Computing Emotion Dynamics from Text and Exploring their use as Biosocial Markers of Overall Well-being

by

Daniela Teodorescu

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

Department of Computing Science

University of Alberta

Abstract

Language is inherently social – it is influenced by our lived experiences and environments, and impacts the way in which we communicate with each other. Therefore, it is not a surprise that our health impacts our language. The patterns with which emotion changes over time – emotion dynamics – is a framework in psychology that is crucial for better understanding overall well-being and mental health. While emotion dynamics (ED) are commonly measured through self-report surveys, recent work has attempted to mitigate limitations of surveys (e.g., limited reach, bias, etc.) through inferring ED from utterances. The utterance emotion dynamics (UED) framework consists of four commonly used metrics, and we expand it by computing two additional metrics inspired by ED work in psychology, for a total of six metrics.

We investigate the relationship between UED and well-being in two domains – poems written by children across grades and data related to mental health. We compute UED metrics on poems written by children in grades 1 to 12 across seven emotions and find trends of emotion change across grades. For valence, these patterns of emotional change across grades are supported by previous work in psychology. Here, we further explored two other dimensions of emotions (arousal and dominance) and four categorical emotions (anger, fear, joy, and sadness), further contributing novel results for these emotions. We extend this work to characterize the UED of groups with a mental health diagnosis on Twitter[1]. Overall, our work builds upon the UED framework by

[1] Towards the end of this thesis Twitter was rename to X. We will continue to use the

computing new metrics and applying it to various domains. We demonstrate that patterns of emotional change in text can act as an indicator of overall well-being across domains.

term Twitter since it is well known.

Preface

This thesis consists of content from four papers:

1. “Utterance Emotion Dynamics in Children’s Poems: Emotional Changes Across Age”

   This paper is published in the 13th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis collocated at ACL 2023.

2. “Language and Mental Health: Measures of Emotion Dynamics from Text as Linguistic Biosocial Markers”

   This paper is currently under review at a conference.

3. “Evaluating Emotion Arcs Across Languages: Bridging the Global Divide in Sentiment Analysis”

   This paper is available on arXiv.

4. “Generating High-Quality Emotion Arcs For Low-Resource Languages Using Emotion Lexicons”

   This paper is currently under review at a journal.

The main content of this thesis consists of papers one and two. The Background chapter contains work on evaluating generated emotion arcs from papers three and four.

I computed utterance emotion dynamics metrics and performed the analyses described here. However, we used existing datasets [46, 116] and built upon the python framework created by Krishnapriya Vishnubhotla [125, 47]. Thank you to Tiffany Cheng who advised on the statistical analyses for the second paper and assisted with the literature review on emotion dynamics and their relationship to mental health diagnoses.

To my family, who are always there to support and care for me, and to Motek for all the snuggles and hugs which always put a smile on my face.

Acknowledgements

A big thank you to my supervisor Alona Fyshe, for her ongoing support and motivation since I was an undergraduate. I am forever grateful for the opportunity to work with you. I have learnt an amazing amount about research and all cool things related to the brain. Your kindness and brilliance are a model of the researcher I would like to be one day. I would like to thank Carrie Demmans Epp, while an at-arm's length examiner for my thesis, you have been an incredible mentor and inspiration during my research journey. I have learnt a great deal about the nitpicky details of presentations, papers, and communication. I appreciate your continued support whenever I have a question, regardless of the topic. I would like to thank Saif Mohammad who has been a tremendous collaborator and mentor over the past year and a bit. I am really grateful for the opportunity to work with you during an internship last year which has led to this incredible journey into the world of *emotion*. I appreciate your detailed feedback, and have learnt a lot about how to write and motivate research problems. I am excited that we will be collaborating in the years to come.

Thank you to Tiffany Cheng for her statistics expertise which helped guide our analyses in Chapter 4 and to Krishnapriya Vishnubhotla for the Emotion Dynamics code-base which set the ground work for computing utterance emotion dynamics.

I would also like to thank my fellow colleagues and CSGSA members for memorable experiences during my time here at the University of Alberta.

Words can not describe how thankful I am to my family for their love and support. It made the hard points a lot more bearable. Thank you for all that you do for me, I would not be where I am today without you.

Lastly, thank you to Amii, NSERC, Alberta Innovates, and DeepMind for your financial support which made a lot of this research possible.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Sections 1.1 and 1.4.2 are currently under review at a conference. Also, parts of Sections 1.2, 1.3, 1.4.1, have been published in Teodorescu *et al.* [118].

1.1 Language as a Biosocial Marker

Language is inherently *social*—from the way in which we say things, the expressions we use and the things we choose to share, to being impacted by our social environment and lived experiences. For centuries, language has been one of the primary means through which we communicate with each other and express ourselves. As our social environments have evolved over time, language has evolved to better support our communication needs and collaborative societies. Therefore, language is also *variable*, as the way in which we use it has adapted to cultures and communities around the world, and it is influenced by an individual's experiences.

Given the prominent role of language in human evolution from hunters–gathers to collaborative societies, and the large extent to which we rely on language today, it is not surprising that our mental health impacts our language usage. Quantitative features in language (e.g., aspects which can be measured) have already been shown to indicate and help clinicians monitor the progression of mental health conditions (MHCs), acting as a *biomarker*.

Linguistic biomarkers are quantitative features of language which have the potential to objectively predict the onset and progression of psychosis [73, 33]. Examples of linguistic biomarkers are: the proportion of pronouns used are an indicator of depression [59], syntax reduction was found in individuals with Anorexia Nervosa [20], lexical and syntactic features differed in those with mild cognitive impairment and dementia [12, 32], and semantic connectedness is an indicator for schizophrenia [18]. Also, the emotions expressed in text have been shown to correlate with mental health diagnoses. For example, more negative sentiment was found for individuals with depression [22, 98, 25]. Other work has shown that r/SuicideWatch, r/Anxiety, and self-harm subreddits had noticeably lower negative sentiment compared to other mental health subreddits such as Autism and Asperger's [35].

While language can be a biomarker for mental health, the social nature of language has implications. Notably, the variability in language use—especially across social groups—means that we should be skeptical about universal biomarkers, and instead realize that linguistic biomarkers alone are not capable of predicting MHCs. A vast amount of contextual and clinical information (often only available to individual's physician) helps determine well-being, and sometimes linguistic markers can aid the process. Further, linguistic biomarkers are more likely to be a stronger indicator among groups with commonalities such as region, culture, and medium of expression (e.g., social media platform). Therefore, a more appropriate term would be that language is a *biosocial* marker for health as it is influenced by social factors and biological factors [73]. For example, social factors such as parental socioeconomic status, neighborhood, and institutionalization (e.g., group foster care by government) influence speech markers such as lexical diversity; and social class influences syntactic complexity [73].

As language is increasingly being studied as a biosocial marker for mental health – accelerated by the ease and availability of Natural Language Processing (NLP) tools and language data online – there are important ethical implications. We must consider the sociolinguistic factors of such markers to ensure less biased and more accessible tools in clinics [73]. If social factors are not considered, then this limits the utility of systems and their ability to predict well-being as they may be capturing confounding variables.

1.2 Emotion Dynamics

In that context, our goal is to understand to what extent are patterns of emotion change a *biosocial* marker for mental health and well-being? *Emotion dynamics* studies the patterns with which emotions change across time and involves the "study of the trajectories, patterns, ... with which emotions ... fluctuate across time" [68, 69]. Emotion dynamics have been shown to correlate with overall well-being, mental health, and psychopathology (the scientific study of mental illness or disorders) [69, 52, 104, 112]. Further, studying emotion dynamics allows us to better understand emotion, and has been shown to have ties with academic success [38, 89], and social interactions (e.g., shyness) in children [111].

Emotion dynamics have been measured in psychology through self-report surveys over a period of time (e.g., 3 times a day for 5 days). Using these self-reports of emotion over time, various metrics can quantify how emotions change over time (e.g., the *average intensity*, *variability*, etc.). We discuss emotion dynamics in Chapter 2.

1.3 Utterance Emotion Dynamics

There are several drawbacks to self-report surveys. For example, limited amounts of data can be collected in terms of duration and reach, and self-reports are prone to biases [64]. Another window through which emotion dynamics can be inferred is through one's *utterances*. Inspired by emotion dynamics in psychology, Hipson and Mohammad [47] 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)*. In UED emotion arcs are generated from streams of text (e.g., sentences in a story, tweets over time, etc.)[1], which in turn can be used to calculate various metrics quantifying emotion change.

In order to generate an emotion arc, emotion scores must be assigned to instances of text (e.g., sentences, tweets). There are two ways in which emotion scores can be assigned to instances: using emotion lexicons or a machine learning (ML) based approach. Emotion lexicons contain word–emotion association scores for thousands of words across various emotions categories, which can be used to determine the average emotion score or the proportion of emotion words in the instance. Alternatively, the ML approach uses a neural network (often a deep learning model) to predict an emotion score per instance. Using emotion lexicons has the benefit of interpretability, accessibility, and efficiency compared to ML models. Thus, we primarily used a lexicon-based approach in our experiments. However, in Chapter 3 we also show that the use of ML models often produces similar trends as the lexicon approach.

There are several metrics in the UED framework inspired by those in psychology (e.g., average emotion, emotion variability, rise rate, and recovery rate) which capture different aspects of emotion change. UED metrics can be computed for a *speaker* over time (e.g., main character in a narrative, tweets

[1] An emotion arc is a series of timestep–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.) [79, 81, 93, 119, 120]. More details are in Chapter 2.

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). More details are in Chapter 2 (Section 2.3.4).

1.4 Our Approach

While emotion dynamics have been studied in psychology for decades, UED were proposed only recently and have been applied to only a small number of domains (literature and tweets). Ties between emotion dynamics metrics and psychological well-being have been shown using surveys in psychology, however it is not known whether this relationship similarly exists for emotion dynamics in *language* usage. To answer this, we explore two use-cases: poems written by children in grades 1–12 [46], and Twitter[2] posts by those who have self-disclosed online as having a mental health diagnosis [116]. Our work allows us to address important unanswered questions such as *how do UED metrics change over development from children to young adults?* and *can UED help provide indicators of mental health?*. Our goal is to apply existing NLP techniques to study patterns of emotion change in the context of emotional development and well-being.

1.4.1 Children's Writing and Poetry

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 [46]. Yet, careful and responsible work such as the work done on the Child Language Data Exchange System (CHILDES) [75] for understanding child language ac-

[2] Towards the end of this thesis Twitter was rename to X. We will continue to use the term Twitter since it is well known.

quisition can be tremendously influential. Similarly, applying UED metrics to children's writing will allow us to infer the typical emotional expression of children across ages. Such work provides important information for psychologists and child development specialists, as emotion dynamics have been shown to underlie well-being, psychopathology, and success [52, 111, 104, 112].

Poetry is a domain of growing interest in NLP (e.g., poem generation [124, 34, 36]). Poems are written to evoke emotions [127, 55] and a medium through which emotions are expressed [129, 8]. The intersection of poems and children's writing is a relatively unexplored area which has the potential to unlock patterns in emotion word usage by children as they age.

1.4.2 Social Media for Mental Health and Well-Being

Lastly, social media is increasingly used as a medium of communication with 4.88 billion users globally in 2023 [1]. Further, social media provides an abundance of data which is available for public health and well-being research. The World Health Organization reports approximately 970 million people globally (or 1 in every 8 people) suffered from a mental health condition in 2019, with numbers only rising drastically since the COVID-19 pandemic [2]. Moreover, in developed countries approximately 35.5% to 50.3%, and in less-developed countries 76.3% to 85.4%, of serious mental health cases go untreated [17]. This lack of treatment could be due to lack of access to resources, stigma surrounding mental health, and more. Social media provides the perfect opportunity to better understand well-being at a larger scale across various languages, geographic regions, and cultures. By studying patterns of emotion change for groups who disclose as having a mental health condition, we see if UED metrics can act as important indicators of well-being. We use a Twitter dataset of users who self-disclose as having a mental health diagnosis online and a control group [116]. For each group of users, the dataset contains their tweets over four years, allowing us to study characteristics of emotion change over time.

To summarize, in this thesis we contribute to the knowledge of patterns of emotion change over time in *language*. More specifically, we explore two domains: poems written by children across grades, and Twitter posts for groups of individuals who have self-disclosed as having a mental health diagnosis. We will show that there are meaningful trends in UED metrics across age and mental health groups, which points to the potential of UED as indicators of emotional development and well-being. The broader implications of further research down this path is that the relationship between UED metrics and overall well-being can be explored across various domains and groups, further contributing to the understanding of UED as *biosocial* markers. We discuss these more in Chapters 3 and 4 respectively. In the upcoming Chapter we review relevant background materials on emotion arcs, emotion dynamics, and utterance emotion dynamics.

Chapter 2

Background

Sections 2.1 and 2.2 in this chapter are currently under review at a conference. Also, Tiffany Cheng helped with some of the writing on the relationship between average emotional state, variability, and reactivity with well-being in Section 2.2.2.

In this Chapter we discuss the relevant background research on tracking emotion change across time through *emotion arcs*, and computing patterns of emotion change through *emotion dynamics*. Afterwards, we look at the associations between emotion dynamics with well-being in psychology and the current literature on studying well-being in children's writing. Lastly, we review how emotion dynamics can be inferred from text (*utterance emotion dynamics*).

2.1 Emotion Arcs

Commercial applications as well as research projects often benefit from accurately tracking the emotions associated with an entity (e.g., person, company, object, etc.) over time. Public health researchers are interested in analyzing social media posts to better understand population-level well-being [125], loneliness [43], and depression [24]. Government Policy makers benefit from tracking public opinion over time to develop effective interventions and laws.

One such example of this is tracking sentiment towards health interventions such as mask mandates and vaccine policies [53]. Researchers in Digital Humanities are interested in understanding basic components of stories such as plot structures, how emotions are associated with compelling characters, and categorizing stories based on emotion changes [93]. In all of these applications, the goal is to determine whether the degree of a particular emotion has remained steady, increased, or decreased from one time step to the next. The time steps of consideration may be days, weeks, years, etc. This series of timestep–emotion value pairs, which can be represented as a time-series graph, is often referred to as an *emotion arc* [57, 79, 93].

When the amount of data involved is large enough that human annotations of emotions are prohibitive, then one can employ automatic methods to estimate emotion arcs. Automatic methods will be less accurate than human assessments but can be applied at scale and easily adjusted to changes in needs (e.g., tracking new, different, or additional entities of interest).
The input to these systems are usually:

- The text of interest where the individual sentences (or instances) are temporally ordered; possibly through available timestamps indicating when the instances were posted/uttered: e.g., all the tweets relevant to (or mentioning) a government policy along with the date and time of posting.

- The emotion dimension/category of interest: e.g., anxiety, anger, valence, arousal, etc.

- The time step granularity of interest (e.g., days, weeks, months etc.)

The automatic methods usually follow the steps listed below to determine the timestep–emotion value pairs (the emotion arc):

- Suitable pre-processing of the text (e.g., converting text to lowercase, removing urls and numbers, tokenizing text, etc.).

- Apply emotion labels to units of text. Two common approaches for labeling units of text are: (1) *The Lexicon-Only (LexO) Method:* to label words using emotion lexicons and (2) *The Machine-Learning (ML) Method:* to label whole sentences using supervised ML models (with or without using emotion lexicons). More details on the *LexO* and the *ML* method are in Section 2.1.1.

- Aggregate the information to compute timestep—emotion value scores; e.g., if using the LexO approach: for each time step, compute the percentage of angry words or average valence of the words in the target text pertaining to each time step, and if using the ML approach: for each time step, compute the percentage of angry sentences or average valence of the sentences for each time step.

The time steps used can be non-overlapping, e.g., months of a year, but they can also be overlapping, e.g., ten-day time steps starting at every day of a year. Here, every adjacent time step has nine overlapping days. Overlapping steps produce smoother emotion arcs, and thus are preferred in some applications. The number of textual instances (usually sentences or tweets) pertaining to a time step are referred to as the *size of the time step* or *bin size*. If the data does not come with associated timestamps, but simply a temporal order from beginning to end (such as the text in a novel), then often a time step is a pre-chosen fixed amount of textual units (e.g., 200 words, 100 sentences, or one chapter). Thus, bins can be created to have the same size based on the chosen amount of textual units.

2.1.1 The Lexicon-Only (LexO) and Machine-Learning (ML) Method

Emotion lexicons can be thought of as lists of thousands of words and their associated emotion scores. Scores can be fine-grain (e.g., continuous number from zero to one) or categorical (e.g., -1, 0, 1). Further, emotion lexicons exist for several emotion dimensions such as valence, arousal, dominance or categories such as anger, fear, joy and sadness. Emotion lexicons such as the NRC Valence, Arousal, Dominance [82] or the NRC Emotion Intensity Lexicon [83] can be used to assign emotion scores to words. Words not in the lexicon can be assigned a neutral score or disregarded. Experiments exploring both of these approaches are detailed in Teodorescu and Mohammad [120]. Since individual words have emotion scores, the emotion score for an instance can be represented as the average of the word emotion scores. Then a time step emotion score can be computed as the average of the instance emotion scores within the time step. This approach is depicted in Figure 2.1.

The ML method uses deep learning models for sentiment analysis. In this scenario emotion scores are predicted for a sentence or phrases of words. This differs from the LexO method as often the whole instance can be assigned an emotion score and then one can simply compute the average of the emotion scores for all instances in the time step.

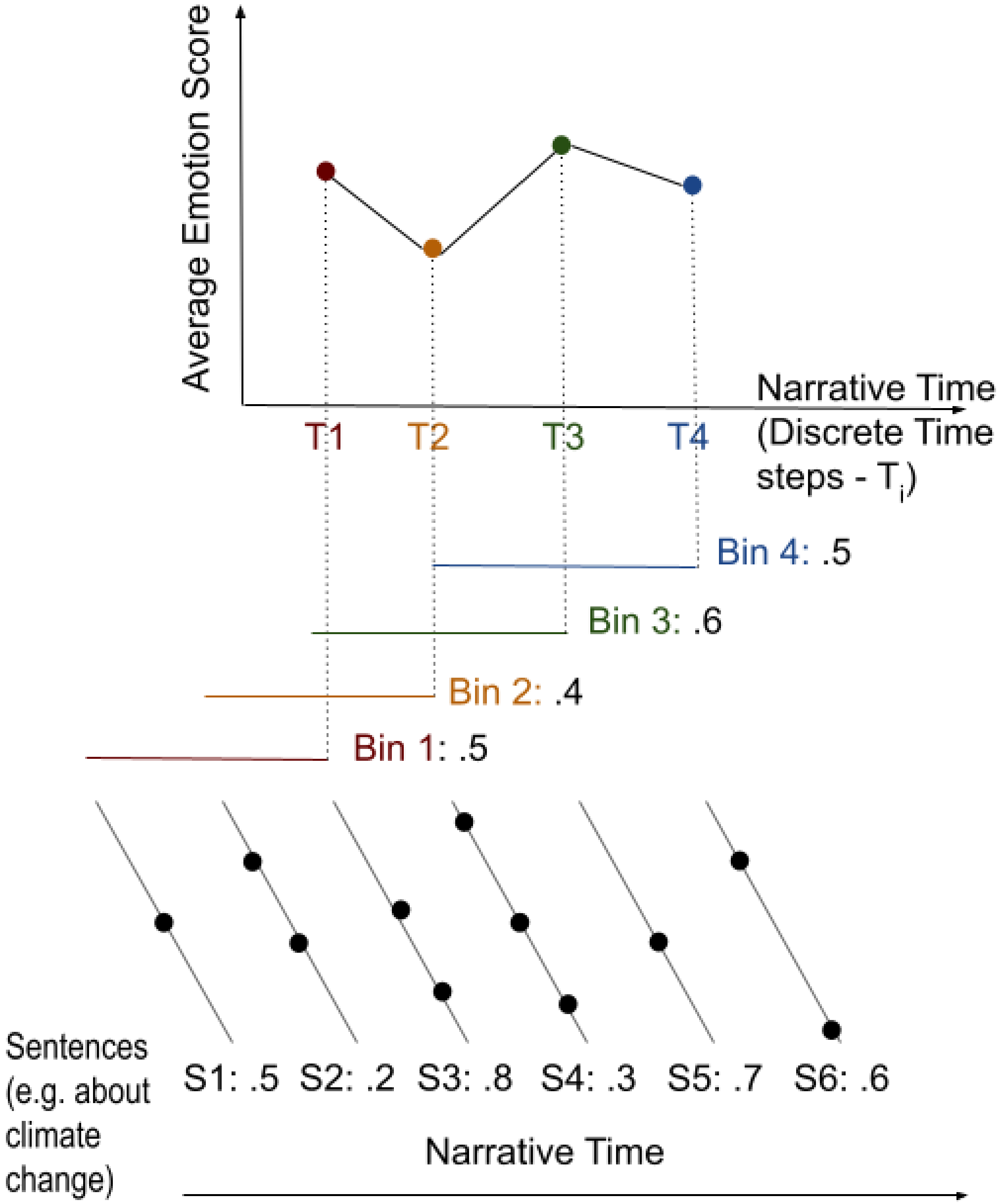


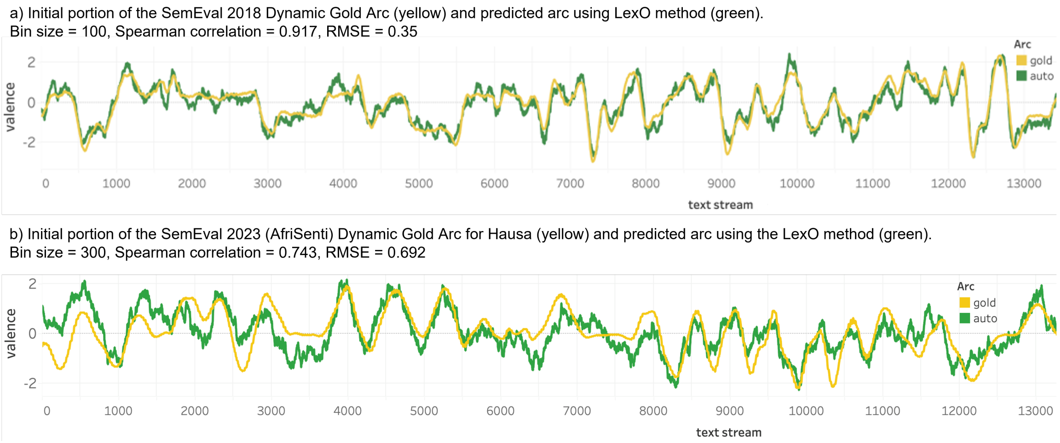
Figure 2.1: Generating an emotion arc using the LexO method.

2.1.2 Related Work

Emotion arcs have commonly been created from literary works and social media content. Alm and Sproat [3] were the first to automatically classify sentences from literary works for emotions using a machine learning paradigm. Mohammad [79] was the first to create emotion arcs and analyze the flow of emotions across the narrative in various novels and books using emotion lexicons. Kim *et al.* [58] built on this work by creating emotion arcs to determine emotion information for various genres using the NRC Emotion Lexicon. Reagan *et al.* [93] and Del Vecchio *et al.* [26] clustered emotion arcs and found evidence for six prototypical arc shapes in stories. Hipson and Mohammad [47] analyzed emotion arcs for individual characters (instead of the whole narrative) in movie dialogues. Emotion arcs of literary works have been used to better understand plot and character development [47, 58, 93]; and also for practical applications such as assisting writers develop and improve their stories [6, 107].

Despite the wide-spread use of emotion arcs in industry and research, there is surprisingly little work on evaluating generated emotion arcs. A key reason for this is that it is hard to determine the true (gold) emotion arc of a story from data annotation. It is hard for people to read a large amount of text, say from a novel, and produce an emotion arc for it. One attempt to evaluate aspects of an emotion arc can be seen in Bhyravajjula *et al.* [9]. They asked one volunteer to read mini-segments of a *'The Lord of the Rings'* novel to determine whether the protagonist's circumstance undergoes a positive or negative shift. They then determined the extent to which the automatic method captured the same shifts.

Emotion arcs commonly employed in commerce and social media research are fundamentally different from the arcs in novels. For products, researchers



a) Initial portion of the SemEval 2018 Dynamic Gold Arc (yellow) and predicted arc using LexO method (green).
Bin size = 100, Spearman correlation = 0.917, RMSE = 0.35

b) Initial portion of the SemEval 2023 (AfriSenti) Dynamic Gold Arc for Hausa (yellow) and predicted arc using the LexO method (green).
Bin size = 300, Spearman correlation = 0.743, RMSE = 0.692

Figure 2.2: Gold and predicted arcs for English (a) and Hausa (b) using the LexO method. This figure is from Teodorescu and Mohammad [120].

are often interested in the arcs associated with posts that mention a product, such as a certain brand of cellphone [88], government policy [106], a person [16], or entity such as Uber [100]. The gold emotion arc can be considered to simply be the average of the manually annotated emotion scores of all the tweets pertaining to the time steps. For example, a gold emotion arc can be generated by averaging human-annotated emotion scores in tweets about the iPhone posted every day[1]. Using this approach, Teodorescu and Mohammad [119] performed experiments on 36 diverse datasets to show that the quality of emotion arcs generated using emotion lexicons is comparable to those generated using ML methods. The lexicon approach performs 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 an English lexicon into low-resource languages, such as indigenous African languages [120]. An example of the predicted arc using the LexO method and the gold arc for both English and Hausa are shown in Figure 2.2. The predicted arc follows the gold arc closely, even when the gold arc has emotion

[1] This averaging of scores of instances may also be applied to character dialogues in novels, but it should be noted that the character's true emotional state is determined through other literary aspects as well, such as author narration and prior character development.

surges and dips of varying widths and heights. These works show that emotion arcs can be accurately generated using emotion lexicons which have the benefit of interpretability, accessibility, and efficiency compared to ML models.

2.2 Emotion Dynamics

Emotions play a key role in overall well-being [69, 51, 104, 112]. People's emotional states are constantly changing in response to internal and external events, and the way in which we regulate emotions also changes throughout child development [131, 77]. 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) [51, 112, 97, 102, 40], academic success [38], and social interactions in children (e.g., shyness) [111].

Several psychopathology studies have introduced metrics to quantify and understand the trajectories and patterns in emotions across time [69]. 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, inertia) [69, 121]. Affective chronometry is a field of research that examines the temporal properties of emotions (i.e., emotion dynamics) and increasingly its relation to mental health and well-being have been studied [64].

2.2.1 Emotion Dynamics Metrics

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).

There are several metrics in the emotion dynamics framework: *emotion intensity*, *emotion variability*, *emotion reactivity*, *emotion regulation*, *emotional inertia*, and *emotional instability*. Each of these metrics can be define as follows:

- **Emotion intensity** is the average emotion over time.

- **Emotion variability** is the variance, or the range of amplitude of emotional states in terms of its intensity over time [64, 52]. This variation can occur over multiple timescales (e.g., seconds, hours, days, weeks [64]).

- **Emotion reactivity** is the emotional response to an event (where individuals will have varying intensities), the time to reach the peak response, the maintenance time of the peak response, and the recovery time to baseline (i.e., emotion regulation) [21, 96]. This process can be seen through the emotional episode in Figure 2.3.

- **Emotion regulation** is the management of emotional responses to situations [56]. This involves attending to and activating processes to modulate emotional experiences (whether effortful or implicitly) by initiating, maintaining, or modulating the intensity, duration or occurrence of feelings [103, 122]. Work in psychology has proposed emotion regulation consists of three stages: identification, selection, and implementation [39]. While emotion reactivity and regulation interact with each other, they are distinct processes [101].

- **Emotional inertia** is how well the intensity of an emotional state can be predicted from a previous emotional state [65, 117]. It can also be thought of as the resistance to emotional changes over time [64]. This can lead to carryover effects of emotional experience despite changes in the external and internal environment [69].

- **Emotional instability** is the magnitude of emotional changes from one moment to the next [52]. Higher instability means experiencing higher emotional shifts from one moment to the next [52].

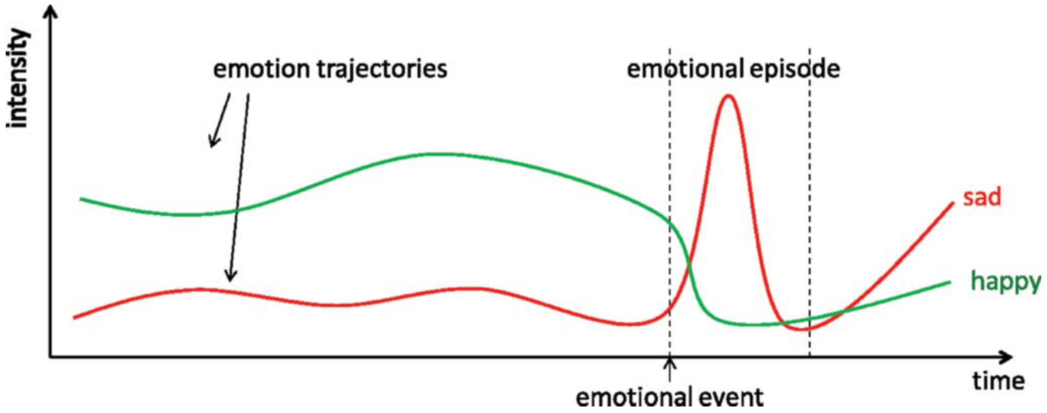


Figure 2.3: Emotional reactivity and regulation can be seen in the above emotional episode. When an emotional event occurs, an emotion whether it be happiness or sadness reaches a peak state (irrespective of direction), and then emotion regulation strategies help modulate the response back to baseline levels. This figure is from Kuppens and Verduyn [68].

2.2.2 Emotion Dynamics and Ties with Well-Being in Psychology

The relationship between various metrics in emotion dynamics and well-being have been the topic of numerous psychology studies. Below we describe how each of the following emotion dynamics metrics have been proposed to be predictive of distinguishing between affective disorders [64]. Moreover, specific emotion dynamic patterns can contribute to maladaptive emotion regulation strategies and poor psychological health [52]. The meta-analysis by Houben *et al.* [52] has indicated that the timescale of emotion dynamics does not moderate the relation between emotion dynamics and psychological well-being. Therefore, the relationship between psychological well-being and emotion dynamics occurs whether it is examined over seconds, days, months, and so on.

Average Emotional State & Psychopathology: Average or baseline emotional states are related to well-being and mental illnesses. Due to the maladaptive (i.e., dysfunctional) nature of psychopathology, those with mental illnesses tend to have more negative emotional baselines compared to those without. For example, Heller *et al.* [45] found that a higher average positive affect is associated with lower levels of depression but not anxiety, and a higher average negative affect was related to increased depression and anxiety. As well, those with post-traumatic stress disorder (PTSD) have reported lower average positive affect [91].

Emotion Reactivity, Regulation & Psychopathology: Research has found that individuals with psychopathologies tend to take longer to recover from differing emotional states (i.e., emotional resetting or recovery rate) than healthy individuals [64, 103]. Houben *et al.* [52] also proposed that high emotional reactivity and slow recovery to baseline states is a maladaptive emotional pattern indicative of poor psychological well-being and psychopathology. In other words, people with poor psychological health may be highly reactive, emotionally volatile, and take a longer time to return to a baseline state.

Hofmann *et al.* [49] showed that mood and anxiety disorders are a result of emotion dysregulation of negative emotions, along with lacking positive emotions. Moreover, emotion dysregulation is thought of as the core of anxiety disorders [14], [78]. Children with anxiety disorders had higher negative emotion reactivity, and were less successful at implementing emotion regulation strategies [14].

Emotion Variability & Psychopathology: The Houben *et al.* [52] meta-analysis findings also indicate that higher emotional variability is related to lower psychological well-being. In particular, variability was positively correlated with depression, anxiety, and other psychopathologies (e.g., bipolar, borderline personality disorder, etc.). This is supported by Heller *et al.* [45] who found that higher positive and negative affect variability was associated with higher levels of depression and anxiety, these effects persisted for anxiety even after controlling for average positive affect. On the other hand, variability was no longer associated with depression after controlling for average affect and the rate of recovery to baseline. This effect was attributed to anhedonia (the reduced ability to feel pleasure) which is a common symptom of depression that leads to reduced emotionality.

Emotional Inertia & Psychopathology: Houben *et al.* [52] found a significant negative correlation between emotional inertia and psychological well-being; the more inert one's emotions were, the lower one's psychological well-being. These trends even occurred after controlling for differences in emotional contexts (e.g., emotional stimuli in the lab) [63]. That is, difficulty moving between emotional states is associated with lower psychological well-being. Inertia is thought to be related to cognitive preservative tendencies such as rumination [61] and inefficient regulatory strategies such as suppression [60]. Therefore, by resisting change and failing to regulate emotions, it is not a surprise that inertia is associated with poor psychological adjustment. Inertia was also found to be positively associated with "negative emotionality and depression, marginally positively related to borderline personality disorder, and negatively related to positive emotionality, eudaimonic well-being (e.g., purpose for doing meaningful things), and other indicators of high psychological well-being" [52]. No significant association was found with anxiety symptoms [52].

Emotional Instability & Psychopathology: Emotional instability and variability share similar associations with psychological well-being. A significant negative correlation was found between emotional variability, instability, and psychological well-being, so the more variable, unstable one's emotions are the lower one's psychological well-being [52]. Variability and instability are positively related to negative emotionality, externalizing behavior, depression, bipolar/mania symptoms (only marginally for instability), anxiety symptoms, borderline personality disorder, and negatively related to eudaimonic well-being and other indicators of high psychological well-being [52]. Moreover, emotional variability and instability are positively related to negative affect, neuroticism, minor depression diagnosis, depression, anxiety symptoms, borderline personality disorder, and are negatively related to self-esteem (only marginally for instability measures), satisfaction with life, and other indicators of high psychological well-being [51].

Overall, emotion dynamics research suggests that one's average emotional state, emotional variability, reactivity, regulation, inertia, and instability vary by mental health diagnoses. Preliminary research suggests that these metrics may also vary across different mental illnesses or psychopathologies.

2.3 Utterance Emotion Dynamics

Studying emotion dynamics through self-reports is arduous work; limiting the amount of data collected both in terms of sample size and study length. Further, self-reports are prone to a number of biases (e.g., social pressures to be perceived as being happy) [64]. Another window through which emotion dynamics can be inferred is through one's utterances [47]. The Utterance Emotion Dynamics (UED) framework uses various metrics inspired by psychology research to measure patterns of emotion change from the emotion expressed in text. Using a person's utterances allows researchers to analyze emotion dynamics since one's utterances can reasonably reflect one's thought process. UED allows for broader scale analyses across mediums (e.g., narratives, social media, etc.) and regions (e.g., cities, countries, etc.). Although proposed recently, UED metrics have been used to study the emotional trajectories of movie characters [47] and to analyze emotional patterns across geographic regions through Twitter data [126]. Seabrook *et al.* [99] studied the association between depression severity and emotion dynamics metrics such as variability and instability on Facebook and Twitter. It was found that increased negative emotional instability was a significant predictor for greater depression severity on Facebook, and that increased negative emotional variability was an indicator for lower depression severity on Twitter.

2.3.1 Metrics

There are several UED metrics:

- **Average Emotional Intensity**: One's average emotion over time. This is computed as the average of emotion intensity scores over time.

- **Home base**: The steady (most common) state where one's emotional intensity is on average in the emotional space. This is computed as one standard deviation above and below the Average Emotional Intensity.

- **Emotional Variability**: How much one's emotional state changes over time. This is computed as the standard deviation for changes in emotion.

- **Rise Rate**: The rate at which one reaches peak emotional intensity (i.e., emotional reactivity). 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. Often the average rise rate is considered, which is simply the average of the rise rates in a text.

- **Recovery Rate**: The rate at which one recovers from peak emotional intensity to home base, (i.e., emotional regulation). Recovery rate is computed similarly to rise rate, however divides peak distance by the number of words during the recovery period (e.g., where emotion intensity starts to move towards baseline levels). Likewise, recovery rate does not distinguish between peak direction.

2.3.2 New Metrics

Inspired by emotion dynamics in psychology, we propose two additional metrics to the UED framework: instability and inertia. They are defined as follows based on the definitions in psychology.

- **Emotional Instability** captures the amplitude of moment-to-moment changes in emotion similarly to emotion variability, however it is time-structured, quantifying differences between consecutive observations of emotion [52]. Therefore, it is sensitive to the time ordering of observations [98]. So the higher the instability, the higher the variance and less positively related are observations [98]. Instability has been operationalized as the mean squared successive difference (MSSD) [86]. The formula for MSSD is as follows, where x_i is the emotion score at time step i, and n is the total number of time steps in the series:

  $MSSD = \frac{\sum_{i=1}^{n-1}(x_{i+1}-x_i)^2}{n-1}$

- **Emotional Inertia** captures how well a previous emotional state predicts the next emotional state [65, 117]. The higher the inertia, the more temporal dependency between observations [98]. Inertia has been measured through the autocorrelation coefficient (ACF). Autocorrelation can be thought of as the correlation between a time series and the time-lagged version of itself by lag k. The ACF at time lag k can be computed as follows, where $Y_1, ..., Y_N$ are emotion scores at time steps $X_1, ..., X_N$, and N is the total number of time steps in the series:

  $ACF_k = \frac{\sum_{i=1}^{N-k}(Y_i-\overline{Y})(Y_{i+k}-\overline{Y})}{\sum_{i=1}^{N}(Y_i-\overline{Y})^2}$

Altogether, both the existing and new UED metrics are inspired by constructs in psychology - we show this mapping in Table 2.1. With the addition of these new UED metrics we can test if they too provide meaningful indicators for patterns of emotion change in text, as they do for emotion dynamics in psychology.

UED Metric	**Psychology Construct**
Average Emotion	Emotion Intensity
Emotional Variability	Emotional Variability
Rise Rate	Emotional Reactivity
Recovery Rate	Emotional Regulation
Emotional Instability	Emotional Instability
Emotional Inertia	Emotional Inertia

Table 2.1: UED metrics and the corresponding construct in psychology.

2.3.3 Time-Aspect of UED metrics

We compute UED metrics by abstracting away time and considering only the temporal ordering of utterances. We consider only the ordering between instances since emotion dynamics in psychology has been shown to be robust to scales of time (e.g., minutes, hours, days, etc.) [52]. However, if one were to consider the relative time difference between points in the series, the above formulas (Sections 2.3.1 and 2.3.2) could account for this. For example, the average emotion intensity over time could assign a higher weight to points that are closer together in time and assign a lower weight for points farther away in time. Another example is computing *time-adjusted* emotional instability as defined by Jahng *et al.* [54] below, where x_i is the emotion score at time step i, n is the total number of time steps in the series, t_i is the timestamp at point i, and $median(\Delta t)$ is the median of incremental time differences across the time-series:

$$Time - adjustedMSSD = \frac{median(\Delta t)}{(n-1)} \sum_{i=1}^{n-1} \frac{(x_{i+1}-x_i)^2}{(t_{i+1}-t_i)}$$

A higher time-adjusted instability means that there is a larger change in emotion expressed between all consecutive pairs of observations relative to their median time difference [98].

While it would be interesting to explore whether different time scales impact UED, that is outside of the scope for this thesis.

2.3.4 UED Levels

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 levels of UED metrics are described below:

- **Speaker UED Metrics:** The UED metrics for a speaker are determined from placing all the utterance of a speaker in temporal order and computing UED on these ordered utterances. If population UED metrics are of interest, the scores for a particular metric can be averaged across multiple speakers. In the past, speaker UED metrics were determined for characters in movie dialogues [47], and for users on Twitter during the pandemic [126].

- **Meta-Speaker UED Metrics:** We may be interested in analyzing the change of emotions in a discourse between multiple speakers, for example, over a Reddit thread. In this case we can treat each discourse (e.g., each Reddit thread) as text produced by a meta-speaker (all users contribute to the thread). Here we arrange each of the utterances in a discourse (Reddit thread) in temporal order and determine the UED metrics over the ordered utterances. UED metrics can be averaged to determine the average UED metric scores for a set of discourses. In the past, discourse UED metrics have been computed for users from geographic regions. For example, past work has treated all users on Twitter in a country as a speaker [126].

- **Instance UED Metrics:** In this work we propose a third form of

UED metrics not explored before. If one is interested in the change of emotions in individual (largely independent) pieces of text (or instances) such as novels, poems, tweets, or blog posts, we can simply apply the UED metrics to each instance and average the scores across instances to determine average instance UED metrics.

2.4 Children's Writing and Well-Being

Little work studies the emotion dynamics in children's writing due to the limited data available. One of the most commonly known datasets is the Child Language Data Exchange System (CHILDES) [75] and in French, E-CALM [27]. This dataset is comprised of audio and transcripts of conversations between children and parents or teachers. These datasets are limited in that they contain parent-child dialogue for children approximately age one to seven, and have limited quantities of text.

Moreover, very few works look at emotions in children's writing. Manabe *et al.* [76] 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. Perhaps counter-intuitively, the researchers found that youth who had higher tendencies toward psychological distress used significantly more positive words, and therefore had higher valence.

2.5 Social Media and Well-Being

Recently, social media platforms have become a commonly used means of communication world-wide. Social media provides an abundance of data with relative ease on a variety of topics, including informal everyday conversation.

This has made it an avenue through which researchers can study mental health and overall well-being. Applying computational linguistic techniques to such large datasets has allowed researchers to study a wide range of questions in the public health domain, including those on the relationship between mental health and language [59, 22, 25, 35]. This combination of NLP and public health on social media data has become a growing field over the past decades, aiming to help individuals struggling with mental health and provide insights for physicians.

NLP work for well-being on social media data has largely focused on creating systems for predicting mental health status using either simple machine learning models or deep learning models. In a literature review by Chancellor and De Choudhury [15] (focusing on papers for predicting mental health status using social media), it was found that commonly studied social networking sites include Twitter, Reddit, Facebook, and Weibo. Chancellor and De Choudhury [15] reported that the most commonly studied disorder was depression, followed by suicide, and then schizophrenia, eating disorders, and anxiety with similar counts. Studies had varying designs and methods such as for:

1. establishing annotations for ground truth of mental health status

2. obtaining control data (users or dialogue without a mental health diagnoses)

3. ensuring data quality and sampling techniques

4. algorithmic techniques used in modelling

5. features selected for prediction (e.g., language, behavioral, emotion and cognition, demographic, and image features).

A first landmark study in this field was work by De Choudhury *et al.* [24] which built a classifier to predict whether Twitter users had depression based on features of their posts such as social activity, emotion expressed in text, social networks, self-attention, and more. Since then, there have been numerous works in this line of research and various datasets have been developed for studying language and mental health (as detailed by Harrigian *et al.* [44]).

There have been many studies examining the relationship between emotions in text and mental health (which this thesis is specifically more interested in). Due to the large number of studies in this area we do not detail them all, but summarize that across studies and social media domains (e.g., Twitter, Reddit, Facebook), it was found that posts with generally lower positivity or happiness were associated with mental health diagnoses such as depression, self-harm, anorexia, post-partum depression and more [22, 98, 25, 35, 123, 4, 23].

2.6 Summary

Overall, in this chapter we examined background material on emotion arcs and how to create both a gold and predicted *emotion arc*; emotion dynamics in psychology, the various metrics through which emotion dynamics are measured, and their association with psychological well-being; how emotion dynamics can be inferred through text with the utterance emotion dynamics framework and the various ways in which UED can be computed; how little work explore how emotion dynamics in children's writing; and lastly how social media data provides opportunities and another medium through which mental health can be studied.

Chapter 3

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

This chapter has been published at the 13th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis collocated with ACL 2023.

In this chapter 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 and dimensions of 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), recovery rates (i.e., emotional regulation), inertia, and instability with age. These results act as useful baselines for further research in how patterns of emotions expressed by children change with age, and their association with mental health.

3.1 Research Questions

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 *utterance emotion dynamics* (UED), building on work by Hipson and Mohammad [46]. 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 does the fluidity of emotional states from one moment to the next (*inertia*) change across grade? How well a previous emotional state predicts the next is informative of emotion regulation strategies which in turn influences psychological well-being.

- How variable are emotions with respect to time (*instability*) across grade? Higher emotional instability has been associated with lower psychological well-being.

- How do *utterance emotion dynamics* compare for adults vs. children?

These metrics are further described in Section2.3 in the Background Chapter. 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 [46] to calculate various UED metrics and examine how they vary across age. The scores for the metrics

Dataset	# of Poems	#Words per Poem (*SD)*
PoKi	61,330	14.3 (13.0)
Grade 1	900	37.3 (37.7)
Grade 2	3,174	32.1 (23.4)
Grade 3	6,712	35.2 (26.0)
Grade 4	10,899	39.3 (27.9)
Grade 5	11,479	44.5 (35.6)
Grade 6	11,011	49.6 (39.6)
Grade 7	7,831	59.7 (46.0)
Grade 8	4,546	67.6 (53.6)
Grade 9	1,284	91.5 (80.7)
Grade 10	1,171	91.8 (80.3)
Grade 11	667	103.0 (104.0)
Grade 12	1,656	97.2 (106.0)
FPP	50	181.02 (199.3)

Table 3.1: Number of poems and the average lengths of poems with the standard deviation in parentheses for the *PoKi* dataset (by grade) and the *FPP* dataset.

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.

3.2 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 3.1 shows key statistics of each dataset.

Poems Written by Children (PoKi): Hipson and Mohammad [46] 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.[1] In the PoKi dataset each poem is released with the child's school grade and first name. We expect grade to be highly correlated with age.

[1] Hipson and Mohammad [46] obtained permission to use these poems for research.

The average emotional patterns for emotion dimensions (valence, arousal, and dominance), along with discrete emotions (anger, fear, joy and sadness) were analyzed 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 [28, 109]. 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 [46] also compiled and used poems written by adults which were published on the *Famous Poets and Poems* website.[2] 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[3] library. Additionally, stop words were removed.

3.3 Experiments

Our goal is to analyze how patterns of emotion words change with age in children's poems. In order to do so, we generated an emotion arc per poem and compute instance-level UED metrics (as described in Section 2.3.4). In this work we are interested in children's poems (instances of poetic text) across age

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

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

and not how each individual child has a different writing style. Afterwards, we average the UED metrics per grade to compare results across age. We use the Emotion Dynamics toolkit[4] to calculate UED metrics (average emotion, variability, rise and recovery rate) and implemented the two new metrics (inertia and instability). Our code for the experiments is available online.[5]

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. Therefore, a time step corresponds to a text window of five emotion words in a poem, where adjacent time steps have 4 overlapping emotion words between them. We only considered poems that included at least five emotion words[6]. For each research question we computed the corresponding metrics: average emotional state, emotional variability, rise rate, recovery rate, inertia and instability. Analyzing average emotion and variability allows us to build foundational knowledge into changes in patterns of emotion words. We then look at rise rate, recovery rate, inertia, and instability 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.[7] 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,

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

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

[6] as per the NRC VAD lexicon

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

UED Metric	**Val.**	**Arousal**	**Dom.**	**Anger**	**Fear**	**Joy**	**Sadness**
Average	0.228	-0.247	-0.087	0.032	0.045	0.063	0.045
Variability	0.219	0.182	0.167	0.045	0.057	0.063	0.057
Rise Rate	0.134	0.114	0.084	0.106	0.113	0.090	0.075
Recovery Rate	0.127	0.105	0.086	0.033	0.046	0.027	0.028
Inertia	0.513	0.503	0.392	0.723	0.725	0.629	0.691
Instability	0.027	0.020	0.019	0.001	0.002	0.003	0.002

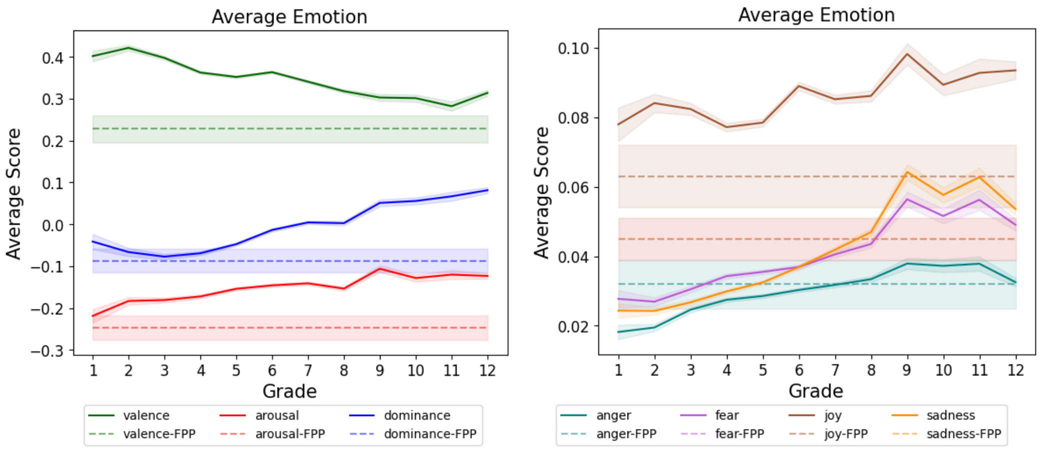
Table 3.2: The values for UED metrics in poems written by adults.

arousal, dominance) and discrete emotions (e.g., anger, fear, joy and sadness). We used the NRC Valence, Arousal, and Dominance (VAD) Lexicon [82] and the NRC Emotion Intensity Lexicon [83] for word-emotion scores.

In Section 4.4 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 3.3.2 we explore the same questions using an ML model for generating emotion arcs. We find similar trends across age with the ML approach as when using the lexicon approach.

3.3.1 Utterance Emotion Dynamics: PoKi

We begin with a question on how average emotion word score changes with grade–a question that Hipson and Mohammad [46] 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 3.2 and contrast them to PoKi below.



(a) Valence, arousal and dominance (b) Anger, fear, joy, and sadness

Figure 3.1: Average emotion across grades. The horizontal dashed lines represent values in poems written by adults. The shaded region around each line represents the standard error of the mean.

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: In text 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 3.1a, we show the average VAD dimensional 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 grade. 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 3.1b, 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 emotion dimensions tend to have a peak around grade 9 and plateau afterwards.

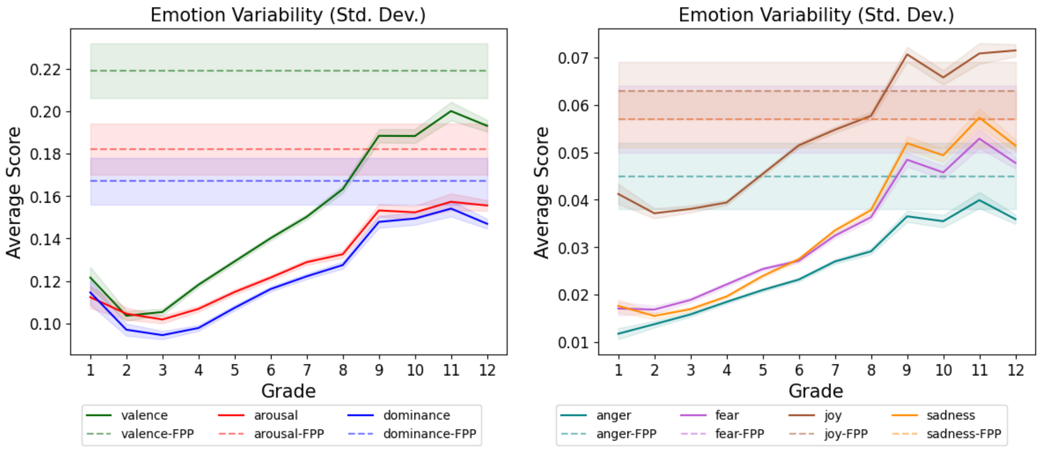
FPP Anger, Fear, Joy, Sadness: Anger, fear and sadness tends to match to those of older children around grade 8 to 9. Children from grade 9 to 12 reach even higher values than adults for anger, fear, and sadness. On the other hand, joy tends to remain below those of children across all ages, and has the most similar values to younger children at 0.063.

Discussion: These findings align with those by Hipson and Mohammad [46] which similarly computed the mean emotion in poems across grade. Numerous works in psychology have found similar trends through self-report studies for valence [31, 70, 105, 128], and arousal [13, 42, 108]. Likewise, as sadness increased with age, Holsen *et al.* [50] 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.

How variable are emotions across age?

Method: Variability is computed as the standard deviation of emotion values for windows in a poem.

Results:



(a) Valence, arousal and dominance

(b) Anger, fear, joy, and sadness

Figure 3.2: Emotional variability across grades. The horizontal dashed lines represent values in poems written by adults. The shaded region around each line represents the standard error of the mean.

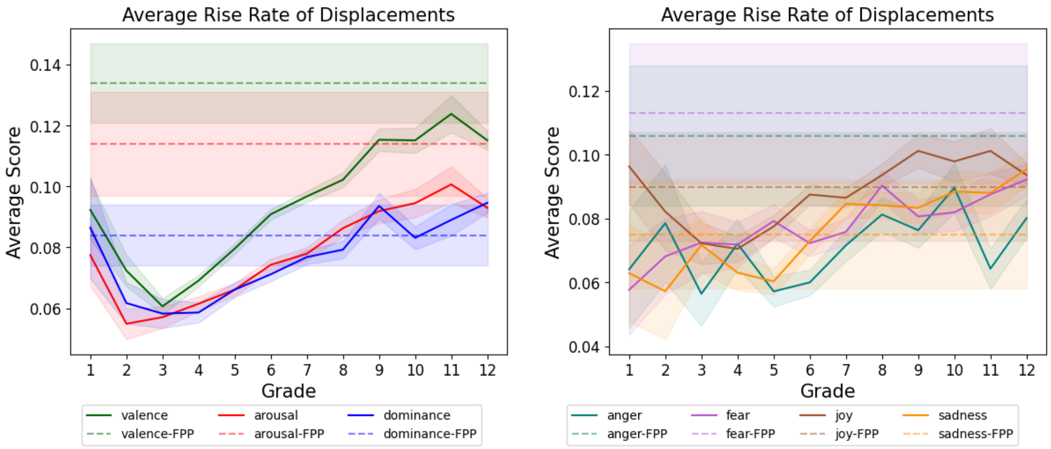
PoKi VAD: In Figure 3.2a, variability for valence, arousal, and dominance all trend upward with age; stabilizing in grades 11 and 12.

FPP VAD: For all three emotions variability was most similar to those of older children, reaching above grades 10–12.

PoKi Anger, Fear, Joy, Sadness: In Figure 3.2b, 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 8–9.

Discussion: The overall trend of increasing emotional variability with age, followed by stabilizing supports findings in psychology. Larson *et al.* [70] 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



(a) Valence, arousal and dominance

(b) Anger, fear, joy and sadness

Figure 3.3: Rise rate in poems across grades. The horizontal dashed lines represent values in poems written by adults. The shaded region around each line represents the standard error of the mean.

variability [11, 5, 114]. Reitsema *et al.* [94] 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 [94].

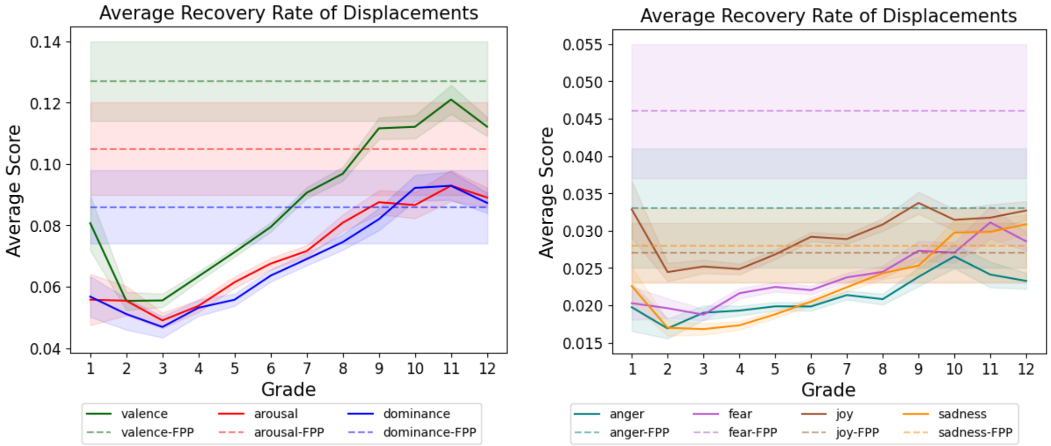
At what rate do emotions change from home to peak state?

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 in emotion score) divided by the number of words during the rise period. The rise rate disregards the direction of the peak.

Results:

PoKi VAD: In Figure 3.3a, 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 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 for children in grades 8 and 10 (grades 9, 11, and 12 have a higher rate than



(a) Valence, arousal and dominance

(b) Anger, fear, joy and sadness

Figure 3.4: Recovery rate in poems across grades. The horizontal dashed lines represent values for poems by adults. The shaded region around each line represents the standard error of the mean.

adults).

PoKi Anger, Fear, Joy, Sadness: In Figure 3.3b, the rise rate for the discrete emotions all increase with grade although there is a fair amount of noise in the trends. Joy has a small dip around grade 4, and then increases until reaching a plateau at grades 9–11. The average rise rates for anger, fear, and sadness all trend upwards overall although have a fairly zigzag pattern. Anger has perhaps the flattest overall trend slope.

FPP Anger, Fear, Joy, Sadness: The rise rate for anger and fear in adult poems is higher than those expressed by children across all grades, with most similar values to older children. Joy and sadness rise rates match slightly younger children in grades 6–7.

Discussion: Rise rate is seen as analogous to reactivity in psychology, which has been found to increase during adolescence [110]. Our findings support these trends. As mentioned in Chapter 2 (Section 2.2.2), emotional reactivity is at the core of anxiety and attention disorder, impacting overall well-being.

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 3.4a 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 3.4b, we similarly see overall increasing average recovery rates for all 4 emotions across grades (although there are some dips around grade 2). 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 10–11. The recovery rate for joy and sadness matches those of slightly younger children, around grade 5 for joy and 9 for sadness.

Discussion: Recovery rate, which is analogous to emotion regulation, has been studied extensively in psychology. Zeman *et al.* [130] 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 academic success of children [38] and adults [89].

At what rate does a previous emotional state predict the next emotional state?

Method: Emotional inertia is computed as the autocorrelation coefficient (ACF). We describe this metric in Chapter 2, however we provide the details here as well for convenience. Autocorrelation can be thought of as the correlation between a time series and the time-lagged version of itself by lag k. The ACF at time lag k can be computed as follows, where $Y_1, ..., Y_N$ are emotion scores at time steps $X_1, ..., X_N$, and N is the total number of time steps in the series:

$$ACF_k = \frac{\sum_{i=1}^{N-k}(Y_i - \overline{Y})(Y_{i+k} - \overline{Y})}{\sum_{i=1}^{N}(Y_i - \overline{Y})^2}$$

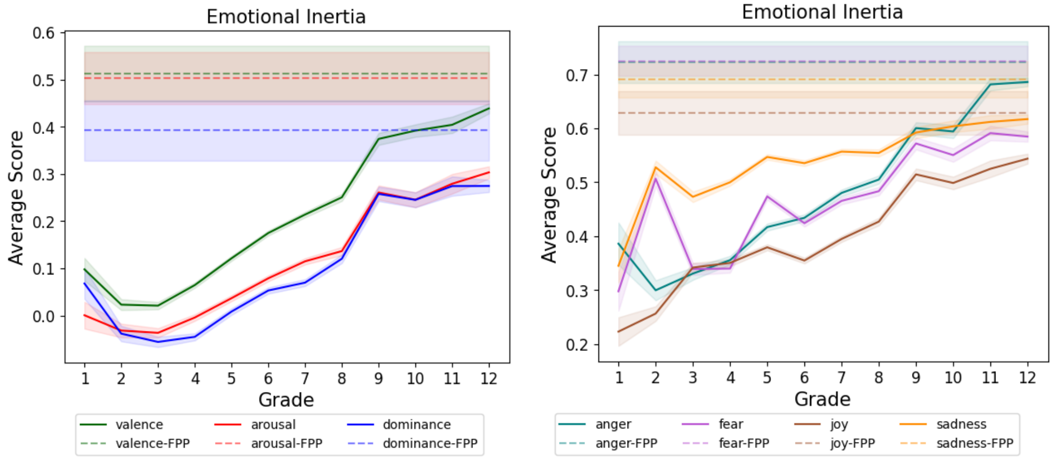
In this scenario we consider a time lag of 1, as it has been used in emotion dynamics work before [112]. A time lag of 1 intuitively means that we are looking at the emotional fluidity between adjacent timesteps. This makes sense in this scenario as the poems are shorter instances of text and have a fewer number of text windows within them.

Results:

PoKi VAD: In Figure 3.5a, we see that emotional inertia for valence, arousal, and dominance increases across grades. There is a local peak at grade 9 for all three emotions. As children reach late-adolescence in grades 11–12, emotional inertia begins to stabilize for all three dimensions.

FPP VAD: The emotional inertia of adults tends to be above those of children across all grades, and is most similar to grades 10–12.

PoKi Anger, Fear, Joy, Sadness: In Figure 3.5b, we see that although there are fluctuations in grades 1–3, overall from grades 4–12 there are increasing levels of emotional inertia across all emotions.



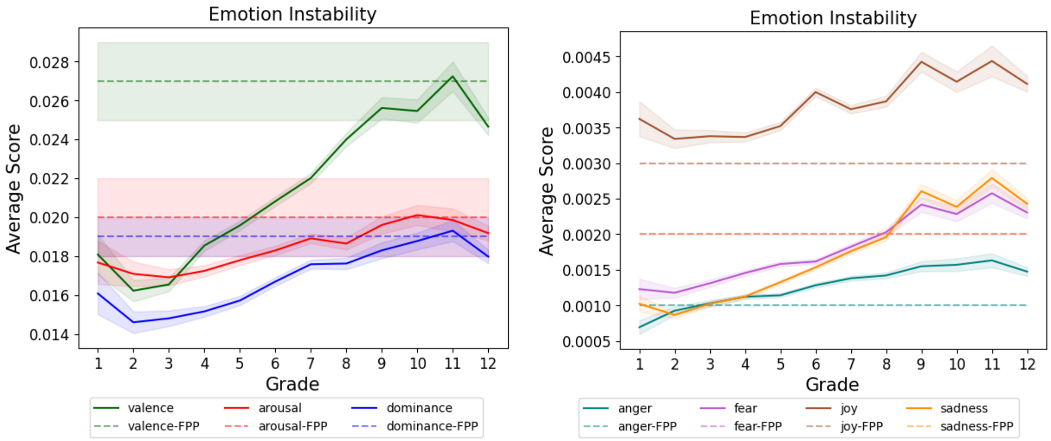
(a) Valence

(b) Anger, fear, joy and sadness

Figure 3.5: Emotional inertia across grades. The horizontal dashed lines represent values for poems by adults. The shaded region around each line represents the standard error of the mean.

FPP Anger, Fear, Joy, Sadness: The emotional inertia of adults tends to be higher than those of children across grades, and match levels for older children in grades 11-12.

Discussion: Work in psychology has found that higher levels of emotional inertia have been associated with lower psychological well-being [52, 61, 60, 63]. Kuppens *et al.* [66] found high emotional inertia in children or adolescents, which points towards potential emotional dysregulation, indicating that they may be vulnerable to the development of psychological problems. We likewise found increasing emotional inertia across grades. While work in psychology focuses on valence, we likewise found this trend also occurs for other emotion dimensions (e.g., arousal, dominance) and categorical emotions (e.g., anger, fear, joy, and sadness). While there is limited work in psychology studying inertia across age ranges, Roekel *et al.* [95] found that no significant difference in positive affect inertia between middle to late adolescents, perhaps reflecting some of the stabilizing results we see nearing older adolescence.



(a) Valence

(b) Anger, fear, joy and sadness

Figure 3.6: Emotion instability across grades. The horizontal dashed lines represent values for poems by adults. The shaded region around each line represents the standard error of the mean.

How variable are emotions with respect to their temporal order across age?

Method: Emotion instability is computed as the mean squared successive difference (MSSD) [86]. More details are in Chapter 2. We provide the formula here for easy access, where x_i represents the emotion score at time-step i and n represents the number of time-steps in the instance.

$$MSSD = \frac{\sum_{i=1}^{n-1}(x_{i+1}-x_i)^2}{n-1}$$

Results:

PoKi VAD: In Figure 3.6a we see that emotion instability increases across grades for valence, arousal, and dominance. Valence has an especially steeper slope for grades 4–11 compared to arousal and dominance. At grades 11-12, instability stabilizes and begins to trend downwards for all emotion dimensions.

FPP VAD: Emotion instability of adults matches and reaches those of older children in grades 10–12.

PoKi Anger, Fear, Joy, Sadness: In Figure 3.6b we see that for all emotions instability increases with age. There are peaks especially at grades 9 and 11 for joy, fear, and sadness. Anger has a relatively flatter slope than the other emotions.

FPP Anger, Fear, Joy, Sadness: For fear and sadness, the emotional instability of adults corresponds with children in grade 8. Whereas for anger, adult instability levels correspond with grade 3 children. For joy, the instability in adults was lower than children across all grades.

Discussion: Higher levels of emotion instability has been shown to correlate with lower psychological well-being, and mental health diagnoses such as depression, bipolar, anxiety, borderline and more [52]. Previous work in psychology has found that higher positive affect instability occurred in middle adolescence compared to late adolescence [95]. Sadness instability was found to increase from childhood to adolescence, and then decrease in later adolescence [94]. These findings in psychology correspond with the patterns we saw for valence as well as other emotions of increasing instabilities followed by stabilizing levels.

3.3.2 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 (LexO approach). Now, we use a *n-gram* windowed 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). ML models are known for their strong performance which lends credibility to the results. In each section we will first describe the ML results, and then compare them to the LexO results.

Datasets: We use the same poem datasets as in Section 3.2, creating n-gram windows of length 5. The only difference is that text windows can contain words not found in the emotion lexicon or neutral words. We chose this approach as ML models are trained on sequential text.

Experiments: We fine-tuned a RoBERTa [74] base model for fine-grain sentiment analysis using the SemEval 2018 Task 1 dataset [85].[8] This means that we were able to predict an emotion score between 0 and 1. After emotion labelling text windows, we performed similar experiments as in Section 4.3: compute the UED metrics per poem and take the average per grade for each metric.

Model Training: We fine-tuned the pretrained RoBERTa [74] base model available on HuggingFace[9]. The SemEval 2018 Task 1 dataset [85] contains tweets annotated with emotion scores for valence, anger, fear, joy and sadness.[10] 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 3.3 and Table 3.4.

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

Table 3.3: Dataset descriptive statistics for the Sem-Eval 2018 Task 1 [85]. The No. of instances includes the train, dev, and test sets for EI-Reg and V-Reg.

[8] 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.

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

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

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

We used the Trainer pipeline from HuggingFace[11] to fine-tune the pre-trained model. For the loss function we used mean-square loss.

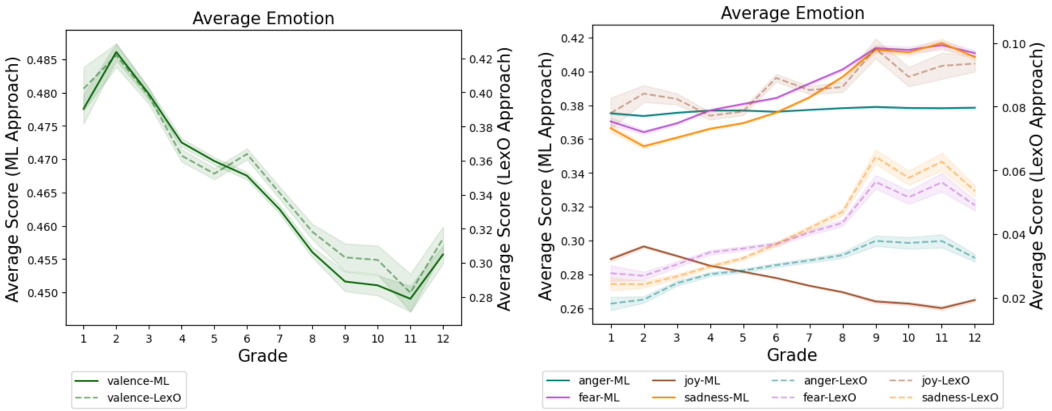
We tuned the following hyperparameters on the development set: learning rate (2e-5, 3e-5), number of epochs (5, 10, 20) and batch size (16, 32). We then selected the best model using mean-square error. Note that our aim here is not to overly fine-tune the model as we are applying it to a different domain (i.e., poems). The best parameters for each emotion model are shown in Table 3.5. After determining the *best* model on the development set we apply it to windows of text in the PoKi poem dataset.

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

Results: Overall we found similar trends as with the *word-level* lexicon approach. We note that a direct comparison between the LexO 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

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



(a) Valence

(b) Anger, fear, joy and sadness

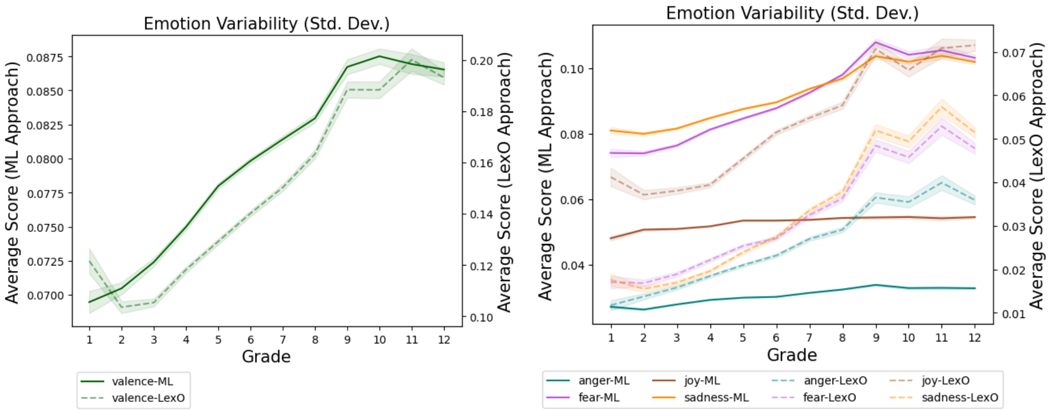
Figure 3.7: Average emotion across grades using the ML *n-gram* approach on the PoKi dataset. The dashed lines represent the values when using the LexO approach (Section 4.4). The shaded region around each line represents the standard error of the mean.

of words). Instead, we can compare the trends in emotion change rather than the magnitude of change or the values themselves. We discuss the results for each UED metric across dimensional (valence, arousal, dominance) and discrete emotions (anger, fear, joy, sadness) below.

Average Emotion: In Figure 3.7a, we see as grade increases, valence similarly trends downwards and there is a stabilization at grades 10–12. In Figure 3.7b anger, fear, and sadness average intensities all increase across grade. We note that the slope across UED metrics for anger are flatter. Perhaps anger is a more challenging emotion for automatic systems to detect [85]. The average intensity for joy decreases, having a similar pattern to that of valence. Perhaps joy and valence appear similar to the ML model resulting in similar trajectories.

Emotional Variability: Older children tend to show increased variability for valence and all categorical emotions (Figure 3.8a & 3.8b). We see a flatter slope for anger and joy, similarly as we did for average emotion (Figure 3.7).

Rise Rate and Recovery Rate:



(a) Valence

(b) Anger, fear, joy and sadness

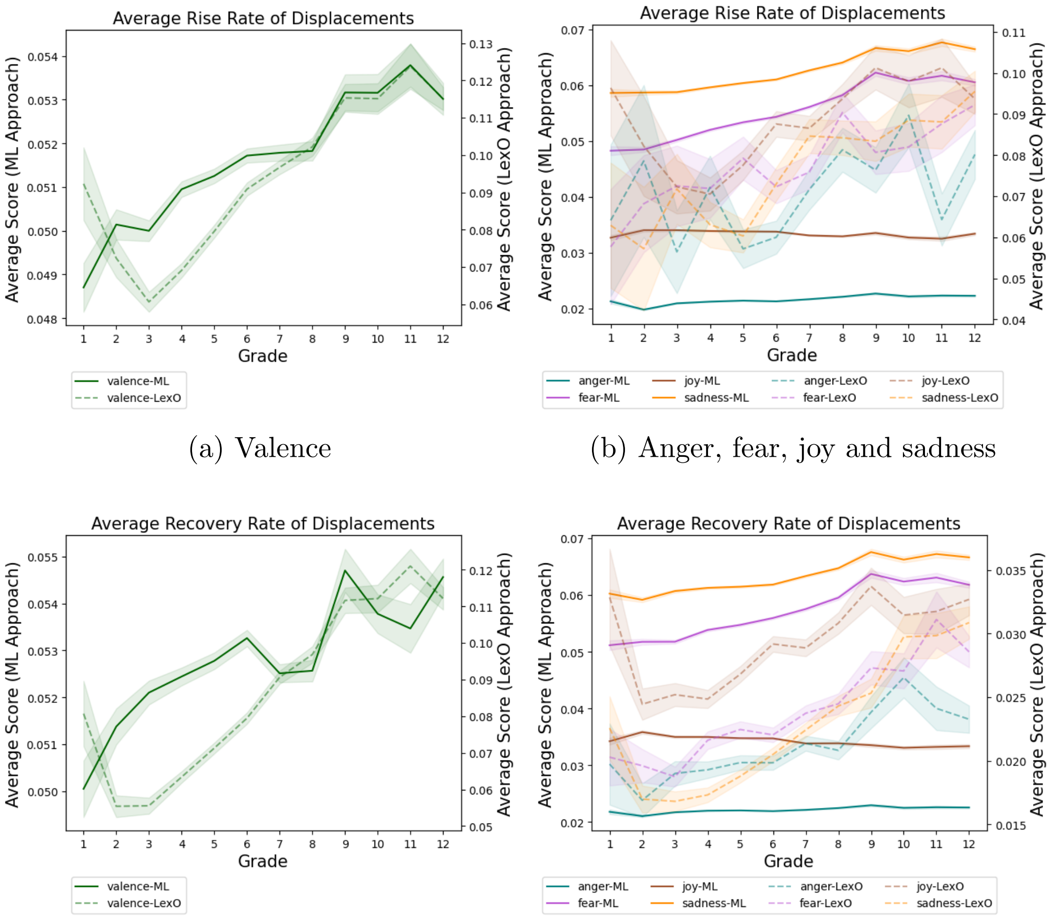
Figure 3.8: Emotional variability across grades using the ML *n-gram* approach on the PoKi dataset. The dashed lines represent the values when using the LexO approach (Section 4.4). The shaded region around each line represents the standard error of the mean.

In Figure 3.9a and 3.9c we show the rise rate and recovery rate for valence using the fine-tuned ML model on the PoKi dataset. With age, children are writing with increased rise and recovery rates. These trends support those seen when using the lexicon approach.

In Figure 3.9b and 3.9d we show rise and recovery rates 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: rise rate and recovery rate increase across grades.

Emotional Inertia: In Figure 3.10a we see increasing valence inertia across grades similarly as with the lexicon approach. In Figure 3.10b, we see increasing trends of inertia for the discrete emotions. Note that the overall patterns seem smoother for the discrete emotions using the ML approach than when using the LexO approach.

Emotional Instability: In Figure 3.11a valence instability increases across grades, similarly as with the LexO approach. In Figure 3.11b, instability for fear and sadness increases across grades (as with the LexO approach).



(a) Valence

(b) Anger, fear, joy and sadness

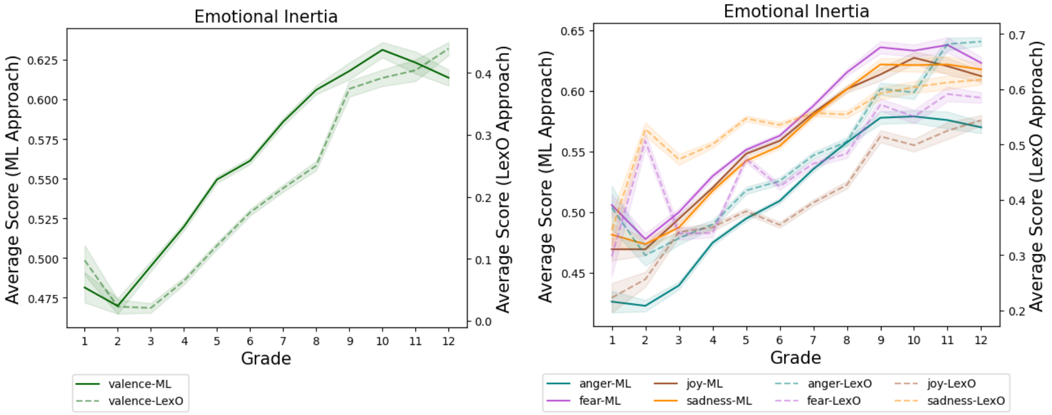
(c) Valence

(d) Anger, fear, joy and sadness

Figure 3.9: Rise rate (a & b) and recovery rates (c & d) for valence (a & c) and discrete emotions (b & d) across grades using the ML *n-gram* approach on the PoKi dataset. The dashed lines represent the values when using the LexO approach (Section 4.4). The shaded region around each line represents the standard error of the mean.

However, joy shows a slight downwards trend and anger has a flat slope. This corresponds with the trends we saw for average emotion, variability, rise rate and recovery rate for anger and joy. This may indicate more difficulty in automatically detecting these emotions.

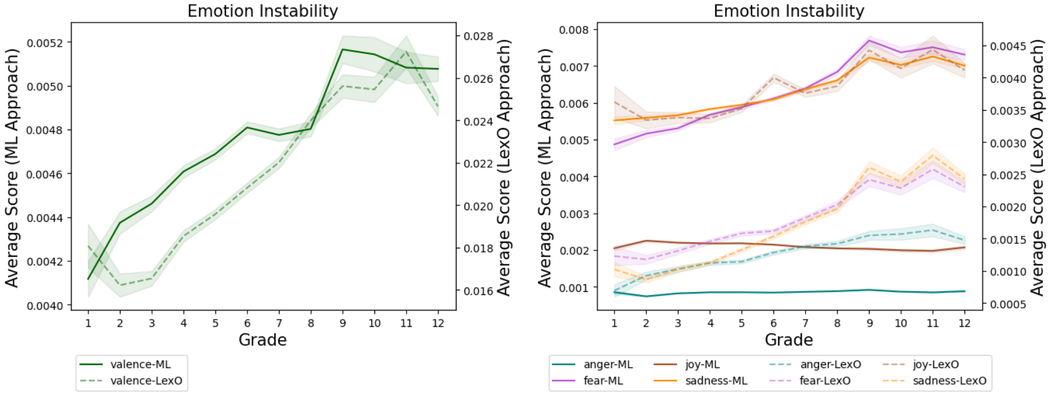
Overall, these results show that there are meaningful patterns of emotion change in children's poems which change across grade, and therefore as children age. The trends found using the lexicon approach are generally replicated using ML models. We see clear similarities in trends for valence, fear, and sad-



(a) Valence

(b) Anger, fear, joy and sadness

Figure 3.10: Emotional inertia across grades using the ML *n-gram* approach on the PoKi dataset. The dashed lines represent the values when using the LexO approach (Section 4.4). The shaded region around each line represents the standard error of the mean.



(a) Valence

(b) Anger, fear, joy and sadness

Figure 3.11: Emotion instability across grades using the ML *n-gram* approach on the PoKi dataset. The dashed lines represent the values when using the LexO approach (Section 4.4). The shaded region around each line represents the standard error of the mean.

ness. When using the LexO method, there appears to be a bit more noise for the discrete emotions (especially for rise rate).

3.4 Summary

We explored six utterance emotion dynamics metrics (average, variability, rise rate, recovery rate, inertia, and instability), and seven emotions (three dimensional and four discrete) in poems written by children and adults. We found that the patterns of emotion change in poetry by children supported previous results and findings in the psychology literature (e.g., increased variability, rise rate, and recovery rates with age). Our results demonstrate that without telling children to write about their feelings, there are patterns of emotion change across age. Namely, there are higher intensities of anger, fear, joy, sadness, arousal, and dominance across age and lower intensities of valence, and increasing emotional variability, rise rates, recovery rates, inertia, and instability for all emotions and dimensions of emotion across age. This means that as children age, their poems express more intense emotions and less happiness, while reaching peaks and dips in emotional states more quickly.

As future work, we would like to examine poetry by adults and compare the patterns of emotion change between experts vs. novices. More generally, we would like to explore how do UED compare across geographic regions, and time periods.

Limitations

The poems written by adults used in Section 4.3 are written by highly accomplished adult writers who are often known for their poetry. These poems may not be representative of poems written by adults in general, which could affect the patterns and trends in emotion words we see. Future work could expand the collection of poems to include those written by novice adults as well.

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.

Chapter 4

Language and Mental Health: Measures of Emotion Dynamics from Text as Linguistic Biosocial Markers

This chapter is currently under review at a conference.

Research in psychopathology has shown that, at an aggregate level, the patterns of emotional change over time—emotion dynamics—are indicators of one's mental health [52, 64, 69, 104, 112]. However, it is not yet known whether measures of *utterance emotion dynamics (UED)* correlate with mental health diagnoses. Here, for the first time, we study the relationship between tweet emotion dynamics and mental health disorders, at an aggregate level. We examine whether UED could be considered a *biosocial* marker for mental health, and what differences or commonalities may exist between control and mental health diagnoses. Twitter provides an abundant amount of text data from the public. By considering tweets from users who have self-disclosed as having a mental health condition (MHC) [116], we can analyze the differences in UED metrics across diagnoses.

We found that each of the UED metrics studied varied by the user's self-disclosed diagnosis. For example: average valence was significantly higher (i.e., more positive) in the control group compared to users with attention-deficit/hyperactivity disorders (ADHD), depression, and post-traumatic stress disorder (PTSD). Valence variability was significantly lower in the control group compared to ADHD, depression, bipolar disorder, major depressive disorder (MDD), PTSD, and obsessive-compulsive disorder (OCD) but not postpartum depression (PPD). Rise and recovery rates, inertia, and instability of valence also exhibited significant differences from the control. This work provides important early evidence for how linguistic cues pertaining to emotion dynamics can play a crucial role as biosocial markers for mental illnesses and aid in the understanding, diagnosis, and management of mental health disorders.

4.1 Research Questions

We describe how utterance emotion dynamics (UED) metrics compare between different diagnoses (e.g., depression, bipolar, etc.) and a control group by exploring the following research questions. Comparing MHCs to the control:

- Does the emotional state averaged across time differ?
- Does the emotional variability differ?
- Does the rate at which emotions reach peak emotional state (i.e., rise rate) differ?
- Does the rate at which emotions recover from peak emotional state back to steady state (i.e., recovery rate) differ?
- Does the fluidity of emotions from one moment to the next (i.e., inertia) differ?
- Does the emotional variability with respect to time (i.e., instability) differ?

We explore each of the above research questions for three dimensions of emotions – valence, arousal, and dominance – further building on the findings in psychology which focus on valence. Our work provides baseline measures for UED across MHCs and insights into new linguistic biosocial markers for mental health. These findings are important for clinicians because they provide a broader context for overall well-being and can help contribute to indicators of early detection and diagnosis management.

4.2 Twitter Dataset

We use a recently compiled dataset—Twitter-STMHD [116]. It comprises of tweets from 27,003 users who have self-reported as having a mental health diagnosis on Twitter. The diagnoses include: depression, major depressive disorder (MDD), post-partum depression (PPD), post-traumatic stress disorder (PTSD), attention-deficit/hyperactivity disorders (ADHD), anxiety, bipolar, and obsessive-compulsive disorder (OCD).

4.2.1 STMHD

Suhavi *et al.* [116] created a regular expression pattern to identify posts which contained a self-disclosure of a diagnosis and the diagnosis name (using a lexicon of common synonyms, abbreviations, etc.), such as 'diagnosed with X'. They collected a large set of tweets using the regex. This resulted in a preliminary dataset of users with potential MHC diagnoses. To handle false positives (e.g., 'my family member has been diagnosed with X', or 'I was not diagnosed with X'), the dataset was split into two non-overlapping parts. One of these parts was annotated by hand, and the other using an updated and high-precision regex. In the part that was annotated by hand, each tweet was annotated by two members of the team. A user was only included in the dataset if both annotations were positive as self-disclosing for a particular class. A licensed clinical psychologist helped verify a sample of 500 tweets from this part of the dataset. Comparing the authors' annotations to those of the psychologist's, the authors annotated the 500 tweet sample with 99.2% precise. The manual annotations were used to refine the regular expressions and diagnosis name lexicon. This updated search pattern was applied to the other dataset split. As a verification of the updated regex, the authors applied it to the manually annotated dataset split. When considering the manual annotations as correct, the regex achieved 94% precision.

The control group consisted of users identified from a random sample of tweets (posted during roughly the same time period as the MHC tweets). These tweeters did not post any tweets that satisfied the MHC regex described above. Additionally, users who had any posts about mental health discourse were removed. We would note here that this does not guarantee that these users did not have an MHC diagnosis, but rather the set as a whole may have very few MHC tweeters. The number of users in the control group was selected to match the size of the depression dataset, which has the largest number of users.

For the finalized set of users, four years of tweets were collected for each user: two years before self-reporting a mental health diagnosis and two years after. For the control group tweets were randomly sampled from between January 2017 and May 2021, the same date range as the other MHC classes.

Users were filtered for their number of tweets as well as their follower count. Users with less than 50 tweets collected were removed so as to allow for more generalizable conclusions to be drawn. Similarly, users with more than 5000 followers were removed so as not to include celebrities, or other organizations that use Twitter to discuss well-being.

Group	#Tweeters Org.	#Tweeters	Avg. #Posts/User
MHC	22,160	10,069	2,177.4
ADHD	8,095	3,866	2,122.2
Bipolar	1,651	721	3,193.3
Depression	6,803	3,017	2,084.0
MDD	325	133	2,402.9
OCD	1,325	605	1,822.9
PPD	547	105	1,671.4
PTSD	3,414	1,622	1,944.9
Control	8,199	4,097	1,613.6

Table 4.1: The number of users in each mental health condition in the original dataset, followed by the number of users and the number of tweets per user in the preprocessed version of the Twitter-STMHD we use for experiments.

4.2.2 Our Preprocessing

We further preprocessed the Twitter-STMHD dataset for our experiments (Section 4.3), as we are specifically interested in the unique patterns of UED for each disorder. Several users self-reported as being diagnosed with more than one disorder, referred to as *comorbidity.* We found a high comorbidity rate between users who self-reported as having anxiety and depression, as is also supported in the literature [90, 37, 48, 19]. Therefore, we removed the anxiety class and only considered the depression class as it was a larger class between the two. We also performed the following preprocessing steps:

- We only consider users who self-reported as having one disorder. We removed 1272 users who had disclosed more than one diagnosis.
- We only consider tweets in English, removing other languages.
- We filtered out tweets that contained URLs.
- We removed retweets (identified through tweets containing ‘RT’, ‘rt’).
- We computed the number of tweets per user, and only considered users whose number of tweets was within the interquartile range (between 25th and 75th percentile) for the diagnosis group they disclosed. This was to

ensure that we are not including users with very little content, or those who use social media extremely frequently.

- We removed punctuation and stop words from the tweets.

Table 4.1 shows key details of the filtered data.

4.3 Experiments

To determine whether UED metrics from tweets can act as biosocial markers for psychopathology, we compare UED metrics for each MHC to the control group to see if they are statistically different. We compute the following UED metrics per user in each condition: average emotion intensity, emotional variability, rise rate, recovery rate, inertia, and instability. We compute UED metrics following the *speaker UED* [118] approach as described in Section 2.3.4.

For each user, we ordered their tweets by timestamp. We used a text window size of five tweets and a rolling window of one. This differs from the previous Chapter (Chapter 3) where instance-level UED metrics were computed and a text window was composed of words in a poem. We used the Emotion Dynamics toolkit [125, 47][1] to compute UED metrics (average emotion, emotional variability, rise rate and recovery rate). We implemented the inertia and instability metrics. When computing the autocorrelation coefficient for inertia, we used a time lag of 1. We chose this value based on previous work [112], and for a more fine-grain analysis. Future work can explore varying values of time lags to better determine its relationship with inertia in UED. We performed analyses for valence, arousal, and dominance. For word-emotion association scores we use the NRC Valence, Arousal, and Dominance (VAD)

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

Emotion	**UED Metric**	**df1**	**df2**	**F-statistic**	**P-value**
Valence	average emotion	7	14158	60.50	p<.001
	emotional variability	7	14158	66.33	p<.001
	rise rate	7	14156	77.35	p<.001
	recovery rate	7	14156	72.58	p<.001
	inertia	7	14158	12.52	p<.001
	instability	7	14158	80.08	p<.001
Arousal	average emotion	7	14158	37.60	p<.001
	emotional variability	7	14158	22.38	p<.001
	rise rate	7	14155	34.76	p<.001
	recovery rate	7	14155	41.28	p<.001
	inertia	7	14158	14.94	p<.001
	instability	7	14158	25.47	p<.001
Dominance	average emotion	7	14158	61.86	p<.001
	emotional variability	7	14158	72.21	p<.001
	rise rate	7	14150	37.69	p<.001
	recovery rate	7	14154	35.64	p<.001
	inertia	7	14158	19.25	p<.001
	instability	7	14158	103.45	p<.001

Table 4.2: The degrees of freedom, F-statistic, and p-value in Levene's test of Homogeneity of Variances for each UED metric and emotion. Levene's test indicated that the assumption for homogeneity of variance was violated for the effect of diagnosis on all UED metrics across all three emotions (valence, arousal, and dominance).

lexicon [82]. Afterwards, we performed an ANOVA to test for significant differences between groups in UED metrics, and post-hoc analyses to determine which groups specifically had significant differences from the control group.

4.4 Results

To analyze potential differences across groups and the control group, we perform an ANOVA for each of the UED metrics (average emotion, emotional variability, rise rate, recovery rate, inertia, and instability) and emotion dimensions (valence, arousal, and dominance). We examined a total of N=14166 users, see Table 4.1 for descriptives.

In order to conduct an ANOVA, several assumptions must be met. The three primary assumptions are: the data for each independent variable are approximately normally distributed, the data are independent of each other,

Emotion	**UED Metric**	**df1**	**df2**	**F-stat.**	**P-val**	**Effect Size ($est\ \omega^2$)**
Valence	average emotion	7	1021.65	14.79	p<.001	0.0068
	emo. variability	7	1021.20	70.30	p<.001	0.0331
	rise rate	7	1026.32	9.93	p<.001	0.0044
	recovery rate	7	1023.62	8.86	p<.001	0.0039
	inertia	7	1025.67	28.10	p<.001	0.0132
	instability	7	1021.03	32.43	p<.001	0.0153
Arousal	average emotion	7	1024.41	33.24	p<.001	0.0157
	emo. variability	7	1029.77	66.23	p<.001	0.0312
	rise rate	7	1025.85	2.84	p=.006	0.0009
	recovery rate	7	1026.95	5.19	p<.001	0.0021
	inertia	7	1029.70	13.68	p<.001	0.0062
	instability	7	1029.25	42.28	p<.001	0.0200
Dominance	average emotion	7	1020.10	56.69	p<.001	0.0268
	emo. variability	7	1023.12	40.50	p<.001	0.0191
	rise rate	7	1025.35	6.31	p<.001	0.0026
	recovery rate	7	1022.99	9.94	p<.001	0.0044
	inertia	7	1031.04	12.40	p<.001	0.0056
	instability	7	1024.38	20.99	p<.001	0.0098

Table 4.3: The degrees of freedom (for the numerator and denominator), F-statistic, p-value, and effect size in Welch's ANOVA test for differences between groups for each UED metric and emotion (valence, arousal and dominance). "emo." is an abbreviation for emotional.

and the distributions have roughly the same variance (homoscedasticity). We can assume the mean is *normally distributed* according to the central limit theorem, due to the large sample size (law of large numbers). Since there are different tweeters in each MHC group we can assume the data are independent of each other. However, we note that generally people are largely not independent of each other e.g., having friends across groups, interacting with content from various mental health groups, etc. The *homogeneity of variance* or homoscedasticity assumption can be tested by looking at the residuals and performing Levene's test. In our case, Levene's test indicated that the assumption for homogeneity of variance was violated for all metrics and emotions (results in Table 4.2). As such, we used Welch's ANOVA test which is an alternative to ANOVA when the equal variance assumption is not met.

The first part of our analyses are Omnibus F-tests, which test for significant differences between groups. This test can not tell us which groups are different from each other, just rather that there is a difference among groups. For each combination of UED metrics and emotions (e.g., emotional variability for valence) there was a significant main effect which means we can conclude that at least one of the mental health diagnoses significantly differed. We show the degrees of freedom, F-statistic, p-value, and corrected effect size in Table 4.3 for valence, arousal, and dominance. The effect size tells us how meaningful the difference between groups is for each metric.[2] For example, 3.31% of the total variance in the emotional variability for valence is accounted for by diagnosis (small effect).

In the next step of our analyses, we would like to know exactly which groups differed from the control group. In order to do this we performed post hoc analyses for pairwise comparisons between groups for each metric across the three dimensions of emotion. We applied a Games-Howell correction since the assumption for homogeneity of variance was violated. In the following Sections we detail how the UED metrics compare across MHCs compared to the control group. We report the results across emotion dimensions (valence, arousal, and dominance) in Table 4.4 for average emotion, emotional variability, rise rate and recovery rate, and in Table 4.5 for inertia and instability. We contextualize our results with previous findings in psychology and studies in NLP. We note that the relationship between patterns of emotion change and well-being for arousal and dominance are under-explored – our findings provide important benchmarks for these dimensional emotions and UED metrics more generally.

[2] An effect size <0.01 is *very small*, between 0.01 to 0.06 is *small*, between 0.06 to 0.14 is *medium*, and greater or equal than 0.14 is *large* [30].

UED **MHC–Control**	**Average Emotion**			**Emotion Variability**			**Rise Rate**			**Recovery Rate**		
Emotion Dimension	V	A	D	V	A	D	V	A	D	V	A	D
ADHD-control	↓	↓	↓	↑	↑	↑	–	–	↑	–	↑	↑
Bipolar-control	–	↓	↓	↑	↑	↑	–	–	–	↑	–	–
Depression-control	–	↓	↓	↑	↑	↑	↑	–	↑	↑	↑	↑
MDD-control	↓	–	↓	↑	↑	↑	↑	–	–	↑	↑	↑
OCD-control	–	↓	↓	↑	↑	↑	–	–	↑	–	↑	↑
PPD-control	–	↓	↓	–	↑	↑	–	–	–	–	–	–
PTSD-control	↓	–	↓	↑	↑	↑	↑	↑	–	↑	↑	↑

Table 4.4: **Valence (V), Arousal (A), Dominance (D)**: The difference in UED metrics across MHC groups compared to the control. If there was a significant difference, the arrow indicates the direction of the difference, otherwise the cell has a dash. E.g., ↓ for ADHD-control average emotion under the 'V' columns means that the ADHD had significantly lower average valence than the control group.

4.4.1 How does the average emotion for an MHC compare to the control?

Valence: The average valence was significantly lower for the ADHD, MDD, and PTSD groups compared to the control group.

Arousal: The ADHD, depression, bipolar, PPD, and OCD groups showed significantly lower arousal compared to the control group.

Dominance: All MHC groups (ADHD, depression, bipolar, MDD, PPD, PTSD, OCD) showed significantly lower dominance compared to the control group.

Discussion: Our findings align with results in psychology, and NLP studies looking at the average dimensional emotions expressed in text. Valence was found to be lower in individuals with depression (of which MDD is a type of depression) through self-reports questionnaires [45, 104] and on social media [98, 22, 25]. Further, work in psychology has found individuals with PTSD and ADHD have lower valence [91, 115]. It has also been shown that the lower arousal and dominance in speech is associated with depression [113, 87, 41]. While average emotion intensity is one of the more commonly explored

UED **MHC–Control**	**Instability**			**Inertia**		
Emotion Dimension	V	A	D	V	A	D
ADHD-control	↑	↑	↑	↑	↑	↑
Bipolar-control	↑	↑	↑	↑	↑	↑
Depression-control	↑	↑	↑	↑	↑	↑
MDD-control	↑	↑	↑	↑	↑	↑
OCD-control	↑	↑	↑	↑	↑	↑
PPD-control	–	–	–	↑	↑	↑
PTSD-control	↑	↑	↑	↑	↑	↑

Table 4.5: **Valence (V), Arousal (A), Dominance (D)**: The difference in UED metrics across MHC groups compared to the control. If there was a significant difference, the arrow indicates the direction of the difference, otherwise the cell has a dash. E.g., ↑ for ADHD-control instability under the 'V' columns means that the ADHD had significantly higher instability for valence than the control group.

measures, there is still relatively few works studying the relationships between arousal and dominance with mental health, compared to valence. Interestingly, dominance appears to differ for many MHCs (all studied here) from the control group, pointing towards an important indicator of well-being.

4.4.2 How does emotional variability for an MHC compare to the control?

Valence: Variability for valence was significantly higher for the ADHD, depression, bipolar, MDD, PTSD, and OCD groups compared to the control. PPD did not show differences from the control.

Arousal: All MHC groups (ADHD, depression, bipolar, MDD, PPD, PTSD, and OCD) showed significantly higher arousal variability.

Dominance: All MHC groups (ADHD, depression, bipolar, MDD, PPD, PTSD, OCD) had significantly higher dominance variability than the control group.

Discussion: In several studies in psychology, it has been shown that higher valence variability occurred for individuals with depression, PTSD [52, 45] and is negatively correlated with overall well-being [52]. Interestingly, Seabrook *et* *al.* [98] found higher valence variability on Twitter indicated lower depression severity which contradicted their findings on Facebook. Kuppens *et al.* [67] report that valence variability was negatively related to self-esteem and was positively related to neuroticism and depression. Overall, our results align with emotional variability having strong ties with well-being. Arousal and dominance variability appear to be *biosocial* markers across several MHCs, although minimally explored in the literature (Ranney *et al.* [92] found higher affective arousal variability was associated with generalized anxiety disorder).

4.4.3 How does emotional rise rate for an MHC compare to the control?

Valence: Rise rate for valence was significantly higher for the depression, MDD, and PTSD groups compared to the control group.

Arousal: PTSD was the only group which had statistically higher arousal rise rates than the control group.

Dominance: The ADHD, depression, and OCD groups had significantly higher rise rates than the control group.

***Discussion**:* Rise rate is analogous to emotional reactivity in psychology, and quickly moving to peak emotional states has been shown in individuals with maladaptive emotion patterns and lower psychological well-being [52]. It is interesting to note that valence and dominance rise rates differed across MHC to the control, whereas not to the same extent for arousal.

4.4.4 How does emotional recovery rate for an MHC compare to the control?

Valence: Recovery rate for valence was significantly higher for the depression, bipolar, MDD, and PTSD groups compared to the control group.

Arousal: The ADHD, depression, MDD, PTSD, and OCD groups showed significantly higher arousal recovery rates compared to the control group.

Dominance: The ADHD, depression, MDD, PTSD, and OCD groups showed significantly higher dominance recovery rates than the control group.

***Discussion**:* Recovery rate can be thought of as a proxy of emotion regulation, and slower recovery from emotional events is associated with psychopathology and poor psychological well-being [52, 64]. Our results, while pointing to higher recovery rates, indicate significant differences from the control group. This is an interesting result that can be further explored if found in other mediums such as Reddit.

4.4.5 How does emotional instability for an MHC compare to the control?

Valence, Arousal, and Dominance: For all three emotions the ADHD, bipolar, depression, MDD, OCD, and PTSD groups had significantly higher instability compared to the control group. The PPD group did not show significant differences from the control for valence, arousal, or dominance.

Discussion: Having higher instability means experiencing more variable emotional shifts from one moment-to-moment [52]. Higher instability has been shown to be associated with mental health diagnoses such as depression, bipolar, and more [52]. It appears that for all of the diagnoses studied here that instability may be a *biosocial* marker, however not for PPD. This points to interesting insights on how PPD may present itself in text. Overall, across UED metrics it appears that PPD is more similar to the control group than the other diagnoses. We note that PPD is not listed as a separate diagnosis from depression in the Diagnostic and Statistical Manual of Mental Disorders (DSM-5), which is a guide of symptoms and descriptions for diagnosing mental disorders [7].

4.4.6 How does emotional inertia for an MHC compare to the control?

Valence, Arousal, and Dominance: All diagnoses (ADHD, bipolar, depression, MDD, OCD, PPD, and PTSD) had significantly higher inertia for valence, arousal, and dominance compared to the control group.

Discussion: Having higher inertia means that it is harder to move from one emotional state to the next [63, 69], and that a previous emotional state can better predict next emotional states [65, 117]. Across emotion dimensions, it appears that inertia is an indicator for all diagnoses. While these findings may seem to conflict with diagnoses having higher instability as well, this is a known finding in psychology referred to as the “inertia–instability paradox” [62, 10]. Previous work in psychology has also found diagnoses such as depression have both higher inertia and higher instability [62, 10]. This paradox can be explained by the statistical overlap between the metrics rather than studying them separately. When correcting the methodological approach, Koval *et al.* [62] demonstrated that depressive symptoms are not associated with instability or inertia of negative affect in daily life but rather with variability. Likewise, Bos *et al.* [10] showed that instability and inertia of negative affect were not associated with depressive symptoms when variability was adjusted for. Since this paradox appears in our findings as well, future work could examine the relationship between diagnoses and UED when accounting for variability.

4.5 Results: *Biosocial* Aspects

We found UED metrics do vary by disclosed mental health disorder. However, language is inherently social (as discussed in the Introduction, Chapter 1), and is an aspect that cannot be ignored when looking at features of language as

Popularity Aspect	**Average**	**Std. Dev.**	**25th Percentile**	**Median**	**75th Percentile**
No. Followers	626.89	821.42	138.00	317.00	748.00
No. Following	771.19	881.22	235.5	472.0	928.00
Avg. Likes	2.06	8.33	0.67	1.20	2.11

Table 4.6: Descriptives for the distributions of *popularity* aspects on Twitter: number of people following, number of followers, average likes.

potential indicators of well-being. To ensure that the *popularity* of individuals on Twitter is not driving the difference in group UED metrics, we performed analyses accounting for three *popularity* measures. The three measures we consider are the number of people following, the number of followers, and the average number of likes one has.

Metadata for these *popularity* measures is available for each user. In Table 4.6 we show the data descriptives for the distributions of the three *popularity* aspects on Twitter: number of people following, number of followers, average likes. Once extracting the *popularity* measures, we binned users based on their values. We look at two bins, users who fall in the 25th to 50th percentile, and those who are in the 50th to 75th percentile. Then, we performed the same analyses as in Section 4.4. Levene's test indicated that the assumption for homogeneity of variance was violated for all metrics and dimensions of emotions, so we used Welch's ANOVA test for significant differences. In Table 4.7 and Table 4.8 we show the results for each *popularity* aspect across the three emotion dimensions. We also show an example of these differences by graphing the distributions of the average intensity of valence for the ADHD and control group for both the 25%–50% percentile of average likes (Figure 4.1a), and for the 50%–75% percentile of average likes (Figure 4.1b). In this Figure we can see that regardless of average number of likes a user receives, there are significant differences in average valence between the ADHD and control group.

In these exploratory experiments we see that even when we controlled for

Social Aspect	Bin	Average Emotion			Emotional Variability			Rise Rate			Recovery Rate		
Emotion Dimension		V	A	D	V	A	D	V	A	D	V	A	D
No. Following	1	*	*	*	*	*	*				*		
No. Following	2	*	*	*	*	*	*	*					*
No. Followers	1	*	*	*	*	*	*				*		
No. Followers	2	*	*	*	*	*	*	*					*
Avg No. Likes	1	*	*	*	*	*	*	*		*	*		*
Avg No. Likes	2	*	*	*	*	*	*	*		*	*		*

Table 4.7: **Average Emotion, Emotional Variability, Rise Rate and Recovery Rate**: Results for Welch's test for significant differences among the groups when controlling for various social aspects. Statistically significant results are depicted by an asterisk.

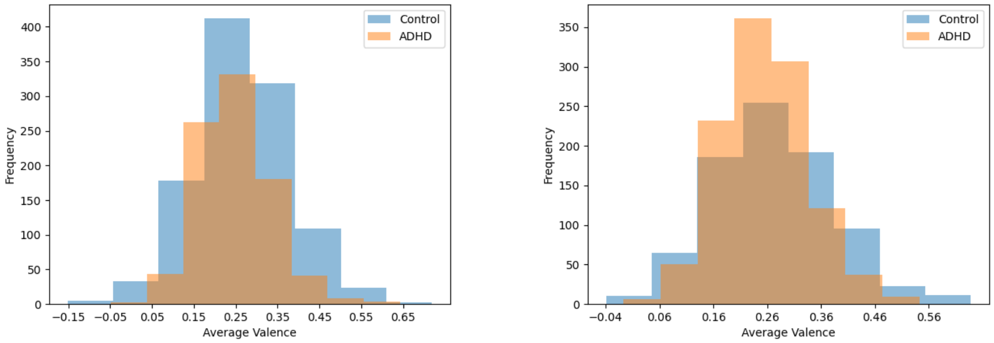
Social Aspect	Bin	Instability			Inertia		
Emotion Dimension		V	A	D	V	A	D
No. Following	1	*	*	*	*	*	*
No. Following	2	*	*	*	*	*	*
No. Followers	1	*	*	*	*	*	*
No. Followers	2	*	*	*	*	*	*
Avg No. Likes	1	*	*	*	*	*	*
Avg No. Likes	2	*	*	*	*	*	*

Table 4.8: **Instability and Inertia**: Results for Welch's test for significant differences among the groups when controlling for various social aspects. Statistically significant results are depicted by an asterisk.

popularity measures (e.g., number of followers, following, average number of likes one receives) we find that groups differed in UED metrics, however popularity measures may play a role for some metrics (especially rise and recovery rate). On the other hand, average emotion, emotional variability, inertia, and instability (for all three emotion dimensions) appear to be robust indicators of mental health even when users have varying levels of these social aspects.

4.6 Summary

In this chapter we showed for the first time that there are significant relationships between patterns of emotion change in text written by individuals with a self-disclosed MHC compared to a control group. By using a Twitter dataset where users have chosen to disclose an MHC diagnosis, we found



(a) Analysis for users with a low average number of likes received on their posts (in the 25%–50% percentile).

(b) Analysis for users with a high average number of likes received on their post (in the 50%–75% percentile).

Figure 4.1: The distributions of average valence for the control and ADHD group when considering users with an average number of likes on their posts in the 25%–50% and 50%–75% percentile. In both of these scenarios, there were significant differences in average valence for the ADHD and control group.

significant differences in six UED metrics (average emotion, emotional variability, rise rate, recovery rate, instability, and inertia) across three emotion dimensions (valence, arousal, and dominance) for MHCs. Our findings provide important contextual information of overall well-being and supporting indicators (in addition to other assessments) to clinicians for diagnosis detection and management.

Building on this work, others could extend our analyses and consider other mediums such as other social media platforms (e.g., Reddit), and explore if UED metrics are biosocial markers across different regions, languages, and socioeconomic statuses. Notably, future work should collaborate with clinicians to gain a more in-depth understanding of various aspects such as recovery and intervention, and the role emotion dynamics may play. Through such a partnership, UED metrics could be studied in the context of clinical information as well.

Limitations

In this study, we used NLP techniques to compare the UED across different psychopathologies and a control group. It is important to be cautious of interpreting these results due to natural limitations within the dataset. Due to the high rates of comorbidity with mental health disorders [29] examining users who only disclosed one diagnosis may not be representative of the population. Furthermore, it is also possible that our dataset included users with more than one disorder but only disclosed one (e.g., a user may have been diagnosed with depression and ADHD but only tweeted "diagnosed with ADHD" or vice versa). Self-disclosure of diagnoses may also be inaccurate due to reasons such as impression management [72] or social desirability [71] where users may disclose having a diagnosis without having received a formal diagnosis. Alternatively, there may have been users included in the control group who have a formal diagnosis of one or more mental disorders but did not disclose their mental disorders on Twitter. In sum, despite the authors' best efforts to collect data accordingly, the users in each group may not be representative of the mental disorders. Future research could replicate this study using a sample of users with confirmed formal diagnoses.

Chapter 5

Conclusion

Studying patterns of emotion change over time provides key insights for industry and research, as it allows us to better understand feelings over time. The characteristics with which our emotions change from one moment to the next provides indicators of psychological well-being and mental health in psychology [52, 64, 69, 104, 112]. In this thesis we explored patterns of emotion change in text – utterance emotion dynamics (UED) – in the context of emotional development and psychological well-being. We also expanded upon the UED framework by introducing two new metrics: inertia and instability. We explored two use-cases: poems written by children across grades 1 to 12, and Twitter posts for those who have self-disclosed as having a mental health diagnosis.

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

In Chapter 3, we demonstrated that UED metrics show trends in children's writing across grades, which support previously reported results using emotion dynamics in psychology. We demonstrated that the average intensity of categorical emotions (e.g., anger, fear, joy, and sadness) and some dimensional emotions (e.g., arousal and dominance) increased across grades, whereas the average intensity of valence decreased. Further, the variability, rise rates, recovery rates, inertia, and instability increased for all categorical and dimensional emotions across age. Previous work in psychology found similar trends: average affect decreased with age [31, 70, 105, 128], arousal [13, 42, 108] and sadness increased with age [50], and variability [70, 94], reactivity [110], regulation [130], inertia [66], and instability [95, 94] increased with age.

While using a ML approach for emotion scoring pointed to largely the same trends as the lexicon-only (LexO) approach, trend lines did not correspond as well with the LexO approach for anger and joy. As discussed in Chapter 3, this could be due to the difficulty in automatically detecting more nuanced emotions, such as anger. Further, the trends for valence and joy were similar which could indicate the ML method was not able to detect the subtleties of joy, or that the LexO approach may be a bit more prone to noise.

Future work could explore poems written by adults more generally (compared to poems by expert writers which we used), and whether similar trends in UED metrics occur across various geographic regions, and languages. Our work provides baseline measures for six UED metrics across an array of emotions in children's writing, which helps set the foundation work for emotional development in children's writing, and its relation to well-being.

5.2 Language and Mental Health: Measures of Emotion Dynamics from Text as Linguistic Biosocial Markers

In Chapter 4, we examined if UED metrics could act as *biosocial marker* for mental health using tweets. We compared the UED metrics for seven mental health conditions (MHCs) (e.g., ADHD, Bipolar, Depression, MDD, OCD, PPD, and PTSD) to a control group on Twitter. Taken as a whole, our results across emotions and UED metrics indicate that social media posts for those who self-disclose as having a mental health diagnosis are less happy, less in-control and less activated (e.g., lower valence, arousal, and dominance), more variable in their emotion intensities, reach and recovery from emotional peaks quicker, are more instable, and emotional states tend to persisted overtime. Overall, these trends in UED metrics remained even after considering social/popularity aspects on Twitter (e.g., number of followers, number of following, and average number of likes received).

Future work could explore if similar trends hold across domains (e.g., Twitter vs. Reddit), languages (i.e., other than English), and geographical regions. Our work demonstrates that UED metrics do have the potential to provide contextual information to clinicians about mental health, however social factors should be considered to better understand the role UED metrics could have as *biosocial* markers.

5.3 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 [84, 80]. 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.

5.4 Concluding Thoughts

Overall, in this thesis we for the first time explored *utterance emotion dynamics* for a novel application – to see if UED can provide meaningful indicators of emotional development in children, and mental health. By exploring poems written by children and tweets of individuals who self-reported as having a MHC, we showed that meaningful characteristics of emotion change can be found in text. Our findings could inform and provide contextual information for emotional development psychologists and clinicians. Future work down this line should collaborate with clinicians to gain a more in-depth understanding, and study UED metrics in the context of clinical information as well.

References

[1] [Online]. Available: https://datareportal.com/social-media-users.

[2] [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/mental-disorders#:~:text=In%202019%2C%201%20in%20every,the%20most%20common%20(1)..

[3] C. O. Alm and R. Sproat, "Emotional sequencing and development in fairy tales," in *Proceedings of the First International Conference on Affective Computing and Intelligent Interaction*, ser. ACII'05, Beijing, China: Springer-Verlag, 2005, pp. 668–674, ISBN: 3540296212. DOI: 10.1007/11573548_86. [Online]. Available: https://doi.org/10.1007/11573548_86.

[4] M. E. Aragón, A. P. López-Monroy, L. C. González-Gurrola, and M. Montes-y-Gómez, "Detecting depression in social media using fine-grained emotions," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 1481–1486. DOI: 10.18653/v1/N19-1151. [Online]. Available: https://aclanthology.org/N19-1151.

[5] J. Arnett, "Adolescent storm and stress, reconsidered," *The American psychologist*, vol. 54, no. 5, pp. 317–326, May 1999, ISSN: 0003-066X. DOI: 10.1037//0003-066x.54.5.317. [Online]. Available: https://doi.org/10.1037//0003-066x.54.5.317.

[6] A. Ashida, M. Tokumaru, and T. Kojiri, "Characters' emotion design support system for writing novels based on story arcs of target readers," *Information and Technology in Education and Learning*, vol. 1, Reg–p005, Jan. 2021. DOI: 10.12937/itel.1.1.Reg.p005.

[7] A. P. Association, *Diagnostic and Statistical Manual of Mental Disorders*, Fifth Edition. American Psychiatric Association, 2013. DOI: 10.1176/appi.books.9780890425596. eprint: https://dsm.psychiatryonline.org/doi/pdf/10.1176/appi.books.9780890425596. [Online]. Available: https://dsm.psychiatryonline.org/doi/abs/10.1176/appi.books.9780890425596.

[8] A. M. Belfi, E. A. Vessel, and G. G. Starr, "Individual ratings of vividness predict aesthetic appeal in poetry.," *Psychology of Aesthetics, Creativity, and the Arts*, vol. 12, no. 3, p. 341, 2018.

[9] S. Bhyravajjula, U. Narayan, and M. Shrivastava, "Marcus: An event-centric nlp pipeline that generates character arcs from narratives," in *Text2Story@ECIR*, 2022.

[10] E. H. Bos, P. de Jonge, and R. F. A. Cox, "Affective variability in depression: Revisiting the inertia-instability paradox," *British journal of psychology (London, England : 1953)*, vol. 110, no. 4, pp. 814–827, Nov. 2019, ISSN: 0007-1269. DOI: 10.1111/bjop.12372. [Online]. Available: https://europepmc.org/articles/PMC6899922.

[11] C. Buchanan, J. Eccles, and J. Becker, "Are adolescents the victims of raging hormones: Evidence for activational effects of hormones on moods and behavior at adolescence," *Psychological bulletin*, vol. 111, no. 1, pp. 62–107, Jan. 1992, ISSN: 0033-2909. DOI: 10.1037/0033-2909.111.1.62. [Online]. Available: https://doi.org/10.1037/0033-2909.111.1.62.

[12] L. Calzà, G. Gagliardi, R. Rossini Favretti, and F. Tamburini, "Linguistic features and automatic classifiers for identifying mild cognitive impairment and dementia," *Computer Speech & Language*, vol. 65, p. 101 113, 2021, ISSN: 0885-2308. DOI: https://doi.org/10.1016/j.csl.2020.101113. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0885230820300462.

[13] L. Carstensen, M. Pasupathi, U. Mayr, and J. Nesselroade, "Emotional experience in everyday life across the adult life span," *Journal of personality and social psychology*, vol. 79, no. 4, pp. 644–655, Oct. 2000, ISSN: 0022-3514. DOI: 10.1037/0022-3514.79.4.644. [Online]. Available: https://doi.org/10.1037/0022-3514.79.4.644.

[14] T. Carthy, N. Horesh, A. Apter, and J. J. Gross, "Patterns of emotional reactivity and regulation in children with anxiety disorders," English, *Journal of Psychopathology and Behavioral Assessment*, vol. 32, no. 1, pp. 23–36, 2010, 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. DOI: 10.1007/s10862-009-9167-8.

[15] S. Chancellor and M. De Choudhury, "Methods in predictive techniques for mental health status on social media: A critical review," *NPJ digital medicine*, vol. 3, p. 43, 2020, ISSN: 2398-6352. DOI: 10.1038/s41746-020-0233-7. [Online]. Available: https://europepmc.org/articles/PMC7093465.

[16] M. Chong and S. Gottipati, "Social media influences and instagram storytelling: Case study of singapore instagram influences," English, *The Journal of Applied Business and Economics*, vol. 22, no. 10, pp. 81–96, 2020, Copyright - Copyright North American Business Press 2020; Last updated - 2021-06-03. [Online]. Available: `https://login.ezproxy.library.ualberta.ca/login?url=https://www.proquest.com/scholarly-journals/social-media-influences-lnstagram-storytelling/docview/2496342719/se-2?accountid=14474`.

[17] T. W. W. M. H. S. Consortium, "Prevalence, Severity, and Unmet Need for Treatment of Mental Disorders in the World Health Organization World Mental Health Surveys," *JAMA*, vol. 291, no. 21, pp. 2581–2590, Jun. 2004, ISSN: 0098-7484. DOI: `10.1001/jama.291.21.2581`. [Online]. Available: `https://doi.org/10.1001/jama.291.21.2581`.

[18] C. M. Corcoran, V. A. Mittal, C. E. Bearden, *et al.*, "Language as a biomarker for psychosis: A natural language processing approach," *Schizophrenia Research*, vol. 226, pp. 158–166, 2020, Biomarkers in the Attenuated Psychosis Syndrome, ISSN: 0920-9964. DOI: `https://doi.org/10.1016/j.schres.2020.04.032`. [Online]. Available: `https://www.sciencedirect.com/science/article/pii/S0920996420302474`.

[19] C. M. Cummings, N. E. Caporino, and P. C. Kendall, "Comorbidity of anxiety and depression in children and adolescents: 20 years after.," *Psychological Bulletin*, vol. 140, no. 3, pp. 816–845, May 2014. DOI: `10.1037/a0034733`. [Online]. Available: `https://doi.org/10.1037%5C%2Fa0034733`.

[20] V. Cuteri, G. Minori, G. Gagliardi, *et al.*, "Linguistic feature of anorexia nervosa: A prospective case-control pilot study," *Eating and weight disorders : EWD*, vol. 27, no. 4, pp. 1367–1375, May 2022, ISSN: 1124-4909. DOI: `10.1007/s40519-021-01273-7`. [Online]. Available: `https://europepmc.org/articles/PMC8311399`.

[21] R. J. Davidson, "Affective style and affective disorders: Perspectives from affective neuroscience," *Cognition and Emotion*, vol. 12, no. 3, pp. 307–330, 1998. DOI: `10.1080/026999398379628`. eprint: `https://doi.org/10.1080/026999398379628`. [Online]. Available: `https://doi.org/10.1080/026999398379628`.

[22] M. De Choudhury, S. Counts, and E. Horvitz, "Social media as a measurement tool of depression in populations," in *Proceedings of the 5th Annual ACM Web Science Conference*, ser. WebSci '13, Paris, France: Association for Computing Machinery, 2013, pp. 47–56, ISBN: 9781450318891. DOI: `10.1145/2464464.2464480`. [Online]. Available: `https://doi.org/10.1145/2464464.2464480`.

[23] M. De Choudhury, S. Counts, E. J. Horvitz, and A. Hoff, "Characterizing and predicting postpartum depression from shared facebook data," in *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*, ser. CSCW '14, Baltimore, Maryland, USA: Association for Computing Machinery, 2014, pp. 626–638, ISBN: 9781450325400. DOI: `10.1145/2531602.2531675`. [Online]. Available: `https://doi.org/10.1145/2531602.2531675`.

[24] M. De Choudhury, M. Gamon, S. Counts, and E. Horvitz, "Predicting depression via social media," in *Seventh international AAAI conference on weblogs and social media*, 2013, pp. 128–137.

[25] M. De Choudhury, M. Gamon, S. Counts, and E. Horvitz, "Predicting depression via social media," *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 7, no. 1, pp. 128–137, Aug. 2021. DOI: `10.1609/icwsm.v7i1.14432`. [Online]. Available: `https://ojs.aaai.org/index.php/ICWSM/article/view/14432`.

[26] M. Del Vecchio, A. Kharlamov, G. Parry, and G. Pogrebna, "The Data Science of Hollywood: Using Emotional Arcs of Movies to Drive Business Model Innovation in Entertainment Industries," *arXiv e-prints*, arXiv:1807.02221, arXiv:1807.02221, Jul. 2018. arXiv: `1807.02221 [cs.CL]`.

[27] C. Doquet, "Ancrages théoriques de l'analyse génétique des textes d'élèves," in *L'Ecole, l'écriture et la création. Etudes françaises et brésiliennes.* Ser. Sciences du langage - Carrefour et points de vue, C. B. et Eduardo Calil, Ed., Academia Bruylant, 2013, pp. 33–53. [Online]. Available: `https://hal.science/hal-01236152`.

[28] M. Dreyfuss, K. Caudle, A. T. Drysdale, *et al.*, "Teens impulsively react rather than retreat from threat," *Developmental neuroscience*, vol. 36, no. 3-4, pp. 220–227, 2014, ISSN: 0378-5866. DOI: `10.1159/000357755`. [Online]. Available: `https://europepmc.org/articles/PMC4125471`.

[29] B. G. Druss and E. R. Walker, "Mental disorders and medical comorbidity," *The Synthesis project. Research synthesis report*, no. 21, pp. 1–26, Feb. 2011, ISSN: 2155-370X. [Online]. Available: `http://europepmc.org/abstract/MED/21675009`.

[30] A. Field, *Discovering Statistics Using IBM SPSS Statistics* (Introducing statistical methods). SAGE Publications, 2013, ISBN: 9781446249178. [Online]. Available: `https://books.google.de/books?id=srb0a9fmMEoC`.

[31] A. Frost, L. T. Hoyt, A. L. Chung, and E. K. Adam, "Daily life with depressive symptoms: Gender differences in adolescents' everyday emotional experiences," *Journal of Adolescence*, vol. 43, pp. 132–141, 2015, ISSN: 0140-1971. DOI: `https://doi.org/10.1016/j.adolescence.2015.06.001`. [Online]. Available: `https://www.sciencedirect.com/science/article/pii/S0140197115001232`.

[32] G. Gagliardi and F. Tamburini, "Linguistic biomarkers for the detection of mild cognitive impairment," it, *Lingue e linguaggio, Rivista semestrale*, no. 1/2021, pp. 3–31, 2021, ISSN: 1720-9331. DOI: 10.1418/101111. [Online]. Available: https://www.rivisteweb.it/doi/10.1418/101111.

[33] G. Gagliardi and F. Tamburini, "The automatic extraction of linguistic biomarkers as a viable solution for the early diagnosis of mental disorders," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, Marseille, France: European Language Resources Association, Jun. 2022, pp. 5234–5242. [Online]. Available: https://aclanthology.org/2022.lrec-1.561.

[34] M. Ghazvininejad, X. Shi, J. Priyadarshi, and K. Knight, "Hafez: An interactive poetry generation system," in *Proceedings of ACL 2017, System Demonstrations*, Vancouver, Canada: Association for Computational Linguistics, Jul. 2017, pp. 43–48. [Online]. Available: https://aclanthology.org/P17-4008.

[35] G. Gkotsis, A. Oellrich, T. Hubbard, *et al.*, "The language of mental health problems in social media," in *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*, San Diego, CA, USA: Association for Computational Linguistics, Jun. 2016, pp. 63–73. DOI: 10.18653/v1/W16-0307. [Online]. Available: https://aclanthology.org/W16-0307.

[36] H. Gonçalo Oliveira, "A survey on intelligent poetry generation: Languages, features, techniques, reutilisation and evaluation," in *Proceedings of the 10th International Conference on Natural Language Generation*, Santiago de Compostela, Spain: Association for Computational Linguistics, Sep. 2017, pp. 11–20. DOI: 10.18653/v1/W17-3502. [Online]. Available: https://aclanthology.org/W17-3502.

[37] J. M. Gorman, "Comorbid depression and anxiety spectrum disorders," *Depression and Anxiety*, vol. 4, no. 4, pp. 160–168, 1996. DOI: https://doi.org/10.1002/(SICI)1520-6394(1996)4:4<160::AID-DA2>3.0.CO;2-J. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/\%28SICI\%291520-6394\%281996\%294\%3A4\%3C160\%3A\%3AAID-DA2\%3E3.0.CO\%3B2-J. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/%5C%28SICI%5C%291520-6394%5C%281996%5C%294%5C%3A4%5C%3C160%5C%3A%5C%3AAID-DA2%5C%3E3.0.CO%5C%3B2-J.

[38] P. A. Graziano, R. D. Reavis, S. P. Keane, and S. D. Calkins, "The role of emotion regulation in children's early academic success," *Journal of School Psychology*, vol. 45, no. 1, pp. 3–19, 2007, ISSN: 0022-4405. DOI: https://doi.org/10.1016/j.jsp.2006.09.002. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0022440506000859.

[39] J. J. Gross, "The extended process model of emotion regulation: Elaborations, applications, and future directions," *Psychological Inquiry*, vol. 26, no. 1, pp. 130–137, 2015. DOI: 10.1080/1047840X.2015.989751. eprint: https://doi.org/10.1080/1047840X.2015.989751. [Online]. Available: https://doi.org/10.1080/1047840X.2015.989751.

[40] J. J. Gross and R. F. Muñoz, "Emotion regulation and mental health," *Clinical Psychology: Science and Practice*, vol. 2, no. 2, pp. 151–164, 1995. DOI: https://doi.org/10.1111/j.1468-2850.1995.tb00036.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1468-2850.1995.tb00036.x. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1468-2850.1995.tb00036.x.

[41] M. Gumus, D. D. DeSouza, M. Xu, C. Fidalgo, W. Simpson, and J. Robin, "Evaluating the utility of daily speech assessments for monitoring depression symptoms," *DIGITAL HEALTH*, vol. 9, p. 20 552 076 231 180 523, 2023. DOI: 10.1177/20552076231180523. eprint: https://doi.org/10.1177/20552076231180523. [Online]. Available: https://doi.org/10.1177/20552076231180523.

[42] M. R. Gunnar, S. Wewerka, K. Frenn, J. D. Long, and C. Griggs, "Developmental changes in hypothalamus-pituitary-adrenal activity over the transition to adolescence: Normative changes and associations with puberty," *Development and psychopathology*, vol. 21, no. 1, pp. 69–85, 2009, ISSN: 0954-5794. DOI: 10.1017/s0954579409000054. [Online]. Available: https://europepmc.org/articles/PMC3933029.

[43] S. C. Guntuku, R. Schneider, A. Pelullo, *et al.*, "Studying expressions of loneliness in individuals using Twitter: An observational study," *BMJ open*, vol. 9, no. 11, e030355, 2019.

[44] K. Harrigian, C. Aguirre, and M. Dredze, "On the state of social media data for mental health research," in *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*, Online: Association for Computational Linguistics, Jun. 2021, pp. 15–24. DOI: 10.18653/v1/2021.clpsych-1.2. [Online]. Available: https://aclanthology.org/2021.clpsych-1.2.

[45] A. S. Heller, A. S. Fox, and R. J. Davidson, "Parsing affective dynamics to identify risk for mood and anxiety disorders," en, *Emotion*, vol. 19, no. 2, pp. 283–291, Jun. 2018.

[46] W. Hipson and S. M. Mohammad, "PoKi: A large dataset of poems by children," English, in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, Marseille, France: European Language Resources Association, May 2020, pp. 1578–1589, ISBN: 979-10-95546-34-4. [Online]. Available: https://aclanthology.org/2020.lrec-1.196.

[47] W. E. Hipson and S. M. Mohammad, "Emotion dynamics in movie dialogues," *PLOS ONE*, vol. 16, no. 9, pp. 1–19, Sep. 2021. DOI: 10.1371/journal.pone.0256153. [Online]. Available: https://doi.org/10.1371/journal.pone.0256153.

[48] R. M. A. Hirschfeld, "The comorbidity of major depression and anxiety disorders: Recognition and management in primary care," *Primary care companion to the Journal of clinical psychiatry*, vol. 3, no. 6, pp. 244–254, Dec. 2001, ISSN: 1523-5998. DOI: 10.4088/pcc.v03n0609. [Online]. Available: https://europepmc.org/articles/PMC181193.

[49] S. G. Hofmann, A. T. Sawyer, A. Fang, and A. Asnaani, "Emotion dysregulation model of mood and anxiety disorders," *Depression and Anxiety*, vol. 29, no. 5, pp. 409–416, 2012. DOI: https://doi.org/10.1002/da.21888. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/da.21888. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/da.21888.

[50] I. Holsen, P. Kraft, and J. Vitterso, "Stability in depressed mood in adolescence: Results from a 6-year longitudinal panel study," English, *Journal of Youth and Adolescence*, vol. 29, no. 1, pp. 61–78, Feb. 2000, Copyright - Copyright Plenum Publishing Corporation Feb 2000; Last updated - 2023-02-07; CODEN - JYADA6. [Online]. Available: https://login.ezproxy.library.ualberta.ca/login?url=https://www.proquest.com/scholarly-journals/stability-depressed-mood-adolescence-results-6/docview/204521468/se-2.

[51] M. Houben, W. V. D. Noortgate, and P. Kuppens, "The relation between short-term emotion dynamics and psychological well-being: A meta-analysis.," *Psychological Bulletin*, vol. 141, no. 4, pp. 901–930, Jul. 2015. DOI: 10.1037/a0038822. [Online]. Available: https://doi.org/10.1037%2Fa0038822.

[52] M. Houben, W. Van Den Noortgate, and P. Kuppens, *The relation between short-term emotion dynamics and psychological well-being: A meta-analysis*, US, 2015.

[53] T. Hu, S. Wang, W. Luo, *et al.*, "Revealing public opinion towards covid-19 vaccines with twitter data in the united states: Spatiotemporal perspective," *Journal of Medical Internet Research*, vol. 23, no. 9, e30854, 2021.

[54] S. Jahng, P. K. Wood, and T. J. Trull, "Analysis of affective instability in ecological momentary assessment: Indices using successive difference and group comparison via multilevel modeling," *Psychological methods*, vol. 13, no. 4, pp. 354–375, Dec. 2008, ISSN: 1082-989X. DOI: 10.1037/a0014173. [Online]. Available: https://doi.org/10.1037/a0014173.

[55] P. N. Johnson-Laird and K. Oatley, "How poetry evokes emotions," *Acta Psychologica*, vol. 224, p. 103 506, 2022, ISSN: 0001-6918. DOI: https://doi.org/10.1016/j.actpsy.2022.103506. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S000169182200021X.

[56] J. Karrass, T. A. Walden, E. G. Conture, *et al.*, "Relation of emotional reactivity and regulation to childhood stuttering," *Journal of communication disorders*, vol. 39, no. 6, pp. 402–423, 2006, ISSN: 0021-9924. DOI: 10.1016/j.jcomdis.2005.12.004. [Online]. Available: https://europepmc.org/articles/PMC1630450.

[57] E. Kim and R. Klinger, "A survey on sentiment and emotion analysis for computational literary studies," *Zeitschrift für digitale Geisteswissenschaften*, 2019. DOI: 10.17175/2019_008. [Online]. Available: https://zfdg.de/2019_008_v1.

[58] E. Kim, S. Padó, and R. Klinger, "Investigating the relationship between literary genres and emotional plot development," in *Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, Vancouver, Canada: Association for Computational Linguistics, Aug. 2017, pp. 17–26. DOI: 10.18653/v1/W17-2203. [Online]. Available: https://aclanthology.org/W17-2203.

[59] S. Koops, S. G. Brederoo, J. N. de Boer, F. G. Nadema, A. E. Voppel, and I. E. Sommer, "Speech as a biomarker for depression," *CNS& neurological disorders drug targets*, vol. 22, no. 2, pp. 152–160, 2023, ISSN: 1871-5273. DOI: 10.2174/1871527320666211213125847. [Online]. Available: https://doi.org/10.2174/1871527320666211213125847.

[60] P. Koval, E. A. Butler, T. Hollenstein, D. Lanteigne, and P. Kuppens, "Emotion regulation and the temporal dynamics of emotions: Effects of cognitive reappraisal and expressive suppression on emotional inertia," *Cognition & emotion*, vol. 29, no. 5, pp. 831–851, 2015, ISSN: 0269-9931. DOI: 10.1080/02699931.2014.948388. [Online]. Available: https://doi.org/10.1080/02699931.2014.948388.

[61] P. Koval, P. Kuppens, N. B. Allen, and L. Sheeber, "Getting stuck in depression: The roles of rumination and emotional inertia," *Cognition & emotion*, vol. 26, no. 8, pp. 1412–1427, 2012, ISSN: 0269-9931. DOI: 10.1080/02699931.2012.667392. [Online]. Available: https://doi.org/10.1080/02699931.2012.667392.

[62] P. Koval, M. L. Pe, K. Meers, and P. Kuppens, "Affect dynamics in relation to depressive symptoms: Variable, unstable or inert?" *Emotion (Washington, D.C.)*, vol. 13, no. 6, pp. 1132–1141, Dec. 2013, ISSN: 1528-3542. DOI: 10.1037/a0033579. [Online]. Available: https://doi.org/10.1037/a0033579.

[63] P. Koval, S. Sütterlin, and P. Kuppens, "Emotional inertia is associated with lower well-being when controlling for differences in emotional context," *Frontiers in Psychology*, vol. 6, 2016, ISSN: 1664-1078. DOI: 10.3389/fpsyg.2015.01997. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fpsyg.2015.01997.

[64] P. A. Kragel, A. R. Hariri, and K. S. LaBar, "The temporal dynamics of spontaneous emotional brain states and their implications for mental health," *Journal of cognitive neuroscience*, vol. 34, no. 5, pp. 715–728, 2022, May, 2022, ISSN: 0898-929X. DOI: 10.1162/jocn_a_01787. [Online]. Available: https://doi.org/10.1162/jocn_a_01787.

[65] P. Kuppens, N. B. Allen, and L. B. Sheeber, "Emotional inertia and psychological maladjustment," *Psychological Science*, vol. 21, no. 7, pp. 984–991, 2010, PMID: 20501521. DOI: 10.1177/0956797610372634. eprint: https://doi.org/10.1177/0956797610372634. [Online]. Available: https://doi.org/10.1177/0956797610372634.

[66] P. Kuppens, L. B. Sheeber, M. B. H. Yap, S. Whittle, J. G. Simmons, and N. B. Allen, "Emotional inertia prospectively predicts the onset of depressive disorder in adolescence," *Emotion (Washington, D.C.)*, vol. 12, no. 2, pp. 283–289, Apr. 2012, ISSN: 1528-3542. DOI: 10.1037/a0025046. [Online]. Available: https://doi.org/10.1037/a0025046.

[67] P. Kuppens, I. Van Mechelen, J. B. Nezlek, D. Dossche, and T. Timmermans, "Individual differences in core affect variability and their relationship to personality and psychological adjustment," *Emotion (Washington, D.C.)*, vol. 7, no. 2, pp. 262–274, May 2007, ISSN: 1528-3542. DOI: 10.1037/1528-3542.7.2.262. [Online]. Available: https://doi.org/10.1037/1528-3542.7.2.262.

[68] P. Kuppens and P. Verduyn, "Looking at emotion regulation through the window of emotion dynamics," *Psychological Inquiry*, vol. 26, no. 1, pp. 72–79, 2015. DOI: 10.1080/1047840X.2015.960505. eprint: https://doi.org/10.1080/1047840X.2015.960505. [Online]. Available: https://doi.org/10.1080/1047840X.2015.960505.

[69] P. Kuppens and P. Verduyn, "Emotion dynamics," *Current Opinion in Psychology*, vol. 17, pp. 22–26, 2017, Emotion, ISSN: 2352-250X. DOI: https://doi.org/10.1016/j.copsyc.2017.06.004. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352250X16302019.

[70] R. W. Larson, G. Moneta, M. H. Richards, and S. Wilson, "Continuity, stability, and change in daily emotional experience across adolescence," *Child Development*, vol. 73, no. 4, pp. 1151–1165, 2002. DOI: https://doi.org/10.1111/1467-8624.00464. eprint: https://srcd.onlinelibrary.wiley.com/doi/pdf/10.1111/1467-8624.00464. [Online]. Available: https://srcd.onlinelibrary.wiley.com/doi/abs/10.1111/1467-8624.00464.

[71] C. A. Latkin, C. Edwards, M. A. Davey-Rothwell, and K. E. Tobin, "The relationship between social desirability bias and self-reports of health, substance use, and social network factors among urban substance users in baltimore, maryland," *Addictive Behaviors*, vol. 73, pp. 133–136, 2017, ISSN: 0306-4603. DOI: `https://doi.org/10.1016/j.addbeh.2017.05.005`. [Online]. Available: `https://www.sciencedirect.com/science/article/pii/S0306460317301752`.

[72] M. Leary, "Impression management, psychology of," in *International Encyclopedia of the Social & Behavioral Sciences*, N. J. Smelser and P. B. Baltes, Eds., Oxford: Pergamon, 2001, pp. 7245–7248, ISBN: 978-0-08-043076-8. DOI: `https://doi.org/10.1016/B0-08-043076-7/01727-7`. [Online]. Available: `https://www.sciencedirect.com/science/article/pii/B0080430767017277`.

[73] P. Lena, "More than a biomarker: Could language be a biosocial marker of psychosis?" English, *NPJ Schizophrenia*, vol. 7, no. 1, 2021. [Online]. Available: `https://www.proquest.com/scholarly-journals/more-than-biomarker-could-language-be-biosocial/docview/2567801968/se-2`, Copyright - © The Author(s) 2021. This work is published under http://creativecommons.org/licenses/by/4.0/ (the "License"). Notwithstanding the ProQuest Terms and Conditions, you may use this content in accordance with the terms of the License; Last updated - 2023-02-22.

[74] Y. Liu, M. Ott, N. Goyal, *et al.*, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[75] B. MacWhinney, *The CHILDES project: Tools for analyzing talk, Volume II: The database*. Psychology Press, 2014.

[76] M. Manabe, K. Liew, S. Yada, S. Wakamiya, and E. Aramaki, "Estimation of psychological distress in japanese youth through narrative writing: Text-based stylometric and sentiment analyses," *JMIR Form Res*, vol. 5, no. 8, e29500, Aug. 2021, ISSN: 2561-326X. DOI: `10.2196/29500`. [Online]. Available: `http://www.ncbi.nlm.nih.gov/pubmed/34387556`.

[77] K. McRae, J. J. Gross, J. Weber, *et al.*, "The development of emotion regulation: an fMRI study of cognitive reappraisal in children, adolescents and young adults," *Social Cognitive and Affective Neuroscience*, vol. 7, no. 1, pp. 11–22, Jan. 2012, ISSN: 1749-5016. DOI: `10.1093/scan/nsr093`. eprint: `https://academic.oup.com/scan/article-pdf/7/1/11/27106078/nsr093.pdf`. [Online]. Available: `https://doi.org/10.1093/scan/nsr093`.

[78] D. S. Mennin, R. M. Holaway, D. M. Fresco, M. T. Moore, and R. G. Heimberg, "Delineating components of emotion and its dysregulation in anxiety and mood psychopathology," *Behavior Therapy*, vol. 38,

no. 3, pp. 284–302, 2007, ISSN: 0005-7894. DOI: https://doi.org/10.1016/j.beth.2006.09.001. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0005789407000202.

[79] S. Mohammad, "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*, Portland, OR, USA: Association for Computational Linguistics, Jun. 2011, pp. 105–114. [Online]. Available: https://aclanthology.org/W11-1514.

[80] S. Mohammad, "Best practices in the creation and use of emotion lexicons," in *Findings of the Association for Computational Linguistics: EACL 2023*, Dubrovnik, Croatia: Association for Computational Linguistics, May 2023, pp. 1825–1836. [Online]. Available: https://aclanthology.org/2023.findings-eacl.136.

[81] S. M. Mohammad, "From once upon a time to happily ever after: Tracking emotions in mail and books," *Decision Support Systems*, vol. 53, no. 4, pp. 730–741, 2012, 1) Computational Approaches to Subjectivity and Sentiment Analysis 2) Service Science in Information Systems Research : Special Issue on PACIS 2010, ISSN: 0167-9236. DOI: https://doi.org/10.1016/j.dss.2012.05.030. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167923612001418.

[82] S. M. Mohammad, "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, 2018.

[83] S. M. Mohammad, "Word affect intensities," in *Proceedings of the 11th Edition of the Language Resources and Evaluation Conference (LREC-2018)*, Miyazaki, Japan, 2018.

[84] S. M. Mohammad, "Ethics sheet for automatic emotion recognition and sentiment analysis," *To Appear in Computational Linguistics*, Jun. 2022.

[85] S. M. Mohammad, F. Bravo-Marquez, M. Salameh, and S. Kiritchenko, "Semeval-2018 Task 1: Affect in tweets," in *Proceedings of International Workshop on Semantic Evaluation (SemEval-2018)*, New Orleans, LA, USA, 2018.

[86] J. von Neumann, R. H. Kent, H. R. Bellinson, and B. I. Hart, "The Mean Square Successive Difference," *The Annals of Mathematical Statistics*, vol. 12, no. 2, pp. 153–162, 1941. DOI: 10.1214/aoms/1177731746. [Online]. Available: https://doi.org/10.1214/aoms/1177731746.

[87] K. Osatuke, J. K. Mosher, J. Z. Goldsmith, *et al.*, "Submissive voices dominate in depression: Assimilation analysis of a helpful session," *Journal of Clinical Psychology*, vol. 63, no. 2, pp. 153–164, 2007. DOI: https://doi.org/10.1002/jclp.20338. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jclp.20338. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/jclp.20338.

[88] G. Park and M. Kwak, "The life cycle of online smartphone reviews: Investigating dynamic change in customer opinion using sentiment analysis," *ICIC Express Letters*, May 2020. DOI: 10.24507/icicelb.11.05.509.

[89] L. H. Phillips, L. H. Phillips, R. Bull, E. Adams, and L. Fraser, "Positive mood and executive function: Evidence from stroop and fluency tasks," *Emotion (Washington, D.C.)*, vol. 2, no. 1, pp. 12–22, Mar. 2002, ISSN: 1528-3542. DOI: 10.1037/1528-3542.2.1.12. [Online]. Available: https://doi.org/10.1037/1528-3542.2.1.12.

[90] M. H. Pollack, "Comorbid anxiety and depression," *Journal of Clinical Psychiatry*, vol. 66, p. 22, 2005.

[91] C. P. Pugach, C. L. May, and B. E. Wisco, "Positive emotion in posttraumatic stress disorder: A global or context-specific problem?" *Journal of Traumatic Stress*, vol. 36, no. 2, pp. 444–456, 2023, ISSN: 0894-9867. DOI: 10.1002/jts.22928. [Online]. Available: https://doi.org/10.1002/jts.22928.

[92] R. M. Ranney, E. Behar, and A. S. Yamasaki, "Affect variability and emotional reactivity in generalized anxiety disorder," *Journal of Behavior Therapy and Experimental Psychiatry*, vol. 68, p. 101 542, 2020, ISSN: 0005-7916. DOI: https://doi.org/10.1016/j.jbtep.2019.101542. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S000579161830257X.

[93] A. J. Reagan, L. Mitchell, D. Kiley, C. M. Danforth, and P. S. Dodds, "The emotional arcs of stories are dominated by six basic shapes," English, *EPJ Data Science*, vol. 5, no. 1, pp. 1–12, Nov. 2016, Copyright - EPJ Data Science is a copyright of Springer, 2016; Last updated - 2017-02-06. [Online]. Available: https://www.proquest.com/scholarly-journals/emotional-arcs-stories-are-dominated-six-basic/docview/1865288690/se-2.

[94] A. M. Reitsema, B. F. Jeronimus, M. van Dijk, and P. de Jonge, "Emotion dynamics in children and adolescents: A meta-analytic and descriptive review," *Emotion (Washington, D.C.)*, vol. 22, no. 2, pp. 374–396, Mar. 2022, ISSN: 1528-3542. DOI: 10.1037/emo0000970. [Online]. Available: https://doi.org/10.1037/emo0000970.

[95] E. van Roekel, E. C. Bennik, J. A. Bastiaansen, *et al.*, "Depressive symptoms and the experience of pleasure in daily life: An exploration of associations in early and late adolescence," *Journal of abnormal child psychology*, vol. 44, no. 5, pp. 999–1009, Jul. 2016, ISSN: 0091-0627. DOI: `10.1007/s10802-015-0090-z`. [Online]. Available: `https://europepmc.org/articles/PMC4893355`.

[96] M. Rothbart and D. Derryberry, "Development of individual differences in temperament," in Jan. 1981, vol. Vol. 1, pp. 33–86.

[97] L. N. Scott, S. E. Victor, E. A. Kaufman, *et al.*, "Affective dynamics across internalizing and externalizing dimensions of psychopathology," *Clinical psychological science : a journal of the Association for Psychological Science*, vol. 8, no. 3, pp. 412–427, May 2020, ISSN: 2167-7026. DOI: `10.1177/2167702619898802`. [Online]. Available: `https://europepmc.org/articles/PMC7363045`.

[98] E. M. Seabrook, M. L. Kern, B. D. Fulcher, and N. S. Rickard, "Predicting depression from language-based emotion dynamics: Longitudinal analysis of facebook and twitter status updates," *J Med Internet Res*, vol. 20, no. 5, e168, May 2018, ISSN: 1438-8871. DOI: `10.2196/jmir.9267`. [Online]. Available: `http://www.ncbi.nlm.nih.gov/pubmed/29739736`.

[99] E. M. Seabrook, M. L. Kern, B. D. Fulcher, and N. S. Rickard, "Predicting depression from language-based emotion dynamics: Longitudinal analysis of facebook and twitter status updates," *J Med Internet Res*, vol. 20, no. 5, e168, May 2018, ISSN: 1438-8871. DOI: `10.2196/jmir.9267`. [Online]. Available: `http://www.ncbi.nlm.nih.gov/pubmed/29739736`.

[100] T. Seidl, "The politics of platform capitalism. a case study on the regulation of uber in new york," *SocArXiv*, 2020. DOI: `https://doi.org/10.1111/rego.12353`.

[101] B. G. Shapero, L. Y. Abramson, and L. B. Alloy, "Emotional reactivity and internalizing symptoms: Moderating role of emotion regulation," *Cognitive therapy and research*, vol. 40, no. 30, pp. 328–340, Jun. 2016, ISSN: 0147-5916. DOI: `10.1007/s10608-015-9722-4`. [Online]. Available: `https://europepmc.org/articles/PMC4876867`.

[102] G. Sheppes, G. Suri, and J. J. Gross, "Emotion regulation and psychopathology," *Annual Review of Clinical Psychology*, vol. 11, no. 1, pp. 379–405, 2015, PMID: 25581242. DOI: `10.1146/annurev-clinpsy-032814-112739`. eprint: `https://doi.org/10.1146/annurev-clinpsy-032814-112739`. [Online]. Available: `https://doi.org/10.1146/annurev-clinpsy-032814-112739`.

[103] G. Sheppes, G. Suri, and J. J. Gross, "Emotion regulation and psychopathology," *Annual Review of Clinical Psychology*, vol. 11, no. 1, pp. 379–405, 2015, PMID: 25581242. DOI: 10.1146/annurev-clinpsy-032814-112739. eprint: https://doi.org/10.1146/annurev-clinpsy-032814-112739. [Online]. Available: https://doi.org/10.1146/annurev-clinpsy-032814-112739.

[104] J. S. Silk, E. E. Forbes, D. J. Whalen, *et al.*, "Daily emotional dynamics in depressed youth: A cell phone ecological momentary assessment study," *Journal of Experimental Child Psychology*, vol. 110, no. 2, pp. 241–257, 2011, Special Issue: Assessment of Emotion in Children and Adolescents, ISSN: 0022-0965. DOI: https://doi.org/10.1016/j.jecp.2010.10.007. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0022096510002055.

[105] R. Simmons, R. Burgeson, S. Carlton-Ford, and D. Blyth, "The impact of cumulative change in early adolescence," *Child development*, vol. 58, no. 5, pp. 1220–1234, Oct. 1987, ISSN: 0009-3920. DOI: 10.1111/j.1467-8624.1987.tb01453.x. [Online]. Available: https://doi.org/10.2307/1130616.

[106] P. Singh, Y. Dwivedi, K. Kahlon, D. R. S. Sawhney, A. Alalwan, and N. Rana, "Smart monitoring and controlling of government policies using social media and cloud computing," *Information Systems Frontiers*, vol. 22, Apr. 2020. DOI: 10.1007/s10796-019-09916-y.

[107] S. Somasundaran, X. Chen, and M. Flor, "Emotion arcs of student narratives," in *Proceedings of the First Joint Workshop on Narrative Understanding, Storylines, and Events*, Online: Association for Computational Linguistics, Jul. 2020, pp. 97–107. DOI: 10.18653/v1/2020.nuse-1.12. [Online]. Available: https://aclanthology.org/2020.nuse-1.12.

[108] L. H. Somerville, "Special issue on the teenage brain: Sensitivity to social evaluation," *Current directions in psychological science*, vol. 22, no. 2, pp. 121–127, Apr. 2013, ISSN: 0963-7214. DOI: 10.1177/0963721413476512. [Online]. Available: https://europepmc.org/articles/PMC3992953.

[109] L. H. Somerville, T. Hare, and B. Casey, "Frontostriatal maturation predicts cognitive control failure to appetitive cues in adolescents," *Journal of cognitive neuroscience*, vol. 23, no. 9, pp. 2123–2134, Sep. 2011, ISSN: 0898-929X. DOI: 10.1162/jocn.2010.21572. [Online]. Available: https://europepmc.org/articles/PMC3131482.

[110] L. Somerville, "Emotional development in adolescence," *Handbook of emotions*, pp. 350–365, 2016.

[111] L. Sosa-Hernandez, M. Wilson, and H. A. Henderson, "Emotion dynamics among preadolescents getting to know each other: Dyadic associations with shyness," *Emotion (Washington, D.C.)*, Sep. 2022, ISSN: 1528-3542. DOI: 10.1037/emo0001155. [Online]. Available: https://doi.org/10.1037/emo0001155.

[112] S. H. Sperry, M. A. Walsh, and T. R. Kwapil, "Emotion dynamics concurrently and prospectively predict mood psychopathology," *Journal of Affective Disorders*, vol. 261, pp. 67–75, 2020, ISSN: 0165-0327. DOI: https://doi.org/10.1016/j.jad.2019.09.076. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0165032719308006.

[113] B. Stasak, J. Epps, N. Cummins, and R. Goecke, "An investigation of emotional speech in depression classification," in *Understanding speech processing in humans and machines: 17th Annual Conference of the International Speech Communication Association (INTERSPEECH 2016), San Francisco, California, USA, 8-12 September 2016; Volume 1*, 2016, ISBN: 9781510833135. DOI: 10.21437/interspeech.2016-867.

[114] L. Steinberg, "Cognitive and affective development in adolescence," *Trends in Cognitive Sciences*, vol. 9, no. 2, pp. 69–74, 2005, ISSN: 1364-6613. DOI: https://doi.org/10.1016/j.tics.2004.12.005. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1364661304003171.

[115] A. Stickley, A. Koyanagi, H. Takahashi, *et al.*, "Attention-deficit/hyperactivity disorder symptoms and happiness among adults in the general population," *Psychiatry Research*, vol. 265, pp. 317–323, 2018, ISSN: 0165-1781. DOI: https://doi.org/10.1016/j.psychres.2018.05.004. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0165178117317596.

[116] Suhavi, A. K. Singh, U. Arora, *et al.*, "Twitter-stmhd: An extensive user-level database of multiple mental health disorders," *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 16, no. 1, pp. 1182–1191, May 2022. DOI: 10.1609/icwsm.v16i1.19368. [Online]. Available: https://ojs.aaai.org/index.php/ICWSM/article/view/19368.

[117] J. Suls, P. Green, and S. Hillis, "Emotional reactivity to everyday problems, affective inertia, and neuroticism," *Personality and Social Psychology Bulletin*, vol. 24, no. 2, pp. 127–136, 1998. DOI: 10.1177/0146167298242002. eprint: https://doi.org/10.1177/0146167298242002. [Online]. Available: https://doi.org/10.1177/0146167298242002.

[118] D. Teodorescu, A. Fyshe, and S. Mohammad, "Utterance emotion dynamics in children's poems: Emotional changes across age," in *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, Toronto, Canada: Association

for Computational Linguistics, Jul. 2023, pp. 401–415. [Online]. Available: https://aclanthology.org/2023.wassa-1.35.

[119] D. Teodorescu and S. M. Mohammad, *Evaluating emotion arcs across languages: Bridging the global divide in sentiment analysis*, 2023. arXiv: 2210.07381 [cs.CL].

[120] D. Teodorescu and S. M. Mohammad, *Generating high-quality emotion arcs for low-resource languages using emotion lexicons*, 2023. arXiv: 2306.02213 [cs.CL].

[121] R. A. Thompson. U of Nebraska Press, 1990, vol. 36.

[122] R. A. Thompson, "Emotion regulation: A theme in search of definition," *Monographs of the Society for Research in Child Development*, vol. 59, no. 2/3, pp. 25–52, 1994, ISSN: 0037976X, 15405834. [Online]. Available: http://www.jstor.org/stable/1166137 (visited on 07/29/2023).

[123] A.-S. Uban, B. Chulvi, and P. Rosso, "An emotion and cognitive based analysis of mental health disorders from social media data," *Future Generation Computer Systems*, vol. 124, pp. 480–494, 2021, ISSN: 0167-739X. DOI: https://doi.org/10.1016/j.future.2021.05.032. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167739X21001825.

[124] T. Van de Cruys, "Automatic poetry generation from prosaic text," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online: Association for Computational Linguistics, Jul. 2020, pp. 2471–2480. DOI: 10.18653/v1/2020.acl-main.223. [Online]. Available: https://aclanthology.org/2020.acl-main.223.

[125] K. Vishnubhotla and S. M. Mohammad, "Tweet emotion dynamics: Emotion word usage in tweets from us and canada," in *Proceedings of the Thirteenth International Conference on Language Resources and Evaluation (LREC 2022)*, Marseille, France, 2022.

[126] K. Vishnubhotla and S. M. Mohammad, "Tweet Emotion Dynamics: Emotion word usage in tweets from US and Canada," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, Marseille, France: European Language Resources Association, Jun. 2022, pp. 4162–4176. [Online]. Available: https://aclanthology.org/2022.lrec-1.442.

[127] E. Wassiliwizky, S. Koelsch, V. Wagner, T. Jacobsen, and W. Menninghaus, "The emotional power of poetry: Neural circuitry, psychophysiology and compositional principles," *Social cognitive and affective neuroscience*, vol. 12, no. 8, pp. 1229–1240, Aug. 2017, ISSN: 1749-5016. DOI: 10.1093/scan/nsx069. [Online]. Available: https://europepmc.org/articles/PMC5597896.

[128] S. M. Weinstein, R. J. Mermelstein, B. L. Hankin, D. Hedeker, and B. R. Flay, "Longitudinal patterns of daily affect and global mood during adolescence," *Journal of Research on Adolescence*, vol. 17, no. 3, pp. 587–600, 2007. DOI: https://doi.org/10.1111/j.1532-7795.2007.00536.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1532-7795.2007.00536.x. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1532-7795.2007.00536.x.

[129] C. Whissell, "Poetic emotion and poetic style: The 100 poems most frequently included in anthologies and the work of emily dickinson," *Empirical Studies of the Arts*, vol. 22, no. 1, pp. 55–75, 2004. DOI: 10.2190/FWGA-M9DB-P9D4-11X6. eprint: https://doi.org/10.2190/FWGA-M9DB-P9D4-11X6. [Online]. Available: https://doi.org/10.2190/FWGA-M9DB-P9D4-11X6.

[130] J. Zeman, M. Cassano, C. Perry-Parrish, and S. Stegall, "Emotion regulation in children and adolescents," *Journal of developmental and behavioral pediatrics : JDBP*, vol. 27, no. 2, pp. 155–168, Apr. 2006, ISSN: 0196-206X. DOI: 10.1097/00004703-200604000-00014. [Online]. Available: https://doi.org/10.1097/00004703-200604000-00014.

[131] P. Zimmermann and A. Iwanski, "Emotion regulation from early adolescence to emerging adulthood and middle adulthood: Age differences, gender differences, and emotion-specific developmental variations," *International Journal of Behavioral Development*, vol. 38, no. 2, pp. 182–194, 2014. DOI: 10.1177/0165025413515405. eprint: https://doi.org/10.1177/0165025413515405. [Online]. Available: https://doi.org/10.1177/0165025413515405.

Appendix A

Poem Length on UED Metrics

As mentioned in Section 4.3, 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 A.1 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.

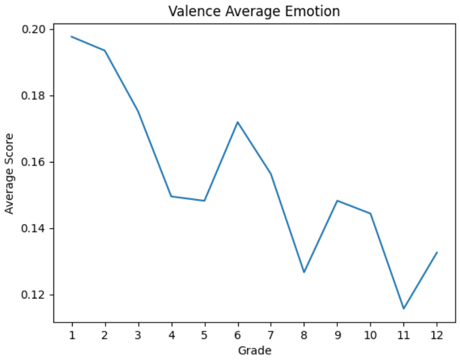


Figure A.1: Average valence across grades for poems of length 10 to 20 words.