Utterance Emotion Dynamics in Children's Poems: Emotional Changes Across Age

Daniela Teodorescu, Alona Fyshe
University of Alberta
{dteodore,alona}@ualberta.ca

Saif M. Mohammad
National Research Council Canada
saif.mohammad@nrc-cnrc.gc.ca

Abstract

Emerging psychopathology studies are showing that patterns of changes in emotional state — *emotion dynamics* — are associated with overall well-being and mental health. More recently, there has been some work in tracking emotion dynamics through one's utterances, allowing for data to be collected on a larger scale across time and people. However, several questions about how emotion dynamics change with age, especially in children, and when determined through children's writing, remain unanswered. In this work, we use both a lexicon and a machine learning based approach to quantify characteristics of emotion dynamics determined from poems written by children of various ages. We show that both approaches point to similar trends: consistent increasing intensities for some emotions (e.g., anger, fear, joy, sadness, arousal, and dominance) with age and a consistent decreasing valence with age. We also find increasing emotional variability, rise rates (i.e., emotional reactivity), and recovery rates (i.e., emotional regulation) with age. These results act as a useful baselines for further research in how patterns of emotions expressed by children change with age, and their association with mental health.

1 Introduction

Emotions play a key role in overall well-being (Kuppens and Verduyn, 2017; Houben et al., 2015; Silk et al., 2011; Sperry et al., 2020). People's emotional states are constantly changing in response to internal and external events, and the way in which we regulate emotions (Zimmermann and Iwanski, 2014; McRae et al., 2012). Patterns of emotion change over time have been shown to be related to general well-being and psychopathology (the scientific study of mental illness and disorders) (Houben et al., 2015; Sperry et al., 2020; Scott et al., 2020; Sheppes et al., 2015), academic success (Graziano et al., 2007), and social interactions in children (Sosa-Hernandez et al., 2022).

Several psychopathology studies have introduced metrics to quantify and understand the trajectories and patterns in emotions across time (Kuppens and Verduyn, 2017). These metrics are referred to as *Emotion Dynamics* and include features of the emotional episode (e.g., duration) and of the emotional trajectory (e.g., emotional variability, covariation, inertia) (Kuppens and Verduyn, 2017). In psychology, emotion dynamics have usually been captured through self-report surveys over periods of time (e.g., five times a day for ten days). However, obtaining such self-reports is arduous work; limiting the amount of data collected. Further, self-reports are prone to a number of biases (e.g., social pressures to be perceived as being happy).

Inspired by the emotion dynamics work in psychology, Hipson and Mohammad (2021) recently introduced the idea that patterns of emotion change can also be explored in the utterances of an individual, which can reflect their inner emotion dynamics. They refer to this as *utterance emotion dynamics (UED)*. They generate emotion arcs from streams of text (e.g., sentences in a story, tweets over time, etc.), which are in turn used to determine various UED metrics.[1] Different UED metrics capture different aspects of emotion change (e.g., variability, rate of change, etc.).

Teodorescu and Mohammad (2022) performed experiments on 36 diverse datasets to show that the quality of emotion arcs generated using emotion lexicons is comparable to those generated from machine learning (ML) methods. The lexicon approach is able to perform well through the power of aggregating information (e.g., 50–100 instances per bin). Moreover, the lexicon approach obtains high performance even when using translations of

[1]An emotion arc is a series of time step–emotion value pairs that acts as a digital representation of how one's emotions change over time. There are several works in NLP that capture emotion arcs from streams of text (e.g., sentences in a story, tweets over time, etc.) (Mohammad, 2011, 2012; Reagan et al., 2016; Teodorescu and Mohammad, 2022, 2023).

Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis, pages 401–415
July 14, 2023 ©2023 Association for Computational Linguistics

an English lexicon into low-resource languages, such as indigenous African languages (Teodorescu and Mohammad, 2023). Emotion lexicons have the benefit of interpretability, accessibility, and efficiency compared to ML models. Thus, we primarily used a lexicon-based approach in our experiments. However, we also show that the use of ML models points to the same trends as discovered by the lexicon approach.[2]

UED metrics, calculated from emotion arcs, can be computed for a *speaker* over time (e.g., main character in a narrative, tweets of a user over time), for multiple speakers at a time (e.g., treating all users in a geographic region as a *speaker* for whom we can compute UED), or at an *instance* level (e.g., independent posts where we compute UED metrics per post). While emotion dynamics have been studied in psychology for the past decades, UED was proposed only recently and has been applied to only a small number of domains (literature and tweets). Important questions such as *how do UED metrics change over development from toddlers to young adults?* and *how do the metrics change across one's adult life?*, remain unanswered.

Generally, children's writing is a less studied domain in NLP, and there is limited data available. Also, research regarding children has guidelines and regulations in place to protect this vulnerable section of society (Hipson and Mohammad, 2020). Yet, careful and responsible work such as the work done on the Child Language Data Exchange System (CHILDES) (MacWhinney, 2014) for understanding child language acquisition can be tremendously influential. Similarly, applying UED metrics to children's writing will allow us to infer the emotional states of children across age. Such work provides important information for psychologists and child development specialists, as emotion dynamics have been shown to underlie well-being, psychopathology, and success.

Poetry is a domain of growing interest in NLP (e.g., poem generation (Van de Cruys, 2020; Gonçalo Oliveira, 2017)). Poems are written to evoke emotions (Wassiliwizky et al., 2017; Johnson-Laird and Oatley, 2022) and a medium through which emotions are expressed (Whissell, 2004; Belfi et al., 2018). The intersection of poems and children's writing is an unexplored area which has the potential to unlock patterns in emotion word usage by children as they age.

[2]We did not find any poem datasets annotated for emotions that could be used to train an ML model; so we fine-tuned a pretrained ML model on emotion annotated tweets.

In this paper we contribute to the knowledge of emotion change over time as children age by studying poems written by children. Our goal is to apply existing NLP techniques to study emotion change in childrens' writing rather than developing a novel algorithm for better detecting emotion. We investigate the following questions:

- How do the *average* emotions vary across grades? How does this compare for discrete emotions (e.g., anger, fear, joy, and sadness) and emotion dimensions (e.g., valence, arousal, and dominance)?
- How *variable* are emotion changes?

These first two questions help us set crucial metrics in UED, building on work by Hipson and Mohammad (2020). Next, to better understand patterns in emotion changes we look at:

- How does the rate at which children reach peak emotional states (*rise rate*) change with age? Rise rate is analogous to emotional reactivity, which is associated with well-being.
- How does the rate at which children recover from peak emotional states back to steady state (*recovery rate*) change with age? Recovery rate plays a role in emotion regulation, which is also associated with well-being.
- How do *utterance emotion dynamics* compare for adults vs. children?

Answers to these questions provide baseline metrics for emotion change in children's poems across age. In order to answer these questions, we use a dataset of $\sim 61K$ poems written by children (Hipson and Mohammad, 2020) to calculate various UED metrics and examine how they vary across age. The scores for the metrics and the analysis will act as useful baselines for further research on emotion dynamics in children's writing, and their implications on mental health and well-being.

2 Related Work

Below we review related work on emotion dynamics and its ties to well-being, the UED framework, and previous work on children's texts.

2.1 Emotion Dynamics

The *emotion dynamics* framework studies change in emotion over time as it is key to the study of emotions and overall well-being (Houben et al., 2015; Silk et al., 2011). Emotion dynamics metrics include *emotion intensity* and *emotion variability*.

Dataset	# of Poems	#Words per Poem
PoKi	61,330	14.3
Grade 1	900	37.3
Grade 2	3,174	32.1
Grade 3	6,712	35.2
Grade 4	10,899	39.3
Grade 5	11,479	44.5
Grade 6	11,011	49.6
Grade 7	7,831	59.7
Grade 8	4,546	67.6
Grade 9	1,284	91.5
Grade 10	1,171	91.8
Grade 11	667	103.0
Grade 12	1,656	97.2
FPP	50	181.02

Table 1: Number of poems and the average lengths of poems in *PoKi* (by grade) and in *FPP*.

Emotion intensity is the average emotion over time. Whereas emotion variability is how much emotion changes from the average, often expressed as the standard deviation. These metrics have been used in various contexts in psychology to better understand well-being, often through self-reports or ecological momentary assessments.

The relationship between various metrics in emotion dynamics and well-being have been the topic of numerous psychology studies. Higher positive and negative affect variability have been shown to be associated with lower psychological well-being and more mental health symptoms in youth (Silk et al., 2003; van Roekel et al., 2016) and adults (Houben et al., 2015). Houben et al. (2015) showed that emotion variability has significant correlation with numerous psychological well-being categories: positive correlation with negative emotionality (e.g., negative affect and neuroticism), depression (e.g., depressive symptoms, depressive diagnosis), anxiety, borderline personality disorder, etc. On the other hand, emotional variability is negatively correlated with self-esteem, quality of life, and other signs of high psychological well-being (Houben et al., 2015).

Similarly, a vast number of studies explored the relationship between emotional regulation and reactivity with overall well-being. Hofmann et al. (2012) showed that mood and anxiety disorders are a result of emotion dysregulation of negative emotions, along with lacking positive emotions. Likewise, emotion dysregulation is thought of as the core of anxiety disorders (Mennin et al., 2007; Carthy et al., 2010). Children with anxiety disorders had higher negative emotion reactivity, and were less successful at implementing emotion regulation strategies (Carthy et al., 2010).

2.2 Utterance Emotion Dynamics

As work in psychology measures emotion dynamics through self-report measures, emotion dynamics can also be determined from text using NLP techniques such as sentiment analysis.

The UED framework (Hipson and Mohammad, 2021) tracks emotions dynamics in utterances, inspired by metrics in psychology. Such metrics include:

- *Home base*: The steady (most common) state where one is on average in emotional space.
- *Variability*: How much one's emotional state changes with time.
- *Rise Rate*: The rate at which one reaches peak emotional intensity, i.e., emotional reactivity.
- *Recovery Rate*: The rate at which one recovers from peak emotional intensity to home base, i.e., emotional regulation.

This framework was used to study emotion arcs of movie characters (Hipson and Mohammad, 2021), and to analyze emotional patterns across geographic regions through Twitter data (Vishnubhotla and Mohammad, 2022). Seabrook et al. (2018) studied the association between depression severity and emotion dynamics metrics such as variability on Facebook and Twitter. It was found that increased negative emotional variability was an indicator for lower depression severity on Twitter.

2.3 Children's Writing

Few work studies children's writing due to the limited data available. One of the most commonly known datasets is the Child Language Data Exchange System (CHILDES) (MacWhinney, 2014) and in French, E-CALM (Doquet, 2013). These datasets are limited in that they contain parent-child dialogue for children approximately age one to seven, and have limited quantities of text.

Very few works look at emotions in children's writing. Manabe et al. (2021) performed sentiment analysis on narratives written by youth for mental illness detection, as self-disclosure is not the norm in some cultures. Participants wrote an imaginative story and answered a questionnaire on their tendencies toward psychological distress. It was found that youth who had higher tendencies toward psychological distress used significantly more positive words, and therefore had higher valence. In this work, we study the patterns of emotion word changes in poems written by children.

3 Poem Datasets

For our experiments, we used a dataset of poems written by children as well as a dataset of poems written by adults (as control). Table 1 shows key statistics of each dataset.

Poems Written by Children (PoKi): Hipson and Mohammad (2020) compiled and curated a dataset of close to 61 thousand poems written by children in grades one to twelve. The poems were published and publicly available on the Scholastic Corporation website.[3] In the PoKi dataset each poem is released with the child's school grade (which can be used as a proxy for age) and first name.

The average emotional patterns for emotion dimensions (valence, arousal, and dominance), along with discrete emotions (anger, fear, joy and sadness) were analysed across grades. Additionally, these patterns were contrasted to those found in poems written by adults (data described below). It was found that as children grow from early childhood into adolescence, valence decreases reaching a minimum at grade 11. Whereas arousal increases with age, aligning with how adults display emotions more visibly (Dreyfuss et al., 2014; Somerville et al., 2011). Likewise, dominance increases with age. Consistently there was higher arousal in poems written by children with names commonly among males compared to those with names common among females. All intensities for anger, fear, joy and sadness increased across grades, with a particularly strong increase in sadness.

Poems Written by Adults (FPP): Hipson and Mohammad (2020) also compiled and used poems written by adults which were published on the *Famous Poets and Poems* website.[4] We will refer to this dataset as *FPP*. The poems are publicly available online and contain works by famous writers such as Edgar Allan Poe, and E.E. Cummings.

Preprocessing: We preprocessed both poem datasets by removing extra whitespace, punctuation, unescaping HTML (if any), tokenizing and lowercasing the text using the Twokenize[5] library. Additionally, stop words were removed.

4 Types of UED Metrics

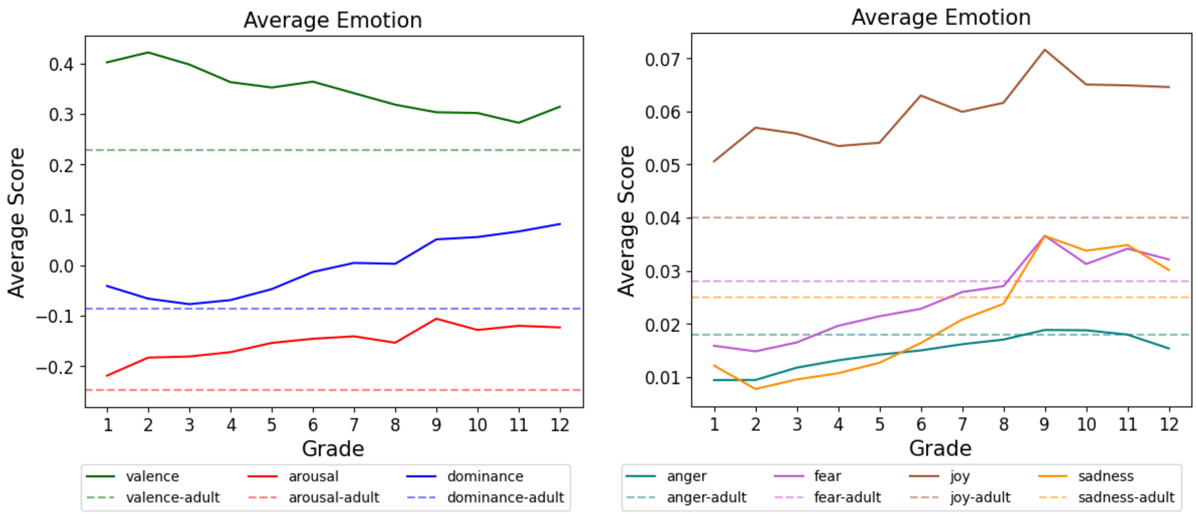
In the past, UED metrics have been calculated for the speaker or jointly for text from a set of speakers (meta-speaker). We propose a third form of UED metrics not explored before — *instance* level UED metrics. All three of these types of UED metrics are described below:

- **Speaker UED Metrics:** Here all available utterances by a speaker are placed in temporal order to form the text from which the UED metrics for the speaker is determined. For each metric, UED scores from multiple speakers can be averaged to determined the average UED score for that metric for the population. In the past, speaker UED metrics have been determined for characters in movie dialogues (Hipson and Mohammad, 2021), and for users on Twitter during the pandemic (Vishnubhotla and Mohammad, 2022).
- **Meta-Speaker UED Metrics:** If one is interested in analyzing change of emotions in a discourse by multiple speakers, for example, analyzing changes in emotion patterns in a Reddit thread, then we can treat each discourse (e.g., each Reddit thread) as text produced by a meta-speaker. Here we arrange each of the utterances in each of the discourses (e.g., Reddit thread) in temporal order and determine the UED metrics for each discourse. UED metrics for all of the discourses can be averaged to determine the average UED metric scores for a set of discourses. In the past, discourse UED metrics have been determined for users from geographic regions, such as treating all users on Twitter in a country as a speaker (Vishnubhotla and Mohammad, 2022).
- **Instance UED Metrics:** If one is interested in the change of emotions in individual pieces of text (or instances) such as a novel, a poem, a tweet, or a blog post, then we can simply apply the UED metrics to each instance. Such a metric is useful at individual instance level if the instance is long enough (otherwise the score for the metric is not a reliable on its own). However, even for smaller pieces of text, the UED scores from a large number of instances can give a reliable estimate of the distribution of these instance-level UED metrics. Such metrics can even be used to compare patterns of emotion change across different sets, where each set is composed of (a) instances from many speakers and (b) instances that are temporally unordered (either because that information is not available or because we are

[3] Hipson and Mohammad (2020) obtained permission to use these poems for research.

[4] http://famouspoetsandpoems.com/top_poems.html

[5] https://github.com/myleott/ark-twokenize-py



(a) Valence, arousal and dominance

(b) Anger, fear, joy, and sadness

Figure 1: Average emotion across grades. The horizontal dashed lines represent values in poems written by adults.

not interested in temporal ordering of items within a set). Examples of instance-level UED use include: determining UEDs of presidential speeches, comparing average UEDs of stream of consciousness essays of different age groups, etc.

In this work we are interested in children's poems (instances of poetic text) across age and not how each individual child has a different writing style. Therefore, we calculate UED metrics for each poem and average the scores for each grade.

5 Experiments

Our goal is to analyze how patterns of emotion words change with age in children's poems. In order to do so, we generate an emotion arc per poem and compute instance-level UED metrics. Afterwards, we average the UED metrics per grade to compare results across age. We use the Emotion Dynamics toolkit[6] to calculate UED metrics and our code for the experiments is available online.[7]

We use text windows of size five (excluding words with a neutral emotion score) and a step size of one to create an emotion arc per poem. We only considered poems that included at least five emotion words[8]. For each research question we computed the corresponding metrics: average emotional state, emotional variability, rise rate and recovery rate. Analyzing average emotion and variability allows us to build foundational knowledge into changes in patterns of emotion words. We then look at rise rate and recovery rate to further our understanding of children's emotion dynamics.

While older children (e.g., grade 10–12) tend to write on average longer poems than younger children (e.g., grade 1–3), these UED metrics are not affected by the length of the poems.[9] Other metrics calculate the number of displacements from home base or the length of displacements to peaks which are affected by poem length. Additionally, because poems are shorter than text streams such as novels, the number of windows that can be created from a poem is limited, so metrics specific to emotional displacement are not computed since they are more suitable for longer texts.

Each metric is computed for both dimensional emotions (e.g., valence, arousal, dominance) and discrete emotions (e.g., anger, fear, joy and sadness). We used the NRC Valence, Arousal, and Dominance (VAD) Lexicon (Mohammad, 2018a) and the NRC Emotion Intensity Lexicon (Mohammad, 2018b) for word-emotion scores.

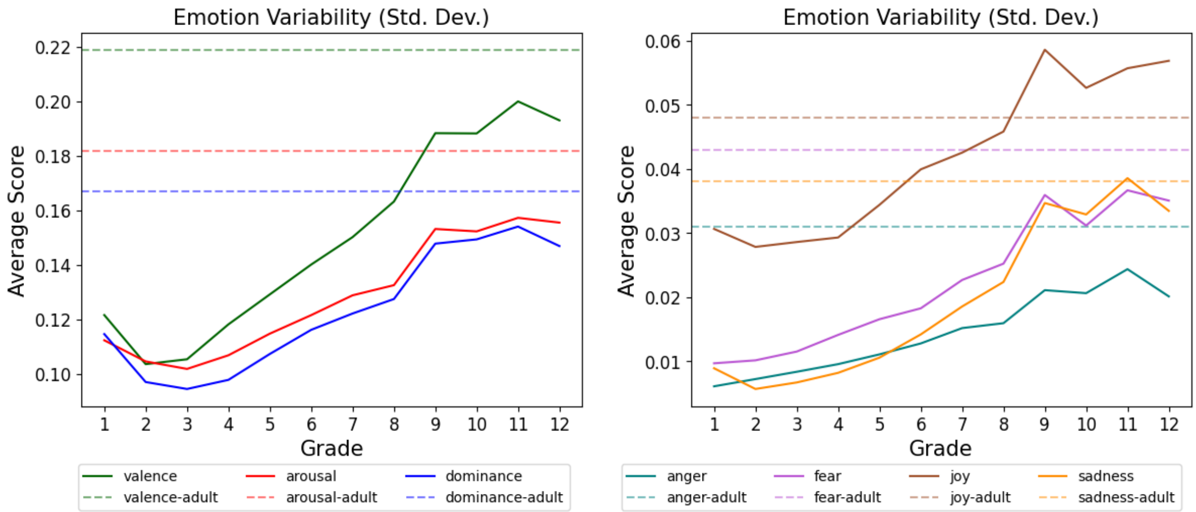
In Section 5.1 we use the lexicon-based approach to generate emotion arcs. We explain how the metrics are computed, contrast the trends across grades and compare the results to poems written by adults. We discuss the ties of these results with work in psychology and implications for emotional development. In Section 5.2 we explore the same questions using an ML model for generating emotion arcs. We find similar trends across grade with the ML approach as when using the lexicon approach.

[6]https://github.com/Priya22/EmotionDynamics

[7]https://github.com/dteodore/EmotionArcs

[8]as per the NRC VAD lexicon

[9]We show in Appendix A that similar patterns in UED metrics hold when controlling for poem length across grades.



(a) Valence, arousal and dominance

(b) Anger, fear, joy, and sadness

Figure 2: Emotion variability across grades. The horizontal dashed lines represent values in poems written by adults.

UED Metric	Valence	Arousal	Dominance	Anger	Fear	Joy	Sadness	Psych. Construct
Average	0.228	-0.247	-0.087	0.018	0.028	0.040	0.025	Intensity
Variability	0.219	0.182	0.167	0.031	0.043	0.048	0.038	Emotional Variability
Rise Rate	0.134	0.114	0.084	0.115	0.109	0.066	0.113	Emotional Reactivity
Recovery Rate	0.127	0.105	0.086	0.024	0.028	0.023	0.020	Emotional Regulation

Table 2: The values for UED metrics in poems written by adults, and the corresponding construct in psychology.

5.1 Utterance Emotion Dynamics: PoKi

We begin with a question on how average emotion word score changes with grade–a question that Hipson and Mohammad (2020) already answered in their work. We replicated the experiment to make sure any differences in preprocessing the data or code development did not lead to different results. We then answer the other questions on how specifically do the trajectories of emotion change across grade differ, which have not been addressed yet. Likewise, we compute the UED metrics on the poems written by adults. We show the results in Table 2 and contrast them to PoKi below.

5.1.1 How does the average emotion expressed change across age?

Method: An average emotion score is calculated per window in the poem using word-emotion scores from the lexicon, and then the average is computed across windows in a poem.

Results: Below we present results on both the valence, arousal, and dominance (VAD) dimensions as well as for discrete emotion categories (Anger, Fear, Joy, Sadness).

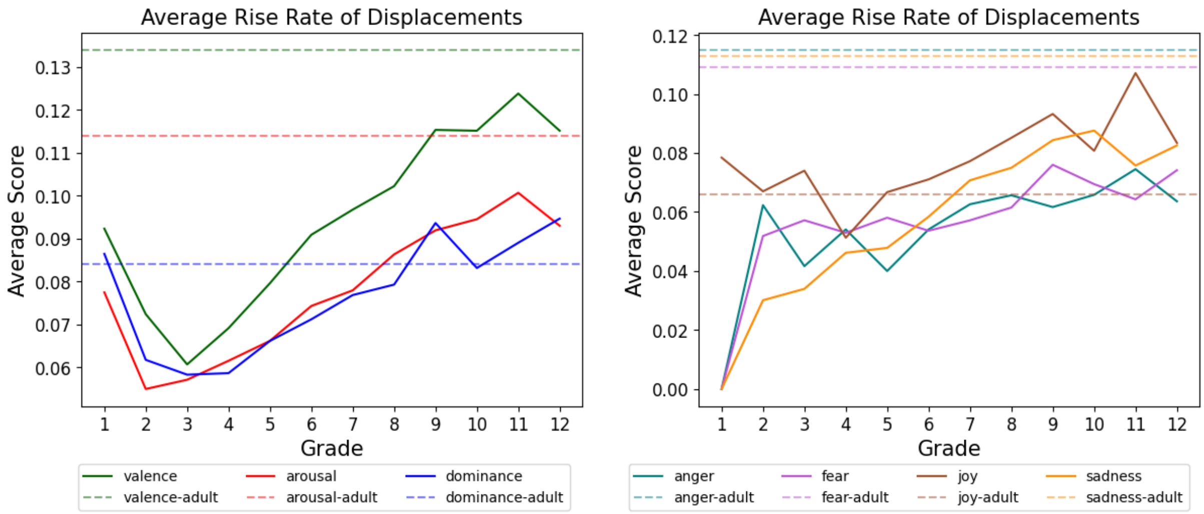
PoKi VAD: In Figure 1a, we show the average VAD emotions expressed across grade. Overall, we see a downward trend in valence from Grade 1 to Grade 12. This means that the poems written by younger children have, on average, more positive emotion words than those written by older children. There is a slight peak at grade 6, however a consistent downwards trend overall. Arousal and dominance similarly both trend upwards with age. There is a steeper increase for arousal and dominance at grade 9. This means that children are expressing more active and powerful emotions in poems as they age.

FPP VAD: The average valence of 0.228 is notably lower than the valence across grades, where the lowest is reached by grade 11s at 0.28. The average arousal at -0.247 and dominance at -0.087 are lower than those of children across all ages, and interestingly most similar to younger children.

PoKi Anger, Fear, Joy, Sadness: In Figure 1b, we see that the average discrete emotions all increase across grades. Anger, while increasing from grade 1 to 9, has a downward trend from grade 10 to 12. All emotions tend to have a peak around grade 9 and plateau afterwards.

FPP Anger, Fear, Joy, Sadness: Anger, fear and sadness tends to match to those of older children around grade 8 to 9. Children from grade 9 to 12 reach even higher values than adults for fear and sadness. On the other hand, joy tends to remain below those of children across all age, and has the most similar values to younger children at 0.04.

Discussion: These findings align with those by



(a) Valence, arousal and dominance

(b) Anger, fear, joy and sadness

Figure 3: Rise rate in poems across grades. The horizontal dashed lines represent values in poems written by adults.

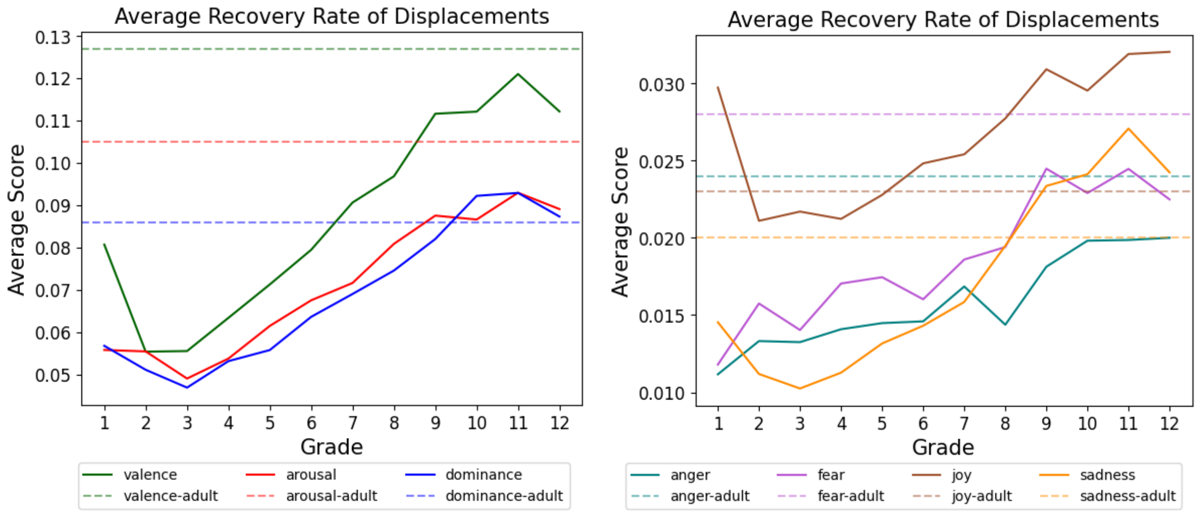
Hipson and Mohammad (2020) which similarly computed the mean emotion in poems across grade. Numerous works in psychology have found similar trends through self-report studies for valence (Frost et al., 2015; Larson et al., 2002; Simmons et al., 1987; Weinstein et al., 2007), and arousal (Carstensen et al., 2000; Gunnar et al., 2009; Somerville, 2013). Likewise, as sadness increased with age, Holsen et al. (2000) have shown that teenagers are more likely to experience a negative and depressed mood. This trend matters because we are seeing similar trends in the emotion words used by children when writing poems as those in psychology self-reports, although they were not told to explicitly talk about how they are feeling. This work further contributes to the current findings on emotional development in children.

5.1.2 How variable are emotions across age?

Method: Variability is computed as the standard deviation of emotion values for windows in a poem.
Results:
PoKi VAD: In Figure 2a, variability for valence, arousal, and dominance all trend upward with age; stabilizing in grades 11 and 12.
FPP VAD: For all three emotions variability was most similar to those of older children, reaching slightly above grades 10–12.
PoKi Anger, Fear, Joy, Sadness: In Figure 2b, we see that variability for all emotions trend upwards from grade 1 to 9, and start to level out around grade 10 to 12. Anger, fear, and sadness all have a peak at grade 9 and grade 11. Joy has an especially pronounced peak at grade 9.
FPP Anger, Fear, Joy, Sadness: Variability in anger, fear and sadness is higher for adults than those expressed by children across all grades, and is most similar to older children around grade 11. Likewise, variability for joy in adults is more similar to older children, however around grade 8.
Discussion: The overall trend of increasing emotional variability with age, followed by stabilizing supports findings in psychology. Larson et al. (2002) found that emotional variability increased over early adolescence and stabilized around mid-adolescence. Further, during adolescence important cognitive, social and psychical changes occur which are thought to increase emotional variability (Buchanan et al., 1992; Arnett, 1999; Steinberg, 2005). Reitsema et al. (2022) found that sadness variability statistically increased with age. These trends are important as they support those found in psychology which are strongly associated with mental well-being (Reitsema et al., 2022).

5.1.3 At what rate do emotions change from home to peak state?

Method: The average rise rate is calculated as the average of the rise rate for windows in a poem. The rise rate is *peak distance* (how far away the peak is from the home base) divided by the number of words during the rise period. The rise rate disregards the direction of the peak.
Results:
PoKi VAD: In Figure 3a, we see that rise rate increases for all three emotions across grade, and plateaus around grade 10 to 12. The rise rate is comparably higher for valence, followed by arousal



(a) Valence, arousal and dominance

(b) Anger, fear, joy and sadness

Figure 4: Recovery rate in poems across grades. The horizontal dashed lines represent values for poems by adults.

and then dominance.

FPP VAD: The rise rate for valence and arousal in adults is higher than those across all grades, and is most similar to older students in grade 11. The rise rate for dominance in adults also matches those of older children, however starting at grade 8 (with grade 9, 11 and 12 having a higher rate than adults).

PoKi Anger, Fear, Joy, Sadness: In Figure 3b, the rise rate for the discrete emotions all increase with grade. Joy has a small dip around grade 4, and then increases matching the average rise rate of anger, fear, and sadness which all started at slightly lower values in grade 2. We note that at grade 1 we could not compute the average rise rate for anger, fear, and sadness as the poems had too few displacements (the number of poems which had displacements was less than our pre-chosen threshold of 5).

FPP Anger, Fear, Joy, Sadness: The rise rate for anger, fear, and sadness in adult poems is higher than those expressed in children across all ages, with most similar values to older children. However, the rise rate for joy is lower and corresponds with younger children around grade 2 to 5.

Discussion: Rise rate is seen as analogous to reactivity in psychology, which has been found to increase during adolescence (Somerville, 2016). Our findings support these trends. As mentioned in 2.1, emotional reactivity is at the core of anxiety and attention disorder, impacting overall well-being.

5.1.4 At what rate do emotions recover?

Method: Recovery rate is computed similarly to rise rate, however divides peak distance by the number of words during the recovery period. Recovery rate does not distinguish between peak direction.

Results:

PoKi VAD: Figure 4a we see the recovery rate increases for all three emotions with age and plateaus around grade 10 to 12. While the valence recovery rate has a larger magnitude than the other emotions, all rates trend upwards. Recovery rate can be thought of *emotion regulation*, indicating that older children are able to return to their home base emotional states after a peak more quickly than younger children.

FPP VAD: The recovery rate of adults for valence and arousal corresponds most closely with older children (e.g., grades 9–12), however is higher than across all grades. The recovery rate of dominance is similar to grade 9 students, and slightly below those of older children.

PoKi Anger, Fear, Joy, Sadness: In Figure 4b, we similarly see increasing average recovery rates for all 4 emotions across age. The magnitude of joy's recovery rate is considerably higher than for the other 3 emotions.

FPP Anger, Fear, Joy, Sadness: The recovery rate for fear and anger is above those across all ages, and is most similar to older children in grade 9–12. On the other hand, the recovery rate for joy and sadness matches those of younger children, around grade 5 for joy and 8 for sadness.

Discussion: Recovery rate, which is analogous to emotion regulation, has been studied extensively in psychology. Zeman et al. (2006) detail the progression of emotional regulation from infancy to adolescence, in which an increase in emotion regulation occurs alongside developments in strategies and motivations. Not only does emotion regulation have ties with well-being, it also plays a role in aca-

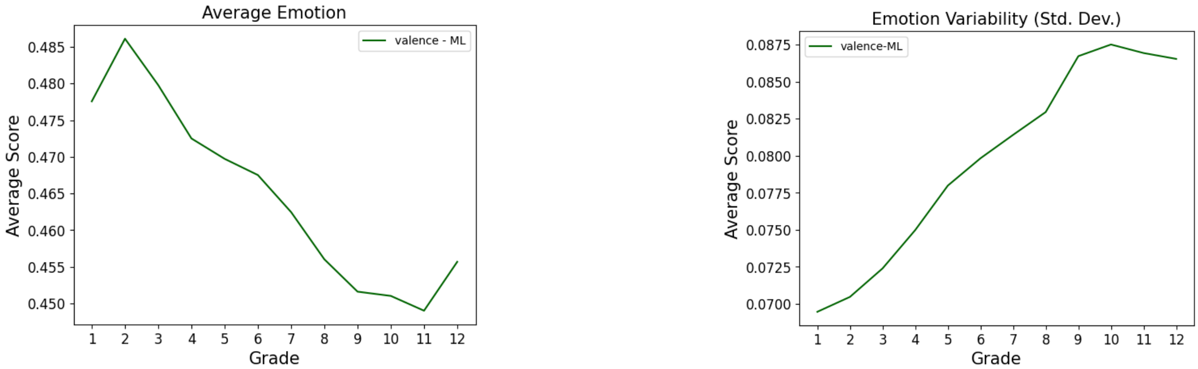


Figure 5: Average emotion and emotion variability for valence using the ML *n-gram* approach on the PoKi dataset.

demic success of children (Graziano et al., 2007) and adults (Phillips et al., 2002).

5.2 Utterance Emotion Dynamics - ML Approach: PoKi

To perform a comprehensive analysis and ensure the trends in emotion change are consistent regardless of the emotion labelling method used, we also performed experiments using a ML model. Previously, individual words were emotion labelled using a lexicon. Now, we use a *n-gram* approach where a ML model assigns emotion scores to windows of text in the poem of length *n*. We are not trying to determine which of these two approaches is *better* at computing UED metrics as this would be challenging - there are no existing annotated datasets for emotion arcs, or UED metrics. Rather, we are supporting the trends found by the word-level lexicon approach, with those found by ML models as they are commonly used on downstream tasks (e.g., sentiment analysis) and are known for their strong performance. If the ML approach did not perform well, we would not expect any trends in UED metrics to appear.

Datasets: We use the same poem datasets as in Section 3, creating n-gram windows of length 5. The only difference is that words not found in the emotion lexicon or neutral words can be included. We choose this approach as ML models are trained on sequential text.

Experiments: We fine-tuned a RoBERTa (Liu et al., 2019) base model for fine-grain sentiment analysis using the SemEval 2018 Task 1 dataset (Mohammad et al., 2018). This means that we were able to predict an emotion score between 0 and 1. Details on the model training are in Appendix B.1. After emotion labelling text windows, we performed similar experiments as in Section 5: compute the UED metrics per poem and take the average per grade for each metric.

Results: Overall we found similar trends as with the *word-level* lexicon approach. We note that a direct comparison between the lexicon and ML approach can not be made as they are using different units of measurement (e.g., windows contain either sequential words found in the lexicon or natural sequences of words). We can simply compare the trends in emotion change rather than the magnitude of change or the values themselves. We discuss the results for valence below (Figure 5 and Appendix B.2). We also show the results for the discrete emotions in Appendix B.2, as the trends were similar to the lexicon approach.

Average Emotion: As grade increases, we see a similar downwards trend and a stabilization at grades 10–12.

Emotional Variability: Older children tend to show increased variability.

Rise Rate and Recovery Rate: With age, children are writing with increased rates of emotional reactivity and also emotion regulation.

Overall, these results show that there are patterns of emotion change in childrens' poems with age, and trends found using the lexicon approach are also replicated using ML models.

6 Conclusion

We explored four utterance emotion dynamics metrics (average, variability, rise rate, and recovery rate), and seven emotions (three dimensional and four discrete) on poems written by children and adults. We found that the patterns of emotion changes in poetry by children supported previous results and findings in the psychology literature (e.g., increased variability, rise rate, and recovery rates with age).

As future work, we would like to examine poetry by adults more in-depth, such as how do patterns of emotion change look for experts vs. novices? And how do UED compare across geographic regions, and time periods.

Limitations

A limitation of this work is that the poems written by adults are by experienced writers who are often known for their poetry. These poems may therefore not be representative of poems written by adults in general, and could affect the patterns and trends in emotion words we see. As future work we would like to expand the collection of poems written by adults to include those written by novices as well.

Ethics Statement

Our research interest is to study emotions at an aggregate/group level. This has applications in emotional development psychology and in public health (e.g., overall well-being and mental health). However, emotions are complex, private, and central to an individual's experience. Additionally, each individual expresses emotion differently through language, which results in large amounts of variation. Therefore, several ethical considerations should be accounted for when performing any textual analysis of emotions (Mohammad, 2022, 2023). The ones we would particularly like to highlight are listed below:

- Our work on studying emotion word usage should not be construed as detecting how people feel; rather, we draw inferences on the emotions that are conveyed by users via the language that they use.
- The language used in an utterance may convey information about the emotional state (or perceived emotional state) of the speaker, listener, or someone mentioned in the utterance. However, it is not sufficient for accurately determining any of their momentary emotional states. Deciphering true momentary emotional state of an individual requires extra-linguistic context and world knowledge. Even then, one can be easily mistaken.
- The inferences we draw in this paper are based on aggregate trends across large populations. We do not draw conclusions about specific individuals or momentary emotional states.

Acknowledgements

We thank Will Hipson for collecting the PoKi poems dataset and for setting the groundwork in this direction of research. We also thank Krishnapriya Vishnubhotla for the toolkit she created to help compute UED metrics.

References

JJ Arnett. 1999. Adolescent storm and stress, reconsidered. *The American psychologist*, 54(5):317—326.

Amy M Belfi, Edward A Vessel, and G Gabrielle Starr. 2018. Individual ratings of vividness predict aesthetic appeal in poetry. *Psychology of Aesthetics, Creativity, and the Arts*, 12(3):341.

CM Buchanan, JS Eccles, and JB Becker. 1992. Are adolescents the victims of raging hormones: evidence for activational effects of hormones on moods and behavior at adolescence. *Psychological bulletin*, 111(1):62—107.

LL Carstensen, M Pasupathi, U Mayr, and JR Nesselroade. 2000. Emotional experience in everyday life across the adult life span. *Journal of personality and social psychology*, 79(4):644—655.

Tal Carthy, Netta Horesh, Alan Apter, and James J. Gross. 2010. Patterns of emotional reactivity and regulation in children with anxiety disorders. *Journal of Psychopathology and Behavioral Assessment*, 32(1):23–36. This study was supported by a research fund of the Adler Research Center in Tel-Aviv University. The authors would like to thank the Anxiety Disorders Clinic in 'Schneider's Children Medical Center of Israel' for support and collaboration. Special thanks to Ronit Jossifoff, Maya Ferber, Yael Tadmor and Hilit Pritsch for their important contribution to the recruitment and examination of the participants.

Claire Doquet. 2013. Ancrages théoriques de l'analyse génétique des textes d'élèves. In Catherine Bore et Eduardo Calil, editor, *L'Ecole, l'écriture et la création. Etudes françaises et brésiliennes.*, Sciences du langage - Carrefour et points de vue, pages 33–53. Academia Bruylant.

Michael Dreyfuss, Kristina Caudle, Andrew T Drysdale, Natalie E Johnston, Alexandra O Cohen, Leah H Somerville, Adriana Galván, Nim Tottenham, Todd A Hare, and BJ Casey. 2014. Teens impulsively react rather than retreat from threat. *Developmental neuroscience*, 36(3-4):220—227.

Allison Frost, Lindsay T. Hoyt, Alissa Levy Chung, and Emma K. Adam. 2015. Daily life with depressive symptoms: Gender differences in adolescents' everyday emotional experiences. *Journal of Adolescence*, 43:132–141.

Hugo Gonçalo Oliveira. 2017. A survey on intelligent poetry generation: Languages, features, techniques, reutilisation and evaluation. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 11–20, Santiago de Compostela, Spain. Association for Computational Linguistics.

Paulo A. Graziano, Rachael D. Reavis, Susan P. Keane, and Susan D. Calkins. 2007. The role of emotion regulation in children's early academic success. *Journal of School Psychology*, 45(1):3–19.

Megan R Gunnar, Sandi Wewerka, Kristin Frenn, Jeffrey D Long, and Christopher Griggs. 2009. Developmental changes in hypothalamus-pituitary-adrenal activity over the transition to adolescence: normative changes and associations with puberty. *Development and psychopathology*, 21(1):69—85.

Will Hipson and Saif M. Mohammad. 2020. PoKi: A large dataset of poems by children. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1578–1589, Marseille, France. European Language Resources Association.

Will E. Hipson and Saif M. Mohammad. 2021. Emotion dynamics in movie dialogues. *PLOS ONE*, 16(9):1–19.

Stefan G. Hofmann, Alice T. Sawyer, Angela Fang, and Anu Asnaani. 2012. Emotion dysregulation model of mood and anxiety disorders. *Depression and Anxiety*, 29(5):409–416.

Ingrid Holsen, Pal Kraft, and Joar Vitterso. 2000. Stability in depressed mood in adolescence: Results from a 6-year longitudinal panel study. *Journal of Youth and Adolescence*, 29(1):61–78. Copyright - Copyright Plenum Publishing Corporation Feb 2000; Last updated - 2023-02-07; CODEN - JYADA6.

Marlies Houben, Wim Van Den Noortgate, and Peter Kuppens. 2015. The relation between short-term emotion dynamics and psychological well-being: A meta-analysis. *Psychological Bulletin*, 141(4):901–930.

Philip N. Johnson-Laird and Keith Oatley. 2022. How poetry evokes emotions. *Acta Psychologica*, 224:103506.

Peter Kuppens and Philippe Verduyn. 2017. Emotion dynamics. *Current Opinion in Psychology*, 17:22–26. Emotion.

Reed W. Larson, Giovanni Moneta, Maryse H. Richards, and Suzanne Wilson. 2002. Continuity, stability, and change in daily emotional experience across adolescence. *Child Development*, 73(4):1151–1165.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Brian MacWhinney. 2014. *The CHILDES project: Tools for analyzing talk, Volume II: The database*. Psychology Press.

Masae Manabe, Kongmeng Liew, Shuntaro Yada, Shoko Wakamiya, and Eiji Aramaki. 2021. Estimation of psychological distress in japanese youth through narrative writing: Text-based stylometric and sentiment analyses. *JMIR Form Res*, 5(8):e29500.

Kateri McRae, James J. Gross, Jochen Weber, Elaine R. Robertson, Peter Sokol-Hessner, Rebecca D. Ray, John D.E. Gabrieli, and Kevin N. Ochsner. 2012. The development of emotion regulation: an fMRI study of cognitive reappraisal in children, adolescents and young adults. *Social Cognitive and Affective Neuroscience*, 7(1):11–22.

Douglas S. Mennin, Robert M. Holaway, David M. Fresco, Michael T. Moore, and Richard G. Heimberg. 2007. Delineating components of emotion and its dysregulation in anxiety and mood psychopathology. *Behavior Therapy*, 38(3):284–302.

Saif Mohammad. 2011. From once upon a time to happily ever after: Tracking emotions in novels and fairy tales. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 105–114, Portland, OR, USA. Association for Computational Linguistics.

Saif Mohammad. 2023. Best practices in the creation and use of emotion lexicons. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1825–1836, Dubrovnik, Croatia. Association for Computational Linguistics.

Saif M. Mohammad. 2012. From once upon a time to happily ever after: Tracking emotions in mail and books. *Decision Support Systems*, 53(4):730–741. 1) Computational Approaches to Subjectivity and Sentiment Analysis 2) Service Science in Information Systems Research : Special Issue on PACIS 2010.

Saif M. Mohammad. 2018a. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words. In *Proceedings of The Annual Conference of the Association for Computational Linguistics (ACL)*, Melbourne, Australia.

Saif M. Mohammad. 2018b. Word affect intensities. In *Proceedings of the 11th Edition of the Language Resources and Evaluation Conference (LREC-2018)*, Miyazaki, Japan.

Saif M. Mohammad. 2022. Ethics sheet for automatic emotion recognition and sentiment analysis. *To Appear in Computational Linguistics*.

Saif M. Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. Semeval-2018 Task 1: Affect in tweets. In *Proceedings of International Workshop on Semantic Evaluation (SemEval-2018)*, New Orleans, LA, USA.

Louise H Phillips, Louise H Phillips, Rebecca Bull, Ewan Adams, and Lisa Fraser. 2002. Positive mood and executive function: evidence from stroop and fluency tasks. *Emotion (Washington, D.C.)*, 2(1):12—22.

Andrew J. Reagan, Lewis Mitchell, Dilan Kiley, Christopher M. Danforth, and Peter S. Dodds. 2016. The emotional arcs of stories are dominated by six basic shapes. *EPJ Data Science*, 5(1):1–12. Copyright -

EPJ Data Science is a copyright of Springer, 2016; Last updated - 2017-02-06.

Anne M Reitsema, Bertus F Jeronimus, Marijn van Dijk, and Peter de Jonge. 2022. Emotion dynamics in children and adolescents: A meta-analytic and descriptive review. *Emotion (Washington, D.C.)*, 22(2):374—396.

Lori N Scott, Sarah E Victor, Erin A Kaufman, Joseph E Beeney, Amy L Byrd, Vera Vine, Paul A Pilkonis, and Stephanie D Stepp. 2020. Affective dynamics across internalizing and externalizing dimensions of psychopathology. *Clinical psychological science : a journal of the Association for Psychological Science*, 8(3):412—427.

Elizabeth M Seabrook, Margaret L Kern, Ben D Fulcher, and Nikki S Rickard. 2018. Predicting depression from language-based emotion dynamics: Longitudinal analysis of facebook and twitter status updates. *J Med Internet Res*, 20(5):e168.

Gal Sheppes, Gaurav Suri, and James J. Gross. 2015. Emotion regulation and psychopathology. *Annual Review of Clinical Psychology*, 11(1):379–405. PMID: 25581242.

Jennifer S. Silk, Erika E. Forbes, Diana J. Whalen, Jennifer L. Jakubcak, Wesley K. Thompson, Neal D. Ryan, David A. Axelson, Boris Birmaher, and Ronald E. Dahl. 2011. Daily emotional dynamics in depressed youth: A cell phone ecological momentary assessment study. *Journal of Experimental Child Psychology*, 110(2):241–257. Special Issue: Assessment of Emotion in Children and Adolescents.

Jennifer S Silk, Laurence Steinberg, and Amanda Sheffield Morris. 2003. Adolescents' emotion regulation in daily life: links to depressive symptoms and problem behavior. *Child development*, 74(6):1869—1880.

RG Simmons, R Burgeson, S Carlton-Ford, and DA Blyth. 1987. The impact of cumulative change in early adolescence. *Child development*, 58(5):1220—1234.

Leah H Somerville. 2013. Special issue on the teenage brain: Sensitivity to social evaluation. *Current directions in psychological science*, 22(2):121—127.

Leah H Somerville, Todd Hare, and BJ Casey. 2011. Frontostriatal maturation predicts cognitive control failure to appetitive cues in adolescents. *Journal of cognitive neuroscience*, 23(9):2123—2134.

LH Somerville. 2016. Emotional development in adolescence. *Handbook of emotions*, pages 350–365.

Linda Sosa-Hernandez, McLennon Wilson, and Heather A Henderson. 2022. Emotion dynamics among preadolescents getting to know each other: Dyadic associations with shyness. *Emotion (Washington, D.C.)*.

Sarah H. Sperry, Molly A. Walsh, and Thomas R. Kwapil. 2020. Emotion dynamics concurrently and prospectively predict mood psychopathology. *Journal of Affective Disorders*, 261:67–75.

Laurence Steinberg. 2005. Cognitive and affective development in adolescence. *Trends in Cognitive Sciences*, 9(2):69–74.

Daniela Teodorescu and Saif M. Mohammad. 2022. Frustratingly easy sentiment analysis of text streams: Generating high-quality emotion arcs using emotion lexicons.

Daniela Teodorescu and Saif M. Mohammad. 2023. Generating high-quality emotion arcs for low-resource languages using emotion lexicons.

Tim Van de Cruys. 2020. Automatic poetry generation from prosaic text. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2471–2480, Online. Association for Computational Linguistics.

Eeske van Roekel, Elise C Bennik, Jojanneke A Bastiaansen, Maaike Verhagen, Johan Ormel, Rutger C M E Engels, and Albertine J Oldehinkel. 2016. Depressive symptoms and the experience of pleasure in daily life: An exploration of associations in early and late adolescence. *Journal of abnormal child psychology*, 44(5):999—1009.

Krishnapriya Vishnubhotla and Saif M. Mohammad. 2022. Tweet Emotion Dynamics: Emotion word usage in tweets from US and Canada. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4162–4176, Marseille, France. European Language Resources Association.

Eugen Wassiliwizky, Stefan Koelsch, Valentin Wagner, Thomas Jacobsen, and Winfried Menninghaus. 2017. The emotional power of poetry: neural circuitry, psychophysiology and compositional principles. *Social cognitive and affective neuroscience*, 12(8):1229—1240.

Sally M. Weinstein, Robin J. Mermelstein, Benjamin L. Hankin, Donald Hedeker, and Brian R. Flay. 2007. Longitudinal patterns of daily affect and global mood during adolescence. *Journal of Research on Adolescence*, 17(3):587–600.

Cynthia Whissell. 2004. Poetic emotion and poetic style: The 100 poems most frequently included in anthologies and the work of emily dickinson. *Empirical Studies of the Arts*, 22(1):55–75.

Janice Zeman, Michael Cassano, Carisa Perry-Parrish, and Sheri Stegall. 2006. Emotion regulation in children and adolescents. *Journal of developmental and behavioral pediatrics : JDBP*, 27(2):155—168.

Peter Zimmermann and Alexandra Iwanski. 2014. Emotion regulation from early adolescence to emerging adulthood and middle adulthood: Age differences,

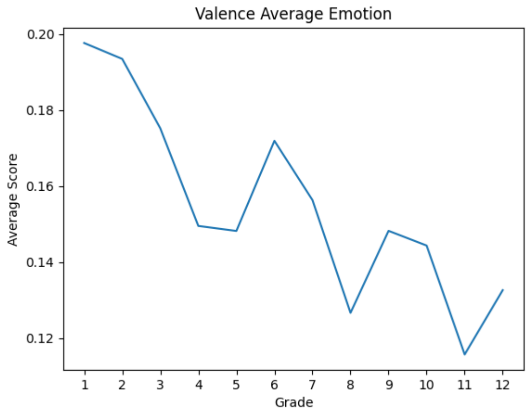


Figure 6: Average valence across grades for poems of length 10 to 20 words.

gender differences, and emotion-specific developmental variations. *International Journal of Behavioral Development*, 38(2):182–194.

A Poem Length on UED Metrics

As mentioned in Section 5, certain UED metrics which rely on distances (e.g., length of displacements to peaks) could be influenced by poem length. Therefore, we selected metrics which are based on rates or averages. To verify these metrics are not impacted by the increasing poem lengths with age, we investigated if the same trends hold when controlling for the length of poems across grade. In Figure 6 we show the results for the average valence across grades for poems of length 10 to 20 words (not including stop words). As grade increases, we similarly see a decrease in valence. Similar trends occur with other metrics.

B Machine Learning Approach

In Section B.1 we detail the model training process and in Section B.2 we show the results on the PoKi dataset using a ML approach.

B.1 Model Training

We fine-tuned the pretrained RoBERTa (Liu et al., 2019) base model available on HuggingFace[10]. For training we used the SemEval 2018 Task 1 dataset (Mohammad et al., 2018) which contains tweets annotated with emotion scores for valence, anger, fear, joy and sadness.[11] The dataset contains both fine-grain emotion scores (real-valued numbers between 0 and 1) and categorical labels (e.g., -1, 0, 1). We use the real-valued emotion scores to compute more fine-grained emotion arcs. More details on this dataset are available in Table 3 and Table 4.

We used the Trainer pipeline from HuggingFace[12] to fine-tune the pretrained model. For the loss function we used mean-square loss.

We tuned the following hyperparameters on the development set and selected the best model using mean-square error: learning rate (2e-5, 3e-5), number of epochs (5, 10, 20) and batch size (16, 32). Note that our aim here is not to overly fine-tune the model as we are applying it to a different domain (i.e., poems). The best parameters for each emotion model are shown in Table 5. After determining the *best* model on the development set we apply it to windows of text in the PoKi poem dataset.

B.2 UED Metric Results

In Figure 7 we show the results for rise rate and recovery rate for valence using the fine-tuned ML model on the PoKi dataset. Both rise rate and recovery rate increase with age. These trends support those seen when using the lexicon approach.

In Figure 8 we show the UED metrics across grade for the discrete emotions (e.g., anger, fear, joy, and sadness). The trends for fear and sadness are similar to trends found when using the lexicon approach: emotion intensity, variability, rise rate and recovery rate increase across grades. Overall, the patterns of emotion change for anger are flatter across metrics. Perhaps anger is a more challenging emotion for automatic systems to detect (Mohammad et al., 2018). The average intensity for joy has a similar pattern to that of valence. These two emotions could appear similar to the ML model resulting in similar trajectories.

[10] `https://huggingface.co/roberta-base`

[11] We could not train models for arousal and dominance as there are no corresponding annotated datasets.

[12] `https://huggingface.co/docs/evaluate/main/en/transformers_integrations#trainer`

Dataset	**Source**	**Domain**	**Dimension**	**Label Type**	**# Instances**
SemEval 2018 (EI-Reg)	Mohammad et al. (2018)	tweets	anger, fear joy, sadness	continuous (0 to 1)	3092, 3627, 3011, 2095
SemEval 2018 (V-Reg)	Mohammad et al. (2018)	tweets	valence	continuous (0 to 1)	2567

Table 3: Dataset descriptive statistics. The No. of instances includes the train, dev, and test sets for the Sem-Eval 2018 Task 1 (EI-Reg and V-Reg).

Emotion	**Train**	**Dev.**	**Test**
Valence	1181	449	937
Anger	1701	388	1002
Fear	2252	389	986
Joy	1616	290	1105
Sadness	1533	397	975

Table 4: The number of tweets in each of the dataset splits for the SemEval 2018 Task 1.

Emotion	**Learning Rate**	**No. Epochs**	**Batch Size**
Valence	3e-05	32	5
Anger	2e-05	32	10
Fear	3e-05	32	10
Joy	2e-05	16	5
Sadness	2e-05	32	10

Table 5: The optimal hyperparameter settings when fine-tuning the RoBERTa base model on the SemEval 2018 Task 1 dataset for each emotion.

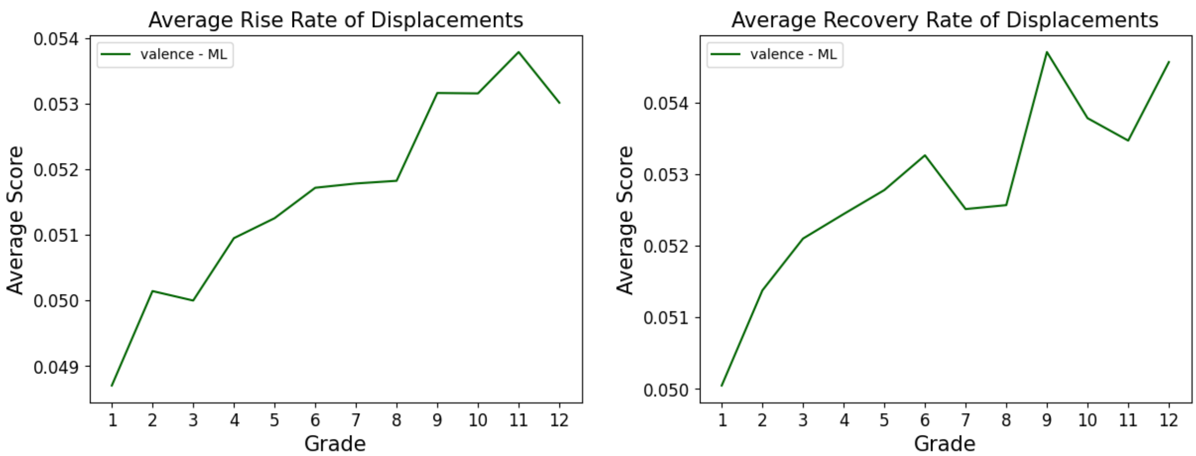


Figure 7: Rise rate and recovery rate for valence using the ML *n-gram* approach on the PoKi dataset.

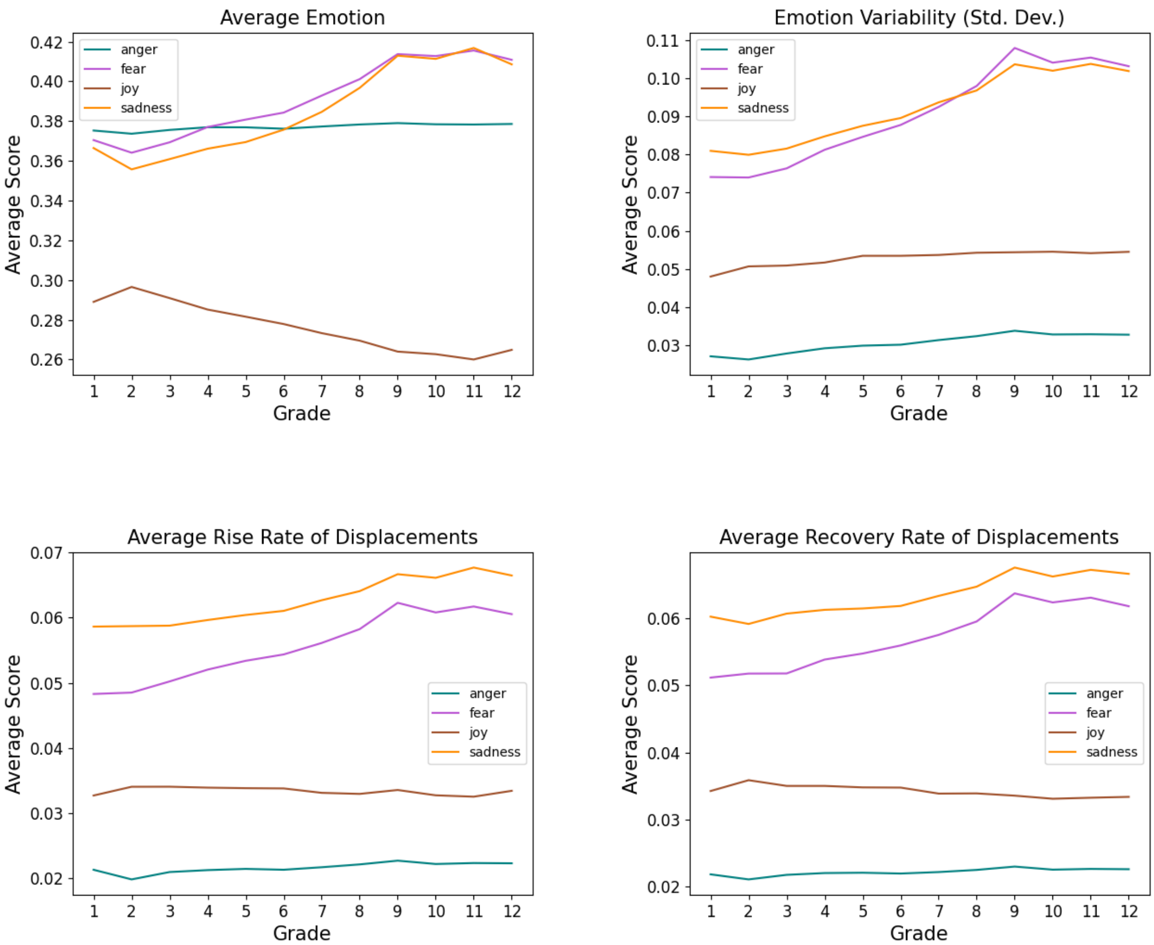


Figure 8: UED metrics for anger, fear, joy and sadness using the ML *n-gram* approach on the PoKi dataset.