The Eras Tour: Investigating the Validity of Music Sentiment Analysis through the Case Study of Taylor Swift

Sarah Baskin

December 16, 2024

Abstract

Is poetic emotion something that can be analyzed – or is it instead something that is intrinsically felt? This is a question that has inspired multitudes of research and debate in music sentiment analysis across the disciplines of Computer Science, Linguistics, Psychology and Musicology, but the disparate and often contradictory nature of such investigations have made it difficult to stake any claims one way or the other. In this study, I attempt to synthesize the various methodologies utilized in music sentiment analysis discussions, using the case study of Taylor Swift's discography as a jumping off point to discuss the capacity of this subfield. Taylor Swift is an ideal subject for a study like this because of an already existing public perception that her music fits into distinct 'eras'. Although there was shown to be a probable negative emotional trend in Swift's music, the findings were generally inconclusive.

1 Introduction

Music has an unparalleled and incredibly unique power to awaken specific emotions in listeners. A single song can inspire complex thought, reminiscence on past events, and positive change in an individual's life and sense of self (Napier and Shamir, 2018). Codifying that ability in music is no easy feat, and has thus contributed to the birth of music sentiment analysis research (Rebora, 2023). Sentiment analysis (referred to from now on as SA) itself is an area of study with a layered background, encompassing many different subdomains and research areas. SA, also known as opinion mining or emotion mining, aims to systematically identify peoples' opinions and attitudes as they're expressed through a particular medium (Wu et al. (2022); Medhat et al. (2014); Napier and Shamir (2018); Nair (2023)). This work does not necessarily need to be wholly computational. In fact, Automatic Sentiment Analysis (henceforth referred to as ASA), is quite new to the field, and theoretical frameworks have a far more storied history (Rebora, 2023). Stylometry is one of these such viewpoints, which counts linguistic features of interest (Whissell, 1996). Stylometry is often used in music SA, with particular deference given to word length, word repetition, negation, and definite article counts (Whissell, 1996), as well as pronoun use (Berg, 2016) and immediacy markers (Petrie et al., 2008). These metrics, when combined, can be used to identify particular emotional arcs in music. Fully complete SA work additionally includes a human element, wherein annotators either corroborate or contradict SA work by analyzing the texts thematically as opposed to through counts or computational analysis (Rebora (2023); Wu et al. (2022); Sacco and Pettijohn II (2009)). For this study, I was unable to complement my ASA and Stylometric work with human annotation, but I have outlined a proposition for how I would go about human annotation in 6.

Music SA must also contend with assumptions, however, that its techniques will always fall up short of truly being able to account for intricacies in emotionality (Rebora, 2023). Poetry is the basis of most modern music imaginings, and thus as a consequence, songs are allowed an inherent complexity not reserved for more traditional targets of SA such as reviews and social media posts (Nair (2023); Kao and Jurafsky (2012)). This leads to natural questions of authenticity in the field, doubts which are exacerbated by the perception of music SA as a "low-cost, low-gain" science (Ma et al., 2024). As of yet, the lack of unification in the field has made it difficult for the necessary reflections to occur, but in this study, I hope to begin to take steps to remedy those oversights.

I have chosen the discography of Taylor Swift (All Music) as the basis for my SA analysis, as I believe that her unique stature in the music industry makes her an ideal candidate for this task. She has not only reached a certain peak of longevity as an artist, but boasts a career crossing multiple genres and a persona built around the idea of having distinct "eras" in her music (Hernández Rodríguez, 2024). I will measure the positive and negative emotional slants in Swift's music over time, as well as analyze more curated notes such as immediacy, agency, sociability and wholesomeness first through stylometric counts of unique words, word length, curse words, and pronoun usage. Then, I will transition to the incorporation of ASA tools, which will measure her songs on a trinary (meaning: positive vs. neutral vs. negative) scale. This task will be set for VADER (Hutto (2024), TextBlob Loria (2013), and AFINN Nielson (2018)), Python libraries designed specifically for SA.

2 Literature Review

I will begin with an overview of previous studies and research data on the topic of music SA. Because this subdomain encompasses so many competing fields and can be approached in such a wide variety of ways, I have further divided the literature review section into more clearly defined units.

2.1 Music Automatic Sentiment Analysis

In the transition to applying ASA techniques to music from the more traditional target of customer reviews, there is generally not much change made to the underlying model architectures. This is an interesting choice, given that many of the more refined models commonly tuned for these sorts of tasks were not originally designed to pick up on emotion keys specifically in music (Napier and Shamir, 2018). The data that these models are generally trained on tends toward bluntness and more overt expressions of opinion. Although it would make sense for fine-tuning tuning to then occur, many researchers' lack of a computational background prevents them from fully understanding the effective steps that will ensure peak model accuracy and performance (Napier and Shamir, 2018).

ASA models tend to be built on at least a semi-psychologically motivated backbone, taking their implementation of emotional benchmarks from theories of emotion realization such as Plutchik (2001). As far as the actual application of an emotion theory-focused framework, a given model will typically assign a valence score (typically falling between -1 and 1 or 0 and 1, depending), based on how present or absent a particular measure (ie. sadness, happiness, fear, extroversion) is in a text (Napier and Shamir, 2018).

lexicon-based models (of the type currently included in my experimentation) specifically provide a mapping for given content words. This relation allows an instance to be associated to a number that dictates the positivity, negativity, or neutrality of a word (Cho et al., 2014). The number itself is typically derived from annotation or polling, meaning that their is some human element underlying the rankings. An algorithm is then applied to perform a summation and normalization of all such content words present in a particular document in order to return overall sentiment scores.

The applications of music ASA are quite varied, and occasionally disjoint. However, the general implementation and structure remains the same. As the most commercial and widely-known genre (see 2.3 for a more in-depth analysis), popular music tends to be on the receiving end of ASA analysis, with song data nearly always ripped directly from the Billboard 100 (Wu et al. (2022); Napier and Shamir (2018). The motivations behind research tend to vary by discipline, but the underlying reasoning is usually either aligned with investigating the symbiotic relationship between music and society (Napier and Shamir (2018); Petrie et al. (2008); Wu et al. (2022); DeWall et al. (2011)) or building more accomplished music recommendation systems (Raschka (2016); Gukasyan (2023); Celma (2006); Schedl (2019)).

When pursuing the former trajectory, the emphasis tends to fall on sentiment change over time, that is, seeing how various emotions have shifted in prevalence in the Billboard charts up until the present day (Napier and Shamir, 2018). This research rests on the principle that society chooses music that best reflects their own emotional leanings, therefore the charts form a representative sample of the popular emotionality of a given point in time (Napier and Shamir (2018); Wu et al. (2022)).

2.2 Music Stylometry Analysis

There are certainly comparisons to be made between Stylometry and ASA. Both can be criticized for being a 'cold' method of textual analysis (Whissell (1996); DeWall et al. (2011)), but unlike ASA, Stylometry only relies on feature counts (Whissell, 1996), without further considering the relation between individual elements or a contribution to a greater whole (Whissell, 1996).

Stylometry has a similarly long history in music analysis, and participates in comparable types of studies. Its techniques are often used as a gauge to track specific societal trends and views over time (Batcho (2007); Sacco and Pettijohn II (2009); Lena and Peterson (2008); Christenson et al. (2019); Pettijohn and Sacco (2009); Christenson et al. (2012); Eastman and Pettijohn II (2015); Madanikia and Bartholomew (2014); Schellenberg and von Scheve (2012)), again with the Billboard Hot 100 as a base point for gathering data. The difference between the stylometric application of this work and the automatic application, however, is that instead of focusing on sentiment over time, the impetus is instead on features that are directly linked to social issues.

Previous work has analyzed the evolution of the depiction of women (Bretthauer et al. (2007); Cooper (1985); Flynn and Craig (2016)), perception of drugs and alcohol (Christenson et al., 2012), and the acceptance of sex and other markers of love and romance (Madanikia and Bartholomew, 2014). Stylometric music research that has involved the explicit analysis of sentiment is generally conducted through the lens of using emotion as a tool to reveal greater clarity on these more distinct topics (Whissell, 1996).

Pronouns are of particular interest in Stylometry. A significant utilization of second person pronouns (ie. "you", "your"), for example, has been linked to many socially-conscious themes, such as the idea of societal cohesion and the de-emphasis of individual narratives in favor of focusing on the whole (Whissell, 1996). Similarly, first person plural pronouns (FPPP) have been shown to create a sense of reduced social distance between a musical artist and a listener (Berg, 2016).

First person singular pronouns (FPSP) show up regularly in club and dance music, and are strikingly present in popular music due to the genre's narrative-heavy style (Berg, 2016). This narrative focus is in large part why high FPSP usage is correlated to high levels of immediacy – wherein immediacy refers to the centering of song in the present moment (Berg, 2016). Short words and present tense markers are similarly linked to immediacy, and are also regularly seen in pop music specifically (Petrie et al., 2008). Independently, sad and melancholy songs also exhibit a pattern of high FPSP rates (Berg, 2016).

Outside of pronouns, tense, and word length, profanity is also tied to specific sentiment realizations. Profanity is closely linked in the popular imagining to 'anti-social' behaviors and genres, a stigma it maintains in its ability to cause perceived sentiment to be more negative (Ballard et al., 1999).

Unlike in ASA, copious amounts of data is not always a necessity for stylometric analysis, since no models are trained or fine-tuned in this process. Therefore, it is in this realm that a greater number of studies focusing on specific artists take shape (Petrie et al. (2008); Whissell (1996); West and Martindale (1996); Whissell (2008)).

The most frequent artist whose work is chosen for this kind of analysis is The Beatles (Petrie et al. (2008); Whissell (1996); West and Martindale (1996)), generally with the intent of quantifying long-thought distinctions between songs written by John Lennon, Paul McCartney, and the Lennon-McCartney partnership. As a key point in these studies, the outside forces that influenced the band are an integral talking point, as many of the stylistic changes can be given better context when also considering where the songwriters were looking for inspiration.

2.3 Defining Genre in Music

It is clear that genre distinctions can do a lot for the realization of sentiment in music. From that understanding, it then becomes necessary to define the ways in which genres differ from each other in order to get a better sense of emotionality tendencies. Since the study deals exclusively with the music of Taylor Swift, it makes the most sense to center this discussion around the two genres in which she has made her career: country and pop.

For country music, a baseline of genre expectation is not hard to locate. Thematically, country songs are generally perceived as being a soapbox for more traditional, conservative, family values (Van Sickel, 2005). In romantic narratives, the focus tends to be on marital hegemony, and the centering of heterosexual, male experiences (Van Sickel, 2005). This is further emphasized by a pride and nostalgia for the American South, complemented by a vague irritation with societal change and an obliviousness to causes such as the fight for racial equality (Van Sickel, 2005). Country is also often characterized as a practically 'verbal' art form, where the rhythmic and harmonic structures are kept purposely simple in order to position the lyrics as the core vehicle behind a song's message (Van Sickel, 2005). All of these individual facets come together to brand country as a more 'wholesome' than other competing genres, with a distinct emphasis on individual narratives and a centering of the artist almost as a storyteller speaking directly to the listener (Ballard et al., 1999).

Popular music is completely different to the relative uniformity of country. This is due, first and foremost, to the inability to dilute pop music down to a simple description that covers the majority of content under its umbrella. Any such definition is inherently unstable, as pop is better conceptualized in terms of its commerciality, distribution patterns, mediation, and constantly shifting fashions and vernacular discourses (Schellenberg and von Scheve (2012); Holt (2007)). The term 'pop music' has been used to encapsulate of a revolving door of genres, ranging from blues and jazz in the 1920s and 30s, to rock and soul in the 70s and 80s, to the dance and hip-hop of today's zeitgeist (Holt, 2007). Therefore, it makes more sense to think of pop music as a generic term to categorize the 'mainstream' music scene, describing the production machine generating hits and stars targeted at a largely adolescent audience (Christenson et al., 2019). At its core, pop is the music found on the Billboard Hot 100, with distinguishing genre characteristics purposely obscured in the name of obtaining wider mass appeal (Lena and Peterson, 2008); (Ortega, 2024).

As I will discuss in 2.4, Swift's rise is generally perceived as a shift from country to pop, with Red and 1989 largely serving as transition albums between the two behemoths. However, instead of envisioning her as a cross-genre artist, it will potentially prove more accurate to say instead that her music 'transformed' into pop music over time as she herself became more enmeshed in the cultural scene. As her star rose, her music consequently become more accessible, slowly shedding the elements that had previously defined it clearly as country. Due to this shift, and due to the wide range of tastes and preferences contained within the pop space, it can be anticipated that there will be greater stylistic, emotional, and creative variety present across Swift's pop discography compared to her country albums.

2.4 The Artist Persona of Taylor Swift

Few artists today can dream of amassing Taylor Swift's level of fame and pop-culture dominance (Hernández Rodríguez, 2024). She has established herself not only as a bastion of Americana (Hernández Rodríguez, 2024), but as a key cultural figure, with her notoriety often compared to that of 'American royalty' (Volpe, 2021).

This image did not coalesce overnight. Rather, it is the result of a marked evolution in her persona as an artist. This change can be traced most noticeably in her music 1, as, since she is the writer of a majority of her own songs, her life experiences have a direct impact on her lyrics (Hernández Rodríguez, 2024). It is because of this deep personal connection that she is most famous for her narrative storytelling (Hernández Rodríguez, 2024), with her lyrics often compared to poetry for their extensive use of irony (Marpaung et al., 2023) and figurative language (Fatikha and Masykuroh, 2022).

Album Title	**Year**	**Record Label**
Taylor Swift (Debut/TS)	2006	Big Machine Records
Fearless	2008	Big Machine Records
Speak Now	2010	Big Machine Records
Red	2012	Big Machine Records
1989	2014	Big Machine Records
Reputation (Rep/rep)	2017	Big Machine Records
Lover	2019	Republic Records
Folklore	2020	Republic Records
Evermore	2020	Republic Records
Midnights	2022	Republic Records
The Tortured Poets Department (TTPD)	2024	Republic Records

Table 1: Taylor Swift Discography

Her skill at crafting narrative songs served her well during her early years as a country artist, given that genre's propensity to value storytelling over musical novelty(Van Sickel, 2005). Now as a bastion of popular music, it is that same narrative focus that sets her apart from many of her contemporaries in the field.

The types of stories that Swift chooses to tell are often indicative of her place and power in her professional career. Although she has perennially focused on romances, the framing has shifted dramatically since her debut. In *Taylor Swift* (2006), *Fearless* (2008), and *Speak Now* (2010), her role within a song was emphasized as passive; a damsel in distress waiting to be saved by a lover (as in "Love Story" (*Fearless*)), pining for a boy who does not notice her (as in "Teardrops on My Guitar" (*Taylor Swift*)), or serving as a second person narrator of sorts to the thoughts of her romantic interest (as in "Mine" (*Speak Now*)). This sort of work is expected by typical genre constraints of country music, however in *Speak Now* there are definite flashes of more independence and a desire for greater control and agency (Hernández Rodríguez, 2024). This is most noticeable in "Mean", where Swift directly challenges press outlets that had previously belittled her, and "Back to December" as well, her first apology song, where she takes responsibility for mistreating a previous boyfriend.

These first glimpses of breaking with genre tradition came to a head with the release of *Red* (2012), which positions Swift as a character with far more agency in her songs – closer to a hero than a damsel (Hernández Rodríguez, 2024). Two key examples of this shift are the darker "I Knew You Were Trouble", and the peppy ballad "22". In her next album, *1989*, (2014) this transformation into a heroine is even more pronounced, with many of the songs on the track list seemingly representing particular steps of the hero's journey (Hernández Rodríguez, 2024).

If 1989 is the apex of Swift's hero persona, then 2016's *Reputation* is the anti-thesis. This transformation was necessitated following a very public feud and falling out with Kanye West, an altercation which received significant media attention and ultimately resulted in a heavy backlash against Swift (Hernández Rodríguez, 2024). In the wake of the downfall of her hero image, the anti-hero had been born (Hernández Rodríguez, 2024). This was accompanied by an embrace of a darker, more tortured aesthetic that has continued through various incarnations into her current work (Hernández Rodríguez, 2024). A desire for vengeance and a greater emphasis on sexuality permeate her recent output (with the noted exception of *Lover*, which is by and large a return to the more mainstream bubblegum pop of *1989*), and a sense of internal torture and ostracization, particularly in Evermore and Folklore (Hernández Rodríguez, 2024).

This trajectory can be used to further clarify the myriad narrative threads of her music through the years, laying the groundwork for analysis and discussion of the differences across Swift's discography.

2.5 Methodology of Music Annotation

Annotation in music studies, though a key component to verifying the results of SA, is usually treated as less of an integral part of an experiment. When annotation metrics *are* incorporated into a music SA study, though, it tends to follow quite predictable procedures.

First and foremost, the team of human annotators must work separately from the rest of the experimental group so as to ensure that they are not influenced in their deliberations. Coders themselves are generally chosen because they are in some way representative of a population, controlled for gender, race, and economic background so as to limit confounding variables in sentiment evaluation (Papp-Zipernovszky et al., 2022). Since academic research takes place on college campuses, annotators are generally college students participating in data collection either for credit or payment (Ballard et al., 1999). In other cases, annotators are simply members of the research team who work asymmetrically from the majority of the group over the course of evaluation (Wu et al., 2022).

The specific tasks that annotators are expected to complete will vary depending on the nature of the experiment. Participants may be asked to rate the sentiment of a text generally on a binary (or trinary scale) (Christenson et al., 2012). In cases featuring an ASA model, coders would instead rate sentiment on a parallel scale to the tool under study (Wu et al., 2022). If the focus is on unlocking the poetic potential of a work, questions about mental imagery, descriptive keywords, and favorite lines may be thrown around instead, as well as free write opportunities to further qualify opinions (Papp-Zipernovszky et al., 2022). When the musical corpus under study is particularly large, the number of annotators tends to be relatively small, usually a group of two to three. In these cases, inter-annotator agreement and understanding becomes key to the process, as it allows for participants to talk through their analyses and come to terms with how they view a particular piece (Christenson et al., 2012).

2.6 Shortcomings of Sentiment Analysis in Music

It has been said that "writing about music is like dancing about architecture" (Barrington et al., 2007). When dealing with an art form that is inherently layered, connotative and specific to each listener, it can be exceedingly difficult to come away with a singular analysis of sentiment that is fully representative of the interpretations contained therein (Tagg (1987); Napier and Shamir (2018); Volpe (2021)).

Additionally, due to the inherently abstract nature of song lyrics, some emotions that are expressed may escape linguistic formalization, and therefore are untraceable by SA standards despite being perfectly understood by listeners (Rebora, 2023). The converse, however, can also be true. Even if an emotion *is* indeed formalized, the true sentiment of a piece can sometimes be completely divorced from the narrative message it conveys, leading to additional incorrect readings by SA algorithms (Rebora, 2023).

This does not even begin to unpack what is perhaps the most glaring current contradiction in music SA studies: the assumption that the writing process begins with lyrics, with an artist only adding the musical part later on in the process (Marouf et al., 2019). Although this line of thought makes it possible to analyze the text of a song and come away with a representation of that song's sentiment, it is an inaccurate representation of the creative process and also ignores the fundamentally layered aspect of music production (Marouf et al., 2019).

Song lyrics will always be interacted with in the context of the musical aspects of performance (Street, 2001). Beat, background vocals, harmonies, riffs, they are just as much a part of a song as the lyrics. This is most keenly observed in the existence of covers, which famously can put entirely new spins on the same lyrics by changing aspects of the production (Ortega (2024); Street (2001)).

Revealingly, when listeners are asked to associate certain trends in music with happiness or cheerfulness, their first thought is not in regards to peppy lyrics, but rather faster tempos and higher pitches; in music, it is quite fittingly the *music* that sets the tone for sentiment (Yeh et al. (2014); Volpe (2021). Adolescents (the major purveyors of pop music) even cite lyrical content as the least important reason for liking a song, preferring rhythm, vocals, and melodic affects (Ballard et al., 1999).There are numerous examples of songs that are falsely perceived as happy due to their tempo and "supra-lyrical" touches, when in fact the narrative of the text is quite depressing (Schellenberg and von Scheve, 2012).

A propensity to rely on models that are limited in scope can also make music SA data more untrustworthy. Traditional positive vs. negative systems, or more generally, systems not built with music in mind at all, will naturally fall up short due to a mismatch in purpose. Music relies on such a complex array of emotional cues, and operates with such a small number of words, that current approaches can often fail to properly represent less pronounced distinctions (Rebora, 2023).

These oversights are not due to some intrinsic issue within music SA, but are largely a by-product of the lack of precedent in such a young field. ASA (and some stylometric) technologies are still quite new, and the idea of music SA did not even fully gain traction until the late 2010s. As a result, most methods still lack a consistent validation method, and connections to the more literary and poetic side of music are not as well-established (Rebora, 2023). Organizational efforts to streamline research processes and establish widely accepted metrics and codes are lacking, and thus every new experiment, although potentially novel, must deal with the question of praising originality over reliability (Ma et al. (2024); Rebora (2023)).

I should note, though, that this criticism comes from a place of optimism. I waylay any immediate refusal of SA as a worthwhile research pipe line, and instead think of the challenges ahead as a testament to the exploratory potential of this field (Rebora, 2023).

3 Method

Uniting together the findings from the preceding sections, I will marry a combination of ASA and stylometric techniques for my experiment, weighing them both in tandem in order to craft a more complete picture of sentiment evolution in Taylor Swift's discography. [1] [2]

I will pay particular attention to the potential pitfalls that I highlighted in 2.6, as well as the career timeline that I laid out in 2.4, dissecting emotion in regards to both of those lines specifically. Due to time constraints, I was unfortunately unable to include a human annotation procedure in my pilot study, however I have laid out a methodology for such an experiment in 6.

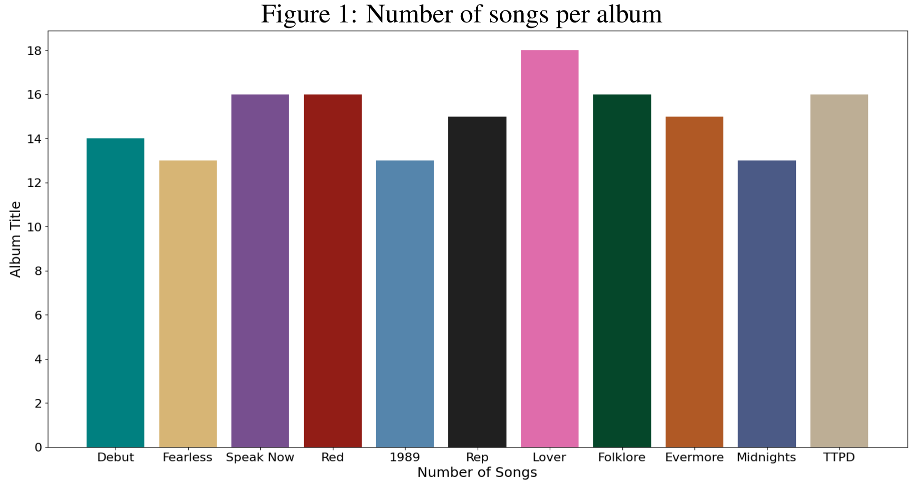
[1] In my graphs, I have chosen to use shortenings for some of Swift's album names for the sake of formatting. All of these shortenings have basis in her fandom, and are regularly used by "Swifties". "Debut" corresponds to *Taylor Swift* (2006); "Rep" corresponds to *Reputation* (2017); "TTPD" corresponds to *The Tortured Poets Department* (2024).

[2] The colors in all graphs are chosen deliberately, and reflect the color scheme that fans tend to associate with a particular album. This is done purely for stylistic effect, and to make it easier to distinguish and identify each album efficiently.

The three sentiment analysis models that I will be using to constitute the ASA portion of the experiment are VADER (Hutto, 2024), TextBlob (Loria, 2013), and AFINN (Nielson, 2018). These are all lexicon-based trinary models, as I discussed previously in 2.1. Although there were other more complex models that I would have liked to have included, I was limited by paywall access.

For stylometric metrics, I am analyzing word count, word length, first person pronoun (FPP) use, and profanity.

Figure 1: Number of songs per album



As can be seen in 1, there is no concerted shift in the number of songs per album over time. Still, there are some deviations, and some albums with a quite high rate of difference than other albums. As a result, counts and sentiment will be calculated as a proportion per song per album. This follows logically since, if there are more songs in an album, the count of unique words in that album will be greater, meaning that there is a greater likelihood for the mass of lexicon-coded words to artificially tip sentiment scores one way or the other even if in reality there is far less of a demonstrated difference.

Please be aware that lyric data was taken from Taylor Swift's re-released albums (marketed as "Taylor's Version") when applicable, meaning *Fearless*, *Speak Now*, *Red*, and *1989*. There are many reasons motivating Swift's decision to re-release her Big Machine Records albums, none of which are expressly relevant to the current study, however it should be noted that I have excluded the "From the Vault" and "Remix" tracks that were included as part of the "Taylor's Version" releases. This is because I feel very strongly that it is important to represent an album as it was at specific points in Swift's career. Although the nature of the re-releases means that the only 'new' songs included are ones that were recorded at the time of original release but were cut, it is a reflection of where Swift was professionally that she was unable to argue for their inclusion (Volpe, 2021). Including those songs in my analysis would not accurately show what her curated persona demanded of her and her songs at that point in time.

All lyric data was adapted from Kothari (2024). The data tokenization process is discussed in A

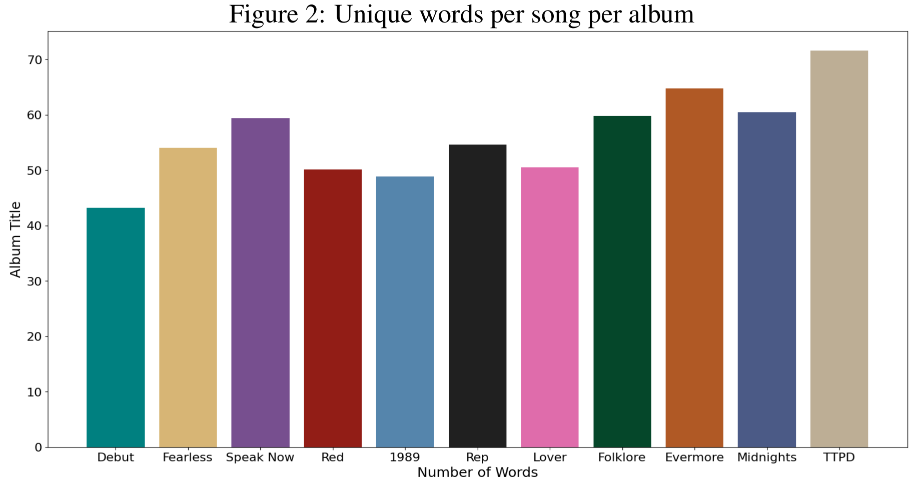
I will be determining the statistical significance of change over time in Swift's albums by using the Pearson

Correlation Coefficient as well as ANOVAs to quantify both linear progression as well as comparison across ’genre’ groups (ie. how different is Swift’s country era from her pop era? How different are Swift’s early pop years from her later pop years?)

4 Results

I will begin by discussing the results from the stylometric portion of my study, before transitioning into the role of ASA.

Figure 2: Unique words per song per album



In 2, it is observed that the average number of unique words per album undergoes a substantial shift over time, with a Pearson Correlation Coefficient of 0.685 and a p-value of 0.01. This is the type of behavior that we would expect to see given previous stylometric work, as pop music has been shown to have higher levels of immediacy than country. The fact that we see higher levels of immediacy in Swift’s later, more pop-heavy work than compared to her days in country is a reflection of this pattern.

3 complicates this assumption somewhat, as it can be seen that there is essentially no variation in word length per song over the eleven albums in question. If there is a northward shift in immediacy, as suggested by 2, we would also expect to see a decrease in word length per song over time. Here, this is not the case, potentially raising the question of which lexical reflection of immediacy is able to most succinctly contribute to a song’s representation of the metric.

4 tells a similarly contradictory story, as neither FPSPs, FPPPs, nor SPPs are correlated over time. We would expect to see greater FPSP use into Swift’s pop era if immediacy increases over those albums.

Considering that a trend is only observable to this point in 2, and the other common markers of increased immediacy are absent, I do not feel that I am able to say that the data shows a definitive change in immediacy, though of course in order to make such a conclusion, I would need to investigate whether or not unique words and pronoun use

Figure 3: Word length per song per album

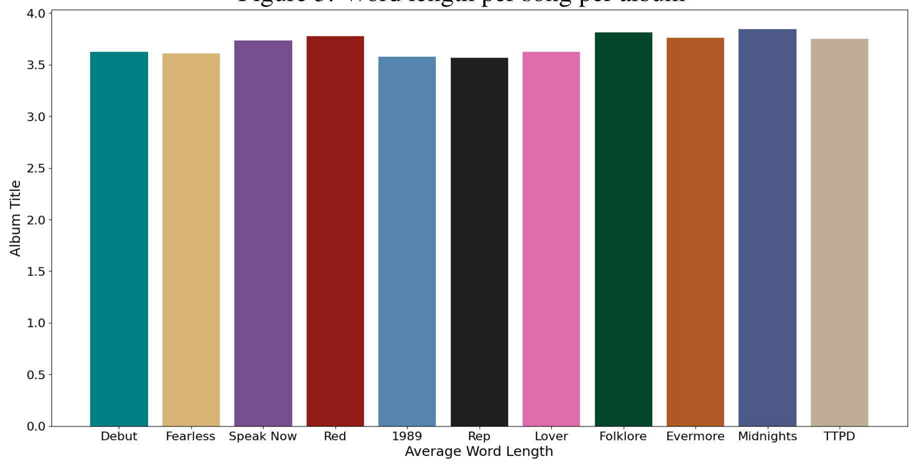
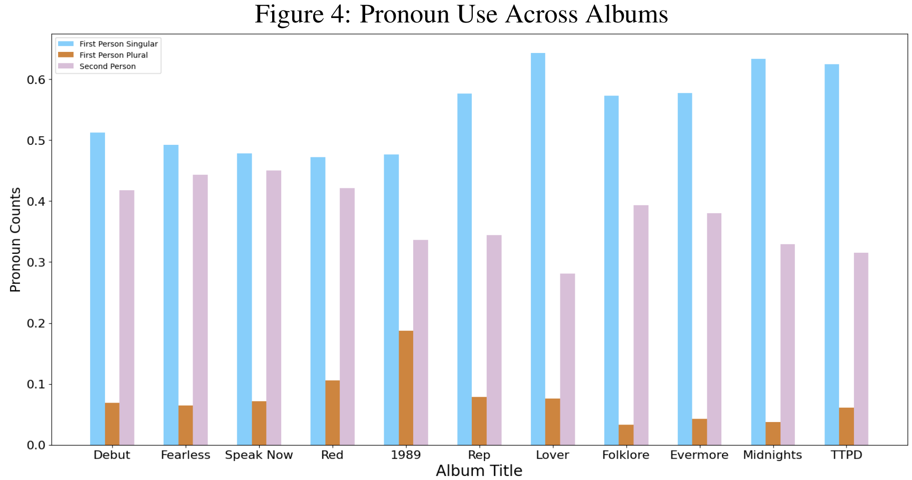
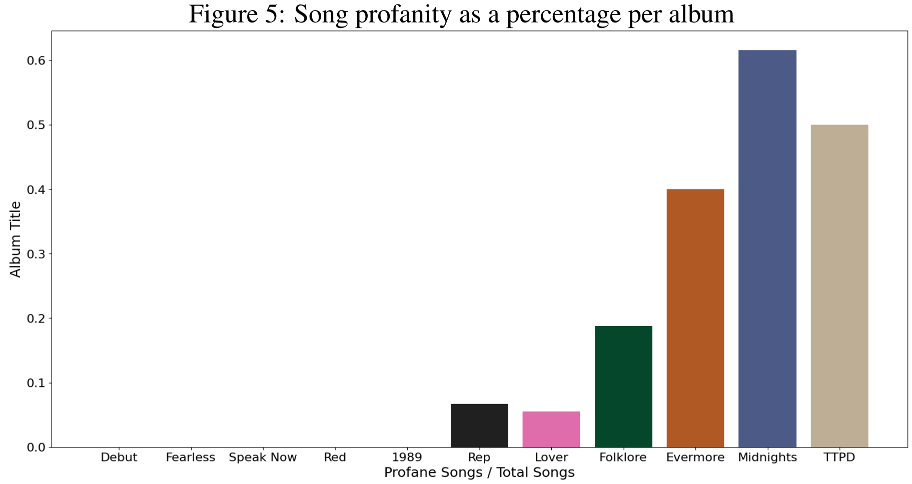


Figure 4: Pronoun Use Across Albums



have different impacts on immediacy in song.

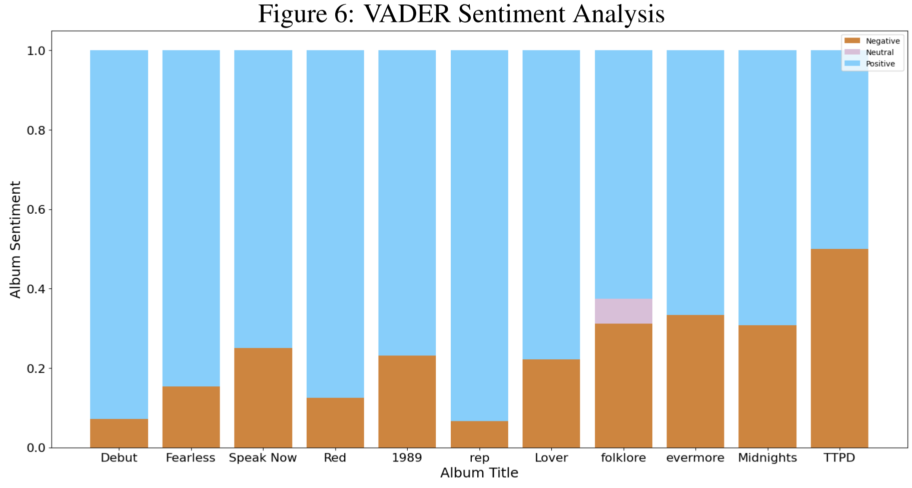
Figure 5: Song profanity as a percentage per album



With song profanity, there is a very clear upward trend in albums containing songs with a higher percentage of swear words, with a Pearson Correlation Coefficient of 0.792 and a p-value of 0.003. This suggests that Swift's songs engage more with profanity the further into her pop era she goes. This is consistent with findings about the nature of country and pop as music-making engines. As was discussed in 2.3, country is perceived as more 'wholesome' than other genres, and upholds traditional values. This is reflected musically in fewer curse words, leading to an overall happier sentiment expression in country songs. When swear words are included in music, the overall effect is generally to hamper the sentiment of a song, adding notes of sadness, anger, and anti-social tendencies. These notes would be consistent with findings of Swift's evolving persona as discussed in 2.4. From the profanity data, it is then suggested that Swift's pop songs exhibit more melancholy emotions than her earlier work.

Transitioning into discussing the results of the ASA models, I will first explain exactly how I set up my ANOVAs. I did this two different ways to test specific processes. I first created four categories: "country" (*Taylor Swift*, *Fearless*, *Speak Now*), "country to pop" (*Red*, *1989*), "pop" (*Reputation*, *Lover*), and "experimental" (*Folklore*, *Evermore*, *Midnights*, *The Tortured Poets Department*) and used the albums sorted into each category to calculate the ANOVA. Additionally, I had also wanted to see if the variation between the pop albums (everything "country to pop" and proceeding), was greater than the variation between the country albums. This exploration was motivated by my findings in 2.3. I set up a different ANOVA calculation for that, using the same procedure as previously, except omitting the country songs from the analysis.

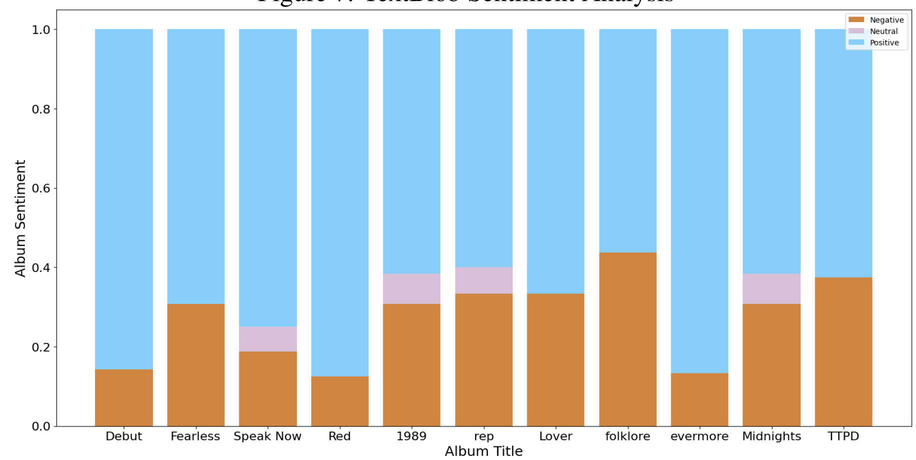
Figure 6: VADER Sentiment Analysis



I will start with the results from VADER. In 6, it can be observed that there was a statistically significant negative lilt over time, with the Pearson Correlation Coefficient being 0.737 with the associated p-value of 0.009.

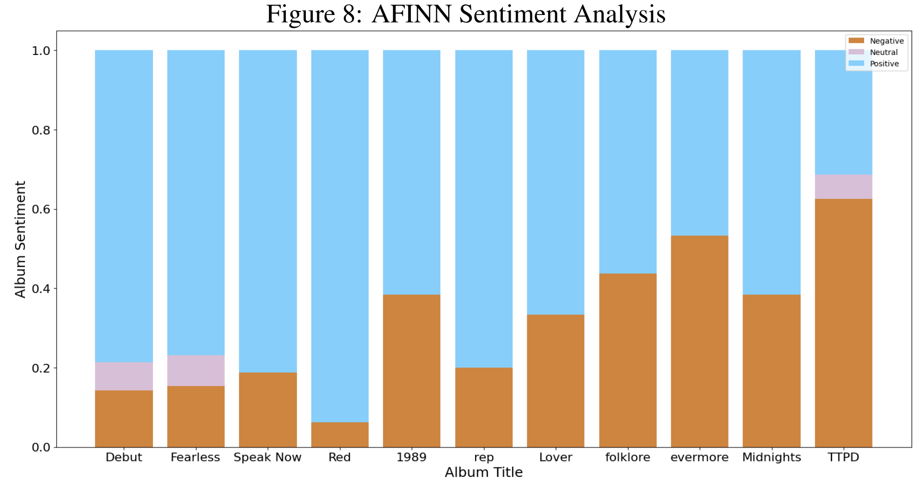
The ANOVA f-stat including country, however, was not found to be statistically significant, and neither was the f-stat for just the pop songs. This implies that relative to the specific "eras" that I created for the experiment, there was more change internally among those albums than when compared to the albums outside of that era.

Figure 7: TextBlob Sentiment Analysis



Unlike in the VADER model, in 7 there is no statistically significant shift in sentiment one way or the other. This is similarly reflected in both ANOVA calculations, where there is no statistically significant trend in sentiment over the albums. This brings into question the findings of the previous model, as it would be expected that another lexicon-based model would in fact be more likely to corroborate the VADER findings.

Figure 8: AFINN Sentiment Analysis



Finally, I will turn to the findings form AFINN in 8. Here there is again an observable linear trend toward greater negativity across the albums. The Pearson Correlation Coefficient for AFINN is 0.826, with a p-value of 0.001. For the ANOVA calculation, there was in fact shown to me greater distinction along the country-pop divide than just within the pop albums themselves. The ANOVA f-stat for the former calculation was 5.4, with a p-value of 0.03. The secondary ANOVA was found to not be statistically significant.

None of the models were able to fully complement one another, with all of them providing slightly contradictory findings. This is likely a reflection of the inconsistency inherent in using trinary architectures to perform ASA, and it is a sign that current methodologies are unable to fully account for nuance in music when evaluating emotional presence.

5 Conclusion

Reviewing the findings asserted in the previous section, it was first found that the number of unique words per song increased over time, implying a potential rise in immediacy in Swift's songs. However, this was not upheld by the counts for pronouns or word length – also correlated with high immediacy – as these numbers stayed stagnant. Profanity did increase in Swift's albums over the years, as expected by her transition away from a 'country' sound, potentially also implying that her lyrics got more sad and anti-social over time.

The ASA models were not universally able to back up this claim, however. The VADER model did observe a linear trend toward greater negativity across Swift's albums, but did not pick up on any major changes across the purported "eras" themselves.

The TextBlob model was largely inconclusive, with no statistically significant linear trend or ANOVA measurement observed.

The AFINN model showed a clear linear trend toward negativity over time, as well as individual evolutionary markers within each era, courtesy of the ANOVA calculation. The AFINN model, however, did not uphold the findings of the VADER model, showing that the sentiment distinctions across eras was indeed noteworthy.

These inconsistencies in some ways demand further data study. For example, other artists' work could be the subject of this kind of analysis in order to determine how Swift's feature counts compare to writers of other genres. This could help to determine relative to a baseline how immediate, anti-social, and person-focused her music is in actuality.

In other ways, the inconclusive data can be seen as a mark of the current shortcomings of music SA. Three relatively similar ASA models fed the exact same music data all came up with completely asymmetrical analyses. This is a sign that the current trends of taking SA models trained for non-music tasks and then feeding them lyrics is not a experimentally sound infrastructure, and more time must be invested into creating SA models, or fine-tuning SA models, to be better equipped to learn from lyrics specifically.

From this pilot study, I conclude that, given the case study of Taylor Swift, music SA is by no means a solved task, and in fact the validity of the field as it stands is deeply tenuous.

6 Future Research Areas

There are a number of aspects of this project that I hope to continue working on following the end of this semester. I want to formally take steps to create some baseline validation criteria for music SA. From my research, teams used a wide variety of statistical metrics and data cleaning methods, none of which seemed to be decided for a particularly astute reason. The first step toward this goal will be to create a list of "music stop words" (ie. words that are incredibly common in music that you would want to remove easily from lyric data so as not to confuse your analysis). In my experimentation, I found that normal stop word libraries were not well-suited for music SA, and so I had to manually fix a number of issues by hand.

I also hope to continue the project by means of locating more complex ASA models. I would like to be able to further test the limits of SA in music, measuring models against one another and seeing how consistent the findings are. I would want to explore beyond the current confines of lexicon-based models, as I think that it would be far more interesting to see how models of different backgrounds perform on this task. I am interested in comparing models that analyze audio sentiment as opposed to lexical sentiment (or both at once) in order to ascertain the current strengths and limitations of lexical-only analysis, which is highly pervasive in the field. I am also considering the possibility of working to build my own SA model that is fully able to take into account all of the aspects that I have discussed, though it would be more ideal to focus on fine-tuning existing models since the former option would necessitate a large time investment.

These are all more incidental next steps that I hope to take on, but in terms of a more formal pilot study, my goal would of course be to include some aspect of human validation in this process, as that adds a greater level of understanding of how accurate ASA and Stylometry actually are in the field of musicology.

My annotator study proposal takes certain aspects from Papp-Zipernovszky et al. (2022); Wu et al. (2022); Rebora (2023); Kim and Klinger (2018), while also incorporating my own ideas and accounting for the perceived oversights of other researchers.

I would organize a team of six annotators total. These annotators would likely be university students, faculty, or staff given that the study would be centralized at Brandeis. The annotators would be pre-screened for previous knowledge of Taylor Swift, with the ideal candidate resting somewhere in the middle ground both in their opinion of her as well as their view of her music. It would be acceptable if people had heard her work (she is incredibly famous, of course, and so it would be very difficult to locate participants completely unfamiliar with her), but more-so on the lines of having heard a song or two on the radio, or recognizing a few lines here or there. Both superfans as well as "super-haters" would be removed from the line-up, as those opinions can color one's view on the sentiment of a song. Additionally, as much as possible, I would want to control for music taste in the annotator group. I would not want all six annotators to be massive pop fans, but rather I would prefer that the majority of the music they enjoy come from a wider spectrum of genres. I am not sure at this point how many other demographic data points I would want to control for, but at the very least I would attempt to get an even sample of age ranges.

From the six-person annotation group, they would be further divided into two teams: one that would evaluate the sentiment of text-only Taylor Swift lyrics, and another that would evaluate the sentiment of audio files containing recordings of Taylor Swift's songs. Both groups would interact with the entirety of Swift's discography, just split over these two distinct platforms. The songs would be staggered and ordered randomly, limiting the possibility of annotators attempting to generalize across albums. The names of the songs would also not be included so as to limit participants from identifying particular songs. The entire experiment would take place over a number of months, with teams listening to at maximum 10 songs per week so as to avoid burnout.

Both groups would have the same rating scales for all of the songs. They would rate each song simply on a trinary scale, before answering more specific questions that would reveal their thoughts. The questionnaire is not yet set, but I would endeavor to include such ideas as: How did this song make you feel?; What images were evoked for you with this song?; What specific lines stuck out as important? Annotators would be encouraged to answer the questions as they read/listened, showing how changes throughout the song had an effect on their impressions.

Weekly meetings would be arranged for the two annotating groups, where each team of three would meet separately to break down any ongoing thoughts related to the songs, share their current opinions on sentiment, and put together a cumulative report of their collective views thus far. If there was major disagreement on any of the sentiments, having three annotators total would make it possible for "tie breaks", where two participants could override a third's opinion for what to write in the report. This would also be a time for annotators to pose questions or express confusion about any of the songs, as applicable.

This pilot study would complement the further computational work that I outlined above, and the two would be combined together in order to come away with a greater understanding of the validity of SA's application to music.

References

AllMusic Taylor Swift. URL `https://www.allmusic.com/artist/taylor-swift-mn0000472102#discography`.

Mary E. Ballard, Alan R. Dodson, and Doris G. Bazzini. Genre of Music and Lyrical Content: Expectation Effects. *The Journal of Genetic Psychology*, 160(4):476–487, December 1999. ISSN 0022-1325, 1940-0896. doi: 10.1080/00221329909595560.

Luke Barrington, Antoni Chan, Douglas Turnbull, and Gert Lanckriet. Audio information retrieval using semantic similarity. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, pages II725–II728, 2007. doi: 10.1109/ICASSP.2007.366338. URL `https://scholars.cityu.edu.hk/en/publications/publication(b8dfb698-4523-4203-bf8a-3fb7ec91c584).html`.

Krystine Irene Batcho. Nostalgia and the Emotional Tone and Content of Song Lyrics. *The American Journal of Psychology*, 120(3):361–381, October 2007. ISSN 0002-9556, 1939-8298. doi: 10.2307/

20445410. URL https://scholarlypublishingcollective.org/ajp/article/120/3/361/258559/Nostalgia-and-the-Emotional-Tone-and-Content-of.

Markus Berg. *DO YOU HAVE THE TIME TO LISTEN TO ME WHINE? First person pronoun use in English pop lyrics*. PhD thesis, Göteborgs Universitet, Göteborg, Sweden, February 2016. URL http://hdl.handle.net/2077/41793. Accepted: 2016-02-05T09:22:02Z.

Brook Bretthauer, Toni Schindler Zimmerman, and James H. Banning. A Feminist Analysis of Popular Music: Power Over, Objectification of, and Violence Against Women. *Journal of Feminist Family Therapy*, 18(4):29–51, February 2007. ISSN 0895-2833, 1540-4099. doi: 10.1300/J086v18n04_02. URL http://www.tandfonline.com/doi/abs/10.1300/J086v18n04_02.

Òscar Celma. Foafing the Music: Bridging the Semantic Gap in Music Recommendation. In Isabel Cruz, Stefan Decker, Dean Allemang, Chris Preist, Daniel Schwabe, Peter Mika, Mike Uschold, and Lora M. Aroyo, editors, *The Semantic Web - ISWC 2006*, pages 927–934, Berlin, Heidelberg, 2006. Springer. ISBN 978-3-540-49055-5. doi: 10.1007/11926078_67.

Heeryon Cho, Songkuk Kim, Jongseo Lee, and Jong-Seok Lee. Data-driven integration of multiple sentiment dictionaries for lexicon-based sentiment classification of product reviews. *Knowledge-Based Systems*, 71:61–71, November 2014. ISSN 0950-7051. doi: 10.1016/j.knosys.2014.06.001. URL https://www.sciencedirect.com/science/article/pii/S0950705114002196.

Peter Christenson, Donald F. Roberts, and Nicholas Bjork. Booze, Drugs, and Pop Music: Trends in Substance Portrayals in the Billboard Top 100—1968–2008. *Substance Use & Misuse*, 47(2):121–129, January 2012. ISSN 1082-6084, 1532-2491. doi: 10.3109/10826084.2012.637433. URL http://www.tandfonline.com/doi/full/10.3109/10826084.2012.637433.

Peter G. Christenson, Siliva de Haan-Rietdijk, Donald F. Roberts, and Tom F. M. ter Bogt. What has America been singing about? Trends in themes in the U.S. top-40 songs: 1960–2010, 2019. URL https://journals.sagepub.com/doi/full/10.1177/0305735617748205.

Virginia W. Cooper. Women in popular music: A quantitative analysis of feminine images over time. *Sex Roles*, 13(9-10):499–506, November 1985. ISSN 0360-0025, 1573-2762. doi: 10.1007/BF00287756. URL http://link.springer.com/10.1007/BF00287756.

C. Nathan DeWall, Richard S. Pond, W. Keith Campbell, and Jean M. Twenge. Tuning in to psychological change: Linguistic markers of psychological traits and emotions over time in popular U.S. song lyrics. *Psychology of*

Aesthetics, Creativity, and the Arts, 5(3):200–207, August 2011. ISSN 1931-390X, 1931-3896. doi: 10.1037/a0023195. URL https://doi.apa.org/doi/10.1037/a0023195.

Jason T. Eastman and Terry F. Pettijohn II. Gone country: An investigation of Billboard country songs of the year across social and economic conditions in the United States. *Psychology of Popular Media Culture*, 4(2):155–171, 2015. ISSN 2160-4142. doi: 10.1037/ppm0000019. Place: US Publisher: Educational Publishing Foundation.

Alya Rizqi Fatikha and Qanitah Masykuroh. FIGURATIVE LANGUAGE IN SELECTED SONGS OF "RED" BY TAYLOR SWIFT. *Language Literacy: Journal of Linguistics, Literature, and Language Teaching*, 6(2):299–308, December 2022. ISSN 2580-9962, 2580-8672. doi: 10.30743/ll.v6i2.5711.

M. A. Flynn and C. N. Craig. Objectification in Popular Music Lyrics: An Examination of Gender and Genre Differences | Sex Roles. *Sex Roles*, 75:164–176, 2016. doi: https://doi.org/10.1007/s11199-016-0592-3. URL https://link.springer.com/article/10.1007/s11199-016-0592-3.

I. I. Gukasyan. *The Poetics of Music: Enhancing Content-Based Music Recommendation with Lyrical Features*. PhD thesis, Erasmus University, Rotterdam, Netherlands, 2023. URL https://thesis.eur.nl/pub/72656/thesis_491911.pdf.

María Hernández Rodríguez. Always Rooting For The Antihero: A Dive Into The Evolution Of Taylor Swift's Poetic Persona. *REDEN. Revista Española de Estudios Norteamericanos*, 5(2), May 2024. ISSN 2695-4168. doi: 10.37536/reden.2024.5.2360.

Fabian Holt. *Genre in Popular Music*. University of Chicago Press, Chicago, IL, October 2007. ISBN 978-0-226-35039-4. URL https://press.uchicago.edu/ucp/books/book/chicago/G/bo5414950.html.

C. J. Hutto. cjhutto/vaderSentiment, December 2024. URL https://github.com/cjhutto/vaderSentiment. original-date: 2014-11-17T16:31:45Z.

Justine Kao and Dan Jurafsky. A Computational Analysis of Style, Affect, and Imagery in Contemporary Poetry. In David Elson, Anna Kazantseva, Rada Mihalcea, and Stan Szpakowicz, editors, *Proceedings of the NAACL-HLT 2012 Workshop on Computational Linguistics for Literature*, pages 8–17, Montréal, Canada, June 2012. Association for Computational Linguistics. URL https://aclanthology.org/W12-2502.

Evgeny Kim and Roman Klinger. Who Feels What and Why? Annotation of a Literature Corpus with Semantic Roles of Emotions. In Emily M. Bender, Leon Derczynski, and Pierre Isabelle, editors, *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1345–1359, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics. URL https://aclanthology.org/C18-1114.

Shayna Kothari. shaynak/taylor-swift-lyrics, November 2024. URL `https://github.com/shaynak/taylor-swift-lyrics`. original-date: 2020-12-16T21:58:32Z.

Jennifer C. Lena and Richard A. Peterson. Classification as Culture: Types and Trajectories of Music Genres. *American Sociological Review*, 73(5):697–718, 2008. ISSN 0003-1224. URL `https://www.jstor.org/stable/25472554`. Publisher: [American Sociological Association, Sage Publications, Inc.].

Steven Loria. TextBlob: Simplified Text Processing — TextBlob 0.18.0.post0 documentation, 2013. URL `https://textblob.readthedocs.io/en/dev/index.html`.

Yiqing Ma, David John Baker, Katherine Vukovics, Connor J. Davis, and Emily M. Elliott. Lyrics and melodies: Do both affect emotions equally? A replication and extension of Ali and Peynircioğlu (2006), 2024. URL `https://journals.sagepub.com/doi/epub/10.1177/10298649221149109`.

Yasaman Madanikia and Kim Bartholomew. Themes of Lust and Love in Popular Music Lyrics From 1971 to 2011. *Sage Open*, 4(3):2158244014547179, July 2014. ISSN 2158-2440, 2158-2440. doi: 10.1177/2158244014547179. URL `https://journals.sagepub.com/doi/10.1177/2158244014547179`.

A. Al Marouf, R. Hossain, M. R. Kabir Rase Sarker, B. Pandey, and S. M. TAnvir Siddiquee. Recognizing Language and Emotional Tone from Music Lyrics using IBM Watson Tone Analyzer. pages 1–6, Coimbatoire, India, 2019. IEEE. doi: 10.1109/ICECCT.2019.8869008. URL `https://ieeexplore.ieee.org/document/8869008`.

Tiara Septinia Marpaung, Rotua Elfrida Pangaribuan, and Donna Ria Pasaribu. Harmonizing Linguistics with Melody: An Analysis of Lexemic Idioms in Taylor Swift's 'Evermore' and Their Pedagogical Implications for EFL. *ELT Worldwide: Journal of English Language Teaching*, 10(2):393, October 2023. ISSN 2503-2291, 2303-3037. doi: 10.26858/eltww.v10i2.53617.

Walaa Medhat, Ahmed Hassan, and Hoda Korashy. Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, 5(4):1093–1113, December 2014. ISSN 2090-4479. doi: 10.1016/j.asej.2014.04.011. URL `https://www.sciencedirect.com/science/article/pii/S2090447914000550`.

Tanya Nair. Natural Language Processing used in sentiment analysis of poetry: a study of six common techniques. *Journal of Student Research*, 12(2), May 2023. ISSN 2167-1907. doi: 10.47611/jsrhs.v12i2.4418. URL `https://www.jsr.org/hs/index.php/path/article/view/4418`.

Kathleen Napier and Lior Shamir. Quantitative Sentiment Analysis of Lyrics in Popular Music. *Journal of Popular Music Studies*, 30(4):161–176, December 2018. ISSN 1533-1598. doi: 10.1525/jpms.2018.300411.

Finn Nielson. AFINN Sentiment Analysis, 2018. URL https://darenr.github.io/afinn/. Section: Articles.

José Luis Ortega. Is Taylor Swift leading a new Pop revolution? A cross- generation analysis of Pop/Rock cover songs [version 2; peer review: 1 approved with reservations, 1 not approved]. *F1000 Research*, 2024. doi: http://dx.doi.org/10.12688/f1000research.125676.1. URL http://hdl.handle.net/10261/362509.

Orsolya Papp-Zipernovszky, Anne Mangen, Arthur Jacobs, and Jana Lüdtke. Shakespeare sonnet reading: An empirical study of emotional responses. *Language and Literature*, 31(3):296–324, August 2022. ISSN 0963-9470. doi: 10.1177/09639470211054647. URL https://doi.org/10.1177/09639470211054647. Publisher: SAGE Publications Ltd.

Keith J. Petrie, James W. Pennebaker, and Borge Sivertsen. Things we said today: A linguistic analysis of the Beatles. *Psychology of Aesthetics, Creativity, and the Arts*, 2(4):197–202, November 2008. ISSN 1931-390X, 1931-3896. doi: 10.1037/a0013117. URL https://doi.apa.org/doi/10.1037/a0013117.

Terry F. Pettijohn and Donald F. Sacco. Tough times, meaningful music, mature performers: popular Billboard songs and performer preferences across social and economic conditions in the USA. *Psychology of Music*, 37(2):155–179, April 2009. ISSN 0305-7356. doi: 10.1177/0305735608094512. URL https://doi.org/10.1177/0305735608094512. Publisher: SAGE Publications Ltd.

Robert Plutchik. The Nature of Emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *Sigma Xi, The Scientific Research Society*, 89(4):344–350, 2001. URL http://www.jstor.org/stable/27857503.

Sebastian Raschka. MusicMood: Predicting the mood of music from song lyrics using machine learning, November 2016. arXiv:1611.00138 [cs].

Simone Rebora. Sentiment Analysis in Literary Studies. A Critical Survey. In *Digital Humanities Quarterly: Sentiment Analysis in Literary Studies. A Critical Survey*, 2023. URL https://api.semanticscholar.org/CorpusID:261066079.

Donald F. Sacco and Terry F. Pettijohn II. The Language of Lyrics: An Analysis of Popular Billboard Songs Across Conditions of Social and Economic Threat - Terry F. Pettijohn, Donald F. Sacco, 2009, 2009. URL https://journals.sagepub.com/doi/10.1177/0261927X09335259.

Markus Schedl. Frontiers | Deep Learning in Music Recommendation Systems. *Frontiers in Applied Mathematics and Statistics*, 5, 2019. doi: https://doi.org/10.3389/fams.2019.00044. URL https://www.frontiersin.org/journals/applied-mathematics-and-statistics/articles/10.3389/fams.2019.00044/full.

E. Glenn Schellenberg and Christian von Scheve. Emotional cues in American popular music: Five decades of the Top 40. *Psychology of Aesthetics, Creativity, and the Arts*, 6(3):196–203, 2012. ISSN 1931-390X. doi: 10.1037/a0028024. Place: US Publisher: Educational Publishing Foundation.

John Street. Rock, pop and politics. In Simon Frith, Will Straw, and John Street, editors, *The Cambridge Companion to Pop and Rock*, pages 243–255. Cambridge University Press, 1 edition, August 2001. ISBN 978-0-521-55660-6 978-0-521-55369-8 978-1-139-00224-0. doi: 10.1017/CCOL9780521553698.015. URL https://www.cambridge.org/core/product/identifier/CBO9781139002240A022/type/book_part.

Philip Tagg. Musicology and the semiotics of popular music. *Semiotica*, 66-1/3:279–298, 1987.

Robert W. Van Sickel. A World without Citizenship: On (the Absence of) Politics and Ideology in Country Music Lyrics, 1960–2000. *Popular Music and Society*, 28(3):313–331, July 2005. ISSN 0300-7766. doi: 10.1080/03007760500105164. Publisher: Routledge _eprint: https://doi.org/10.1080/03007760500105164.

Mila Volpe. Music's Powers: Embodiment & Flourishing with Taylor Swift. *Contemporary Music Review*, 40(1):60–77, January 2021. ISSN 0749-4467, 1477-2256. doi: 10.1080/07494467.2021.1945228. URL https://www.tandfonline.com/doi/full/10.1080/07494467.2021.1945228.

Alan West and Colin Martindale. Creative trends in the content of beatles lyrics. *Popular Music & Society*, December 1996. doi: 10.1080/03007769608591647. URL https://www.tandfonline.com/doi/abs/10.1080/03007769608591647. Publisher: Taylor & Francis Group.

Cynthia Whissell. Traditional and emotional stylometric analysis of the songs of Beatles Paul McCartney and John Lennon. *Computers and the Humanities*, 30(3):257–265, 1996. ISSN 0010-4817, 1572-8412. doi: 10.1007/BF00055109. URL http://link.springer.com/10.1007/BF00055109.

Cynthia Whissell. Emotional Fluctuations in Bob Dylan's Lyrics Measured by the Dictionary of Affect Accompany Events and Phases in his Life, 2008. URL https://journals.sagepub.com/doi/epdf/10.2466/pr0.102.2.469-483.

Chenyang Wu, Scott Le Vine, Elizabeth Bengel, Jason Czerwinski, and John Polak. Sentiment analysis of popular-music references to automobiles, 1950s to 2010s. *Transportation*, 49(2):641–678, April 2022. ISSN 0049-4488, 1572-9435. doi: 10.1007/s11116-021-10189-1.

Chia-Hung Yeh, Wen-Yu Tseng, Chia-Yen Chen, Yu-Dun Lin, Yi-Ren Tsai, Hsuan-I Bi, Yu-Ching Lin, and Ho-Yi Lin. Popular music representation: chorus detection & emotion recognition. *Multimedia Tools and Applications*, 73(3):2103–2128, December 2014. ISSN 1380-7501, 1573-7721. doi: 10.1007/s11042-013-1687-2. URL `http://link.springer.com/10.1007/s11042-013-1687-2`.

A Data Cleaning and Pre-Processing Steps

The data cleaning was an itemized process that I undertook in a carefully designed order to ensure the best possible starting point to achieve my results. All lyric data was adapted from Kothari (2024).

1. In the lyric data I started with, choruses and verses were marked as such (ie. [CHORUS]; [VERSE]) I did not want any of these characters included, and so I removed them from the data before continuing.
2. I wanted "not" to be counted as its own token from contractions in order to properly reflect the word's negative sentiment in the text. Therefore, I expanded all contractions (ie. "can't" → "can" "not")
3. In order to count all varying instances of the same word as one, I transitioned all of the lyrics to lowercase (ie. "Yes", "YES", "yes" –> "yes")
4. I removed all nonsense and non-contentful lyrics (ie. “na-na”, “mm-mm”, “ra-di-di-di-di-di-di-di-di-di-da-da-da”)
5. I removed stop words (words important for grammar that are otherwise non-contentful), except for a set list that I felt were still uniquely able to contribute to my research goals, such as negations and pronouns
6. I removed all punctuation markers
7. I lemmatized the lyrics ("walks", "walking", "walked" → "walk") for the sake of preserving counts