alternate approach of 'Navarasa,' an ontology that analyzes sentiments on short stories in poetry of emotion [through] polarity in each word" [4]. The deep learning method utilized by the researches uses a variety of layers, where the majority of the transformative work from data to label is done in a BiDirectional Long Short Term Memory, which allows input to be viewed from past-to-future contexts or future-to-past contexts [4]. Essentially this is akin to reading a paragraph a sentence at a time regularly, as well as reading the paragraph backwards a sentence at a time, giving future context for your current input, a sentence in the middle of the paragraph. Through their deep learning method along with the usage of BiDirectional Long Short Term Memory, the researchers were able to achieve an accuracy of 84%, precision of 85%, recall of 83%, and a f1 score of 84% [4].

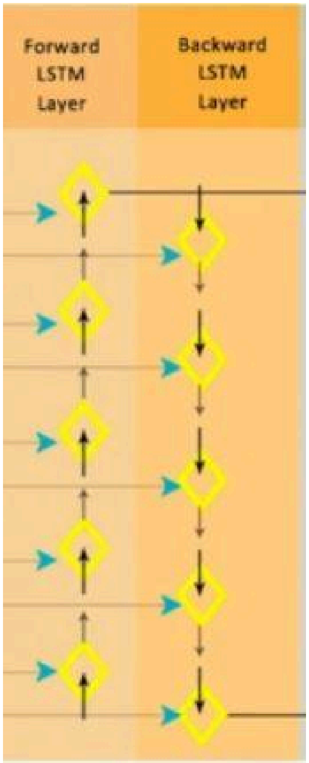


Fig 1. The BiDirectional Long Short Term Memory portion of the model made by researchers. The two orange layers, one being the forward long short term memory layer, and the other the backward component, parse/evaluate the input in different directions. In poetry, this may be akin to evaluating starting from the first line and going to the end as well as evaluating from the last line and going to the first concurrently [4].

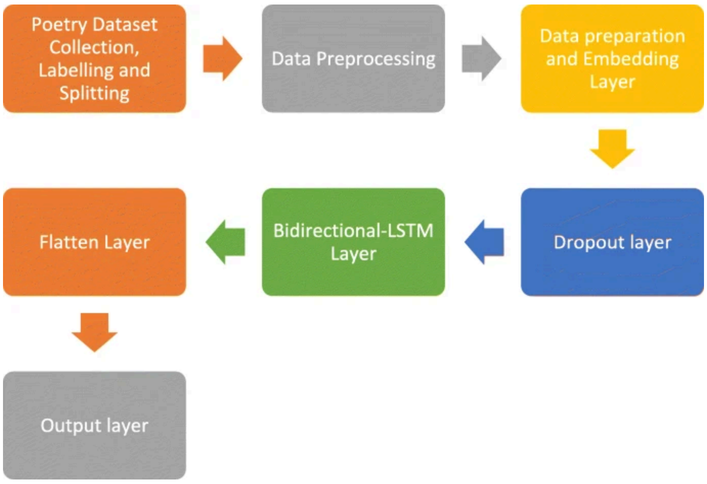
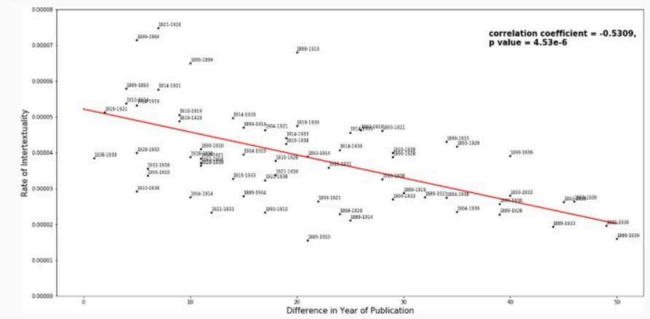

Poetry Dataset Collection, Labelling and Splitting
Data Preprocessing
Data preparation and Embedding Layer
Dropout layer
Bidirectional-LSTM Layer
Flatten Layer
Output layer

Fig 2. The Approach of the Emotional Evaluation Deep Learning Model [4].

B. Text Reuse

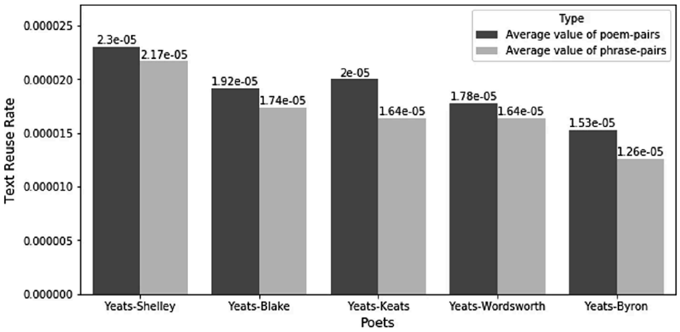
Another subset of analyzing poetry was done by researchers at the University of Illinois Urbana-Champaign regarding text reuse in William Butler Yeats poems as compared to a few of his poetic influences. The study used a term frequency-inverse document frequency, TF-IDF, weighting scheme, which assigns values of importance for words in documents, and compared it to the methods used by the Tesserae Project, a similar project that investigated verbal repetition by Latin poets [5]. Both methods consider instances of text reuse as phrases from the target and source material that share at least 2 words in common. Both methods record the distance between common words, as well as the weight of the words in between those words, to generate a weighted score. Unlike the weighting scheme used by the Tesserae Project, which finds the distance between the two common words with the lowest frequencies first, the researchers of UIUC chose to find distance based on the closest common words. Afterwards, each of Yeats's poems were scored with the average of this weighted score for all poem-pair values. The researchers ultimately came to the conclusion that there was a negative correlation between the publication dates for Yeats's collections of poetry and the text reuse, defined by the average of the weighted score for all poem-pair values [5]. The negative correlation makes sense, as writers tend to develop their own styles and rely less upon their knowledge gained by solely studying previous authors. However, the authors of the research take the conclusion a bit further, believing that the rate of change of the correlation was small, which highlights the gradual change of a writer's voice and diction choice [5]. This conclusion is interesting, as it showcases a trend of creators being prone to gradually changing their styles as opposed to drastic changes.

Fig. 2.



Correlation between the text reuse rate and difference in publications dates of Yeats's collections

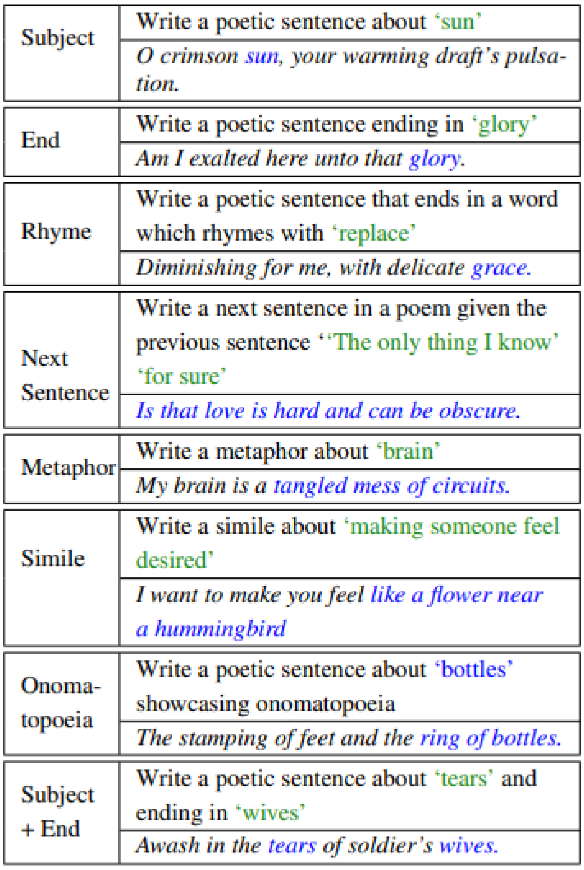
Fig 3. The output data in a graph form by researchers evaluating text reuse by William Butler Yeats. The y-axis is the amount of text reuse compared with 5 poets that were poetic influences on Yeats, and the x-axis is publication dates of Yeat's collections of poetry [4].



Text reuse rate between Yeats and his English Romantic poet influences [4].

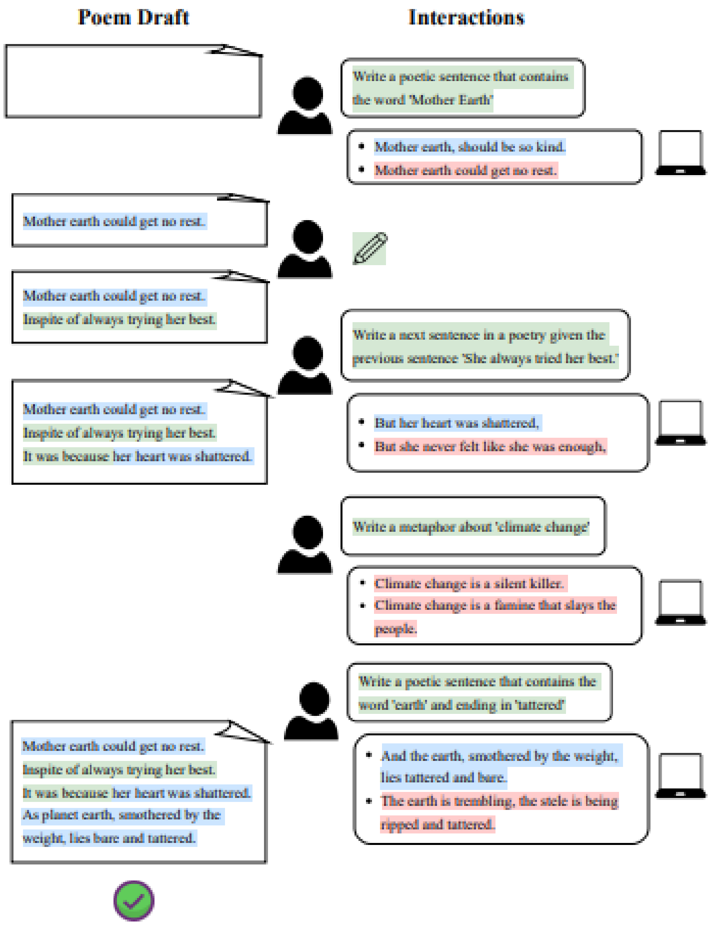
C. Large Language Models

More recently, the commercialization of artificial intelligence, especially for use by the public with art generation such as DALL-E and large language models such as ChatGPT, begs the question of whether such methods can also be applied to poetry, an art that appears to have vague definitions at best. Researchers from Columbia University and New York University attempted to utilize a large language model to create poetry. The researchers acknowledge a major weakness of large language models, that their output usually suffers from long-term coherence, despite being the specifications given to the model. Instead, they present a large language model called CoPoet, which seeks to help in a collaborative writing method between the model and humans. CoPoet was created through a sequence-to-sequence model on instruction-output pairs, where the instructions were created by rules and outputs were synthesized from texts of creative writing [3]. Ultimately, when viewing performance, the researchers found that larger models tended to generalize and perform better, and that CoPoet performed at least 10% better than the 175B version of InstructGPT [3]. Although the researcher's large language model performed significantly better than other large language models, the method of measuring performance and comparison to human poetry appears a bit flawed, as a lot of the measurements are based on human input and human feedback, as well as being based on subjective metrics.



Subject	Write a poetic sentence about 'sun' *O crimson sun, your warming draft's pulsation.*
End	Write a poetic sentence ending in 'glory' *Am I exalted here unto that glory.*
Rhyme	Write a poetic sentence that ends in a word which rhymes with 'replace' *Diminishing for me, with delicate grace.*
Next Sentence	Write a next sentence in a poem given the previous sentence ''The only thing I know' 'for sure' *Is that love is hard and can be obscure.*
Metaphor	Write a metaphor about 'brain' *My brain is a tangled mess of circuits.*
Simile	Write a simile about 'making someone feel desired' *I want to make you feel like a flower near a hummingbird*
Onomatopoeia	Write a poetic sentence about 'bottles' showcasing onomatopoeia *The stamping of feet and the ring of bottles.*
Subject + End	Write a poetic sentence about 'tears' and ending in 'wives' *Awash in the tears of soldier's wives.*

Fig 4. A graphic explaining the instruction by rule used by CoPoet, as well as example outputs [3].



Write a poetic sentence that contains the word 'Mother Earth'

- Mother earth, should be so kind.
- Mother earth could get no rest.

Mother earth could get no rest.

Mother earth could get no rest.
Inspite of always trying her best.

Write a next sentence in a poetry given the previous sentence 'She always tried her best.'

- But her heart was shattered,
- But she never felt like she was enough,

Mother earth could get no rest.
Inspite of always trying her best.
It was because her heart was shattered.

Write a metaphor about 'climate change'

- Climate change is a silent killer.
- Climate change is a famine that slays the people.

Write a poetic sentence that contains the word 'earth' and ending in 'tattered'

- And the earth, smothered by the weight, lies tattered and bare.
- The earth is trembling, the stele is being ripped and tattered.

Mother earth could get no rest.
Inspite of always trying her best.
It was because her heart was shattered.
As planet earth, smothered by the weight, lies bare and tattered.

Fig 5. An example of CoPoet's interactions with a human to create a poem based on Climate Change [3].

D. Machine Translation

Translations of Poetry are extremely difficult for even human poets, who oftentimes must make sacrifices to either the meaning or the structure of the source poem. There are proponents for the greater importance of the meaning or form among translators of poetry. This issue makes it tempting to resort to machine translation of literary works, especially if untranslated thus far. Researchers at Google suggest a pattern-based system that takes into account stress. Stress can be identified through either a pronunciation module of a text-to-speech system or a pronunciation dictionary such as the Carnegie Mellon Pronunciation Dictionary, although the researchers chose a pronunciation dictionary as well as functions to identify rhyme between words [7]. With this phrase-based approach, the researchers trained a baseline French to English poetic translation system and a stress pattern-constrained system to compare the two [7].

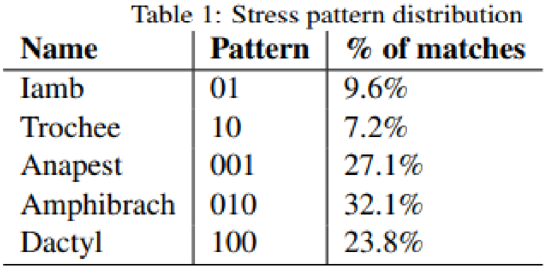


Table 1: Stress pattern distribution

Name	Pattern	% of matches
Iamb	01	9.6%
Trochee	10	7.2%
Anapest	001	27.1%
Amphibrach	010	32.1%
Dactyl	100	23.8%

Fig 6. Identification of Stress Patterns (where 0 is no stress and 1 is stress), within the experiment [7].

Although it would be difficult to identify quality of evaluation in terms of retained meaning, the researchers instead quantified success as two metrics: "percentage of sentences that can be translated while obeying a stress pattern constraint, and the impact of this constraint on BLEU score" [7]. BLEU score, or BiLingual Evaluation Understudy score, is an automatically evaluated percentage that measures the similarity of machine translations to high quality samples. The researchers found that normally, the BLEU score of their stress pattern-constrained system tended to be half of the baseline score, with a baseline score of 35.33 and a score of 18.93 for the stress pattern-constrained system [7]. Furthermore, with the stress pattern-constrained system, sentences were correctly matched 85% of the time, and with allowing 1 stress error, that number reaches 93% [7]. The small difference of 7% between the two proportions above showcase that in a majority of situations, the stress pattern-constraint can be satisfied [7].

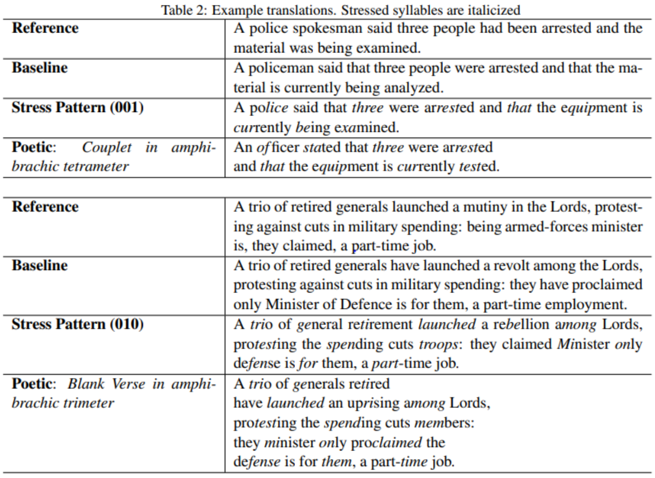


Table 2: Example translations. Stressed syllables are italicized

Reference	A police spokesman said three people had been arrested and the material was being examined.
Baseline	A policeman said that three people were arrested and that the material is currently being analyzed.
Stress Pattern (001)	A po*lice* said that *three* were ar*rest*ed and *that* the e*quip*ment is *cur*rently *be*ing ex*a*mined.
Poetic: *Couplet in amphibrachic tetrameter*	An *of*ficer *sta*ted that *three* were ar*rest*ed and *that* the e*quip*ment is *cur*rently *test*ed.

Reference	A trio of retired generals launched a mutiny in the Lords, protesting against cuts in military spending: being armed-forces minister is, they claimed, a part-time job.
Baseline	A trio of retired generals have launched a revolt among the Lords, protesting against cuts in military spending: they have proclaimed only Minister of Defence is for them, a part-time employment.
Stress Pattern (010)	A *trio* of *ge*neral re*ti*rement *launched* a re*bel*lion a*mong* Lords, pro*test*ing the *spen*ding cuts *troops*: they claimed *Mi*nister *on*ly de*fense* is *for* them, a *part*-time job.
Poetic: *Blank Verse in amphibrachic trimeter*	A *trio* of *ge*nerals re*tir*ed have *launched* an up*ri*sing a*mong* Lords, pro*test*ing the *spen*ding cuts *mem*bers: they *mi*nister *on*ly pro*claimed* the de*fense* is for *them*, a part-*time* job.

Fig 7. An Example of Baseline and Stress Pattern-Constrained output on reference texts [7].

E. Machine Learning

Researchers at the Rajiv Gandhi Institute of Technology used a Support Vector Machine classifier with a dynamic time warping framework to identify poetic meter through visual audio files generated by MELODIA [6].

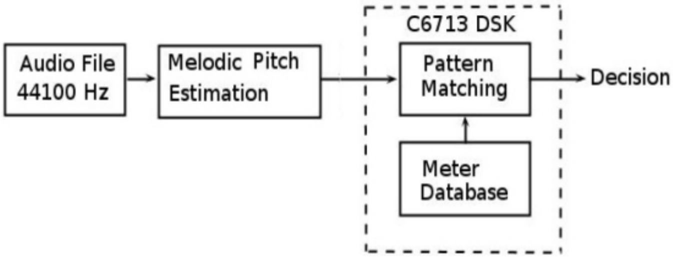


Fig 8. The Approach of the SVM Classifier used by Researchers to classify the meter of the poem in the test set [6].

The identification of poetic meter is fairly difficult, as while many poems have common ways words are broken up into syllables and stressed, different speakers can meterize poems differently. Furthermore, there are far more exceptions with meter due to the creative freedom of poetry, which makes mainstream approaches of using a dictionary with common ways words are broken apart into syllables and stressed, such as the Carnegie Mellon Pronunciation Dictionary, difficult to find perfect accuracy. The use of visualized audio files, as well as time-based comparisons of the distances of pitch contours, is a new and interesting approach to addressing this issue.

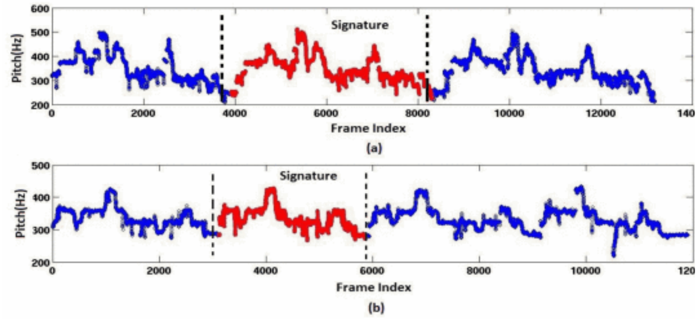


Fig 9. An Example of 2 Extracted Meters, a. dhruthakākali, b. Kēka [6].

V. Challenges

A. *Issues with Identification of Emotion*

A majority of the issues present appear to be discussed by the authors, such as the limited number of emotions present in the set of emotions evaluated for, size of sample set of training data made up of poetry, and that their model only evaluated for one emotion within an entire piece instead of evaluating emotions on smaller intervals such as individual stanzas or lines.

B. *Text Reuse*

Overall, it appeared as though the researcher's statistical approach was relatively solid, especially since the only change from the previous referenced research project was on how weights were calculated based on word importance rather than purely word distance between the 2 common word pairs. The importance of words used for this metric, perhaps should not be based on, or at least only on, the word frequency, but instead also evaluate the importance of words to the messaging behind the literature or document.

C. *Human Input in Evaluation of LLMs*

The evaluation behind the performance of this research was perhaps most problematic. Humans were used to direct the researchers' large language model, CoPoet, on which lines should be kept or revised to craft the poem, with researchers calling the participants "experts" in the subject matter without discussing their abilities or credentials. Furthermore, the evaluation of performance of the model was judged by other experts that again do not have their credentials discussed, which brings up how much can human judgment be used confidently when evaluating art?

D. *Limitations on Output in Machine Translation*

The researchers state that their feature function used in the stress pattern-constrained system cannot avoid hypotheses with impossible rhyming constraints, despite being capable of avoiding hypotheses where length constraints are impossible to implement [7]. As a result, the output may not fit the desired format for the translated poetry, as well as taking more time and effort for the model to finalize its work. Furthermore, as a result of this time and effort, the researchers were not able to provide an online demo, which makes their results difficult to replicate.

E. *Limited Dataset and Similar Meters*

The research presented by Saju and other researchers appears to be fairly limited as the dataset consists of a maximum of 100 Malayalam poems of 10 types of meter [6]. As a result, this research cannot be directly applied to other types of poetry, although its approaches can probably be applied to other poetry as well as other fields, such as music. Further, the publication of this research addressed that it had lower accuracies with certain meters, Kēka being the lowest, as a result of its similarities with other meters, which led to more misclassifications [6]. Despite these issues, this approach of viewing poetry from its vocal features is an interesting approach in the field of computer analysis of poetry.

VI. Conclusion

Although there are minor flaws in a couple of the research conducted in this field, ultimately in evaluating and generating poetry through technology, there are no clear solutions for many of the desired applications. Ultimately, all of this research takes very interesting and novel strides forward regarding research in the field. Despite initial research using these methods seeming relatively basic compared to more developed fields, the use of these approaches regarding art and humanities are interesting developments to witness.

Acknowledgment

The author would like to thank the University of California, Santa Cruz for allowing him to access reference materials within research databases.

References

[1] Orr, Gregory. A Primer for Poets &amp; Readers of Poetry. W.W. Norton &amp; Company, 2018.

[2] Caroline Ardrey, Visualising Voice: Analysing spoken recordings of nineteenth-century French poetry, Digital Scholarship in the Humanities, Volume 35, Issue 4, December 2020, Pages 737–758, https://doi.org/10.1093/llc/fqz073

[3] Chakrabarty, Tuhin, Vishakh Padmakumar, and He He. "Help me write a poem: Instruction Tuning as a Vehicle for Collaborative Poetry Writing." arXiv preprint arXiv:2210.13669 (2022).

[4] Khattak, A., Asghar, M.Z., Khalid, H.A. et al. Emotion classification in poetry text using deep neural network. Multimed Tools Appl 81, 26223–26244 (2022). https://doi.org/10.1007/s11042-022-12902-3.

[5] Shang, W., Underwood, T. (2021). Improving Measures of Text Reuse in English Poetry: A TF–IDF Based Method. In: Toeppe, K., Yan, H., Chu, S.K.W. (eds) Diversity, Divergence, Dialogue. iConference 2021. Lecture Notes in Computer Science(), vol 12645. Springer, Cham. https://doi.org/10.1007/978-3-030-71292-1_36.

[6] Soumya S, Saju S, Rajan R, Sebastian N (2017) Poetic meter classification using TMS320C6713 DSK. In: 2017 international conference on signal processing and communication (ICSPC). IEEE, pp 23–27

[7] D. Genzel, J. Uszkoreit, and F. Och, “"Poetic" Statistical Machine Translation: Rhyme and Meter,” Proc. 2010 Conference on Empirical Methods in Natural Language Processing (ACL), 2010, pp. 158–166.