rijksuniversiteit groningen

faculteit der letteren

Digital Humanities: Tools and Methods — GROUP PROJECT FINAL REPORT

Group 2

Liesa Albers (S3830381)

Natalie Voo Xin Ru (S4228693)

Pragati Prasad (S4064283)

Digital Humanities: Tools and Methods

LHU011M05

Faculty of Arts

University of Groningen

January 2025

Table of Contents

1. Introduction

Music has always been an integral part of human culture, serving as a medium for self-expression, storytelling, and entertainment. As a powerful art form, it offers valuable insights into the social and cultural contexts of its creation (Hao, 2019). In contemporary music, nostalgia has emerged as a popular tool, with many hit songs incorporating elements from the 1970s and 1980s, leveraging nostalgia as a form of escapism (Davies et al., 2022). Stylometry, defined by Eder et al. (2016) as "the relationship between writing style in texts and meta-data about those texts (such as date, genre, gender, authorship)," provides a robust framework for analysing these trends. By employing statistical methods, stylometry can identify differences in literary styles, genres, and time periods, making it a powerful tool for tracking the evolution of pop lyrics over time. This report presents a stylometric analysis of popular music lyrics from the 2000s and 2010s, aiming to identify stylistic similarities and differences between these two decades. By analysing stylometric distance between music lyrics, this study aims to answer the following research question:

"How can the stylometric relationship of music lyrics from the 2000s and 2010s be defined, and what are the key stylometric differences or similarities in lyrics from the 2000's and 2010's?"

To address this question, this study employs a combination of computational text analysis and statistical methods. The analysis focuses on features suitable for the relatively short length of song lyrics, such as Most Frequent Words (MFW), Most Frequent Bi-grams, and function word frequencies. The analysis utilises dendrogram clustering, bootstrap consensus trees, and Craig's Zeta method to provide a comprehensive exploration of lyrical styles across the two decades. By examining these stylometric relationships, this report aims to contribute to the understanding of how popular music has evolved in recent decades, reflecting broader cultural and social changes. This analysis not only provides insights into the craft of songwriting but also offers a unique perspective on the shifting landscape of popular culture in the 21st century.

2. Data

This study undertakes a stylometric analysis of music lyrics from the 2000s and 2010s, employing a combination of computational text analysis and statistical methods. The aim is to identify and compare stylistic differences in song lyrics across these two decades, addressing the challenges posed by the relatively short length of song lyrics.

The corpus for this study was constructed using lyrics from songs listed on the Billboard Year-End Hot 100 charts for the relevant years. At the end of each year, Billboard Magazine publishes a chart of the 100 most popular singles throughout the year based on a combination of weekly physical and digital sales, radio play, and streaming data tracked by the company Nielsen SoundScan. Therefore, the charts present a reliable indicator of general music consumption trends. Specifically, the top 10 songs from each year of the two decades were selected for analysis, resulting in 20 text files (10 per decade). The titles and artists of these songs were extracted using Python's `BeautifulSoup` library to scrape data from Billboard's Wikipedia pages, as the official Billboard Magazine website employs a paywall. The song titles and artists were then formatted for compatibility with their AZLyrics webpage, an online lyrics database, where the lyrics for each song were retrieved using the same web scraping technique. Each year's lyrics were combined into a single text file, a decision made to mitigate the limitations posed by the short length of individual songs. This approach resulted in a corpus with an average text length of 4,606 words per file, allowing for more robust stylometric analysis. The corpus is accompanied by a metadata file in CSV format, which includes details about each song, such as the year, Billboard rank, title, and artist. The metadata file can be found in the GitHub repository that accompanies this report: https://github.com/LiesaAlbers/TM-project.

3. Methodology

Given the characteristics of the corpus, specific stylometric features were chosen to accommodate the limitations of shorter texts. Features that rely on longer texts, such as syntactic tree structures, semantic analysis, and vocabulary richness measures, were excluded. These features require a substantial amount of text to produce meaningful results, and their application to short texts like song lyrics can lead to unreliable findings (Lagutina et al., 2013; López-Escobedo et al., 2013). Instead, the analysis focused on the Most Frequent Words (MFW) and Most Frequent Bi-grams to provide reliable insights into the stylistic patterns of music lyrics from the 2000's and the 2010's.

The analysis was conducted using the `Stylo` package in R. For each of the features, varying ranges were tested to ensure the stability of clusters. Additionally, to reduce noise and improve interpretability, culling was applied to all features. Culling ranges of 0–20% and 10–30% with 5% increments were employed, ensuring that only words appearing in at least 20%

of the texts were included, while still preserving stylistic patterns. This range also accounted for the repetitive and filler structures characteristic of song lyrics, such as refrains and choruses. The stylometric analysis included two main methods: dendrogram clustering and bootstrap consensus trees. Bootstrap consensus trees were used in addition to the dendrograms to assess the stability of clusters across different parameter settings. These methods allowed for a robust comparison of lyrical styles between the two decades. Additionally, Craig's Zeta was employed to identify words that were distinctly overused or underused in the 2000s compared to the 2010s. This method relied on two predefined subcorpora: one for the 2000s making up the primary set (99 songs) and one for the 2010s making up the secondary set (98 songs). Craig's Zeta computed scores for each word, highlighting lexical items that were characteristic of each decade.

The combination of dendrograms, bootstrap consensus trees, and Craig's Zeta provide a comprehensive approach to identify and compare stylistic differences in song lyrics across the two decades. By combining robust corpus construction methods, carefully selected stylometric features, and rigorous analytical techniques, this study ensures a reliable and nuanced exploration of the evolution of lyrical styles in popular music over the 2000s and 2010s.

4. Analysis

Using the discussed methodology, our analysis was conducted with the aim to discover and define the stylometric relationship between popular song lyrics from the 2000s and the 2010s by comparing their stylometric features.

In Figure 1, the dendrogram based on the 300 Most Frequent Words within the corpus reveals that stylometric features often rise and fall in popularity, so much so that popular music from one year can share quite a few similarities with those from even a decade later. Take for example, the (2002, 2014) branch and the (2002, 2011) branch in the dendrogram. No year has popular music that stood out within the two decades. It also explains why sometimes, even if you do not take the actual "musical" qualities of a song, it can reflect a certain year's "vibe", eg: "This song is so 2016.". Sometimes it is the aural qualities but oftentimes it can be the style of the lyrics – how certain things are phrased, and what kind of "function" words were common at a certain time.

Figure 1. Dendrogram 300 MFW, 0-20% Culling, 5% increments

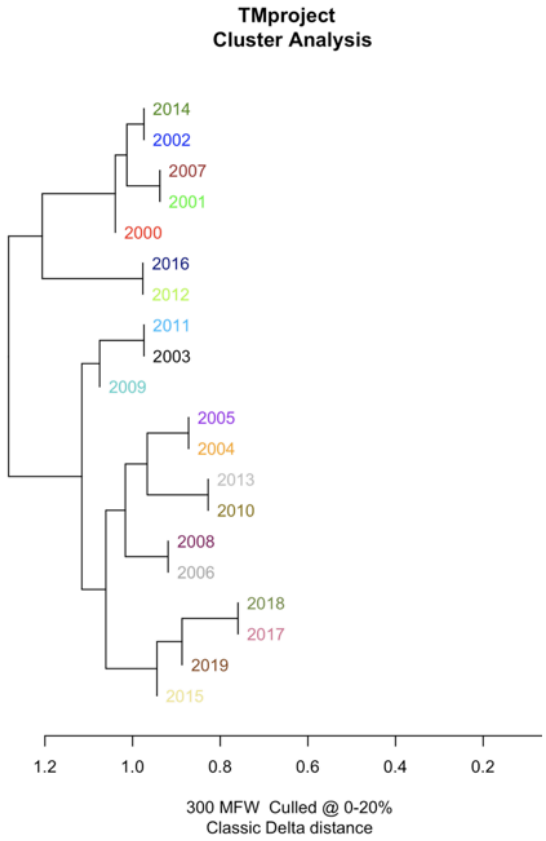
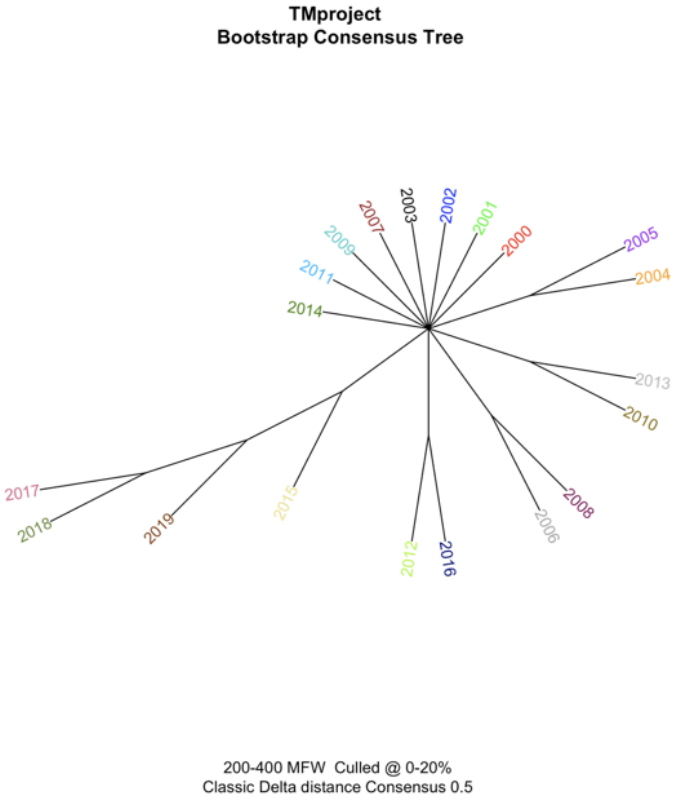


Figure 2. Bootstrap Consensus Tree 200-400 MFW, 0-20% Culling, 5% increments



The bootstrap consensus tree in Figure 2 above similarly proves that sometimes there are some lyrical trends that remain popular for an extended period of time before morphing slightly to fit current trends, The more obvious connections are songs from 2004 and 2005 splitting at the same distance from a single branch, in the same way as (2020, 2013), (2006, 2008), (2010, 2013), and (2012, 2016). The consensus tree also makes it clear with its outlier branch being splintered out that music from 2015, 2017, 2018, and 2019 share quite a few similarities but are stylometrically distinct from the rest of the music released between 2000 and 2020. It also reveals that what makes music “trendy” or “popular” has remained relatively stable over the course of the two decades.

To further differentiate music released in different years, we looked at bi-grams—where two words occur together within the corpus. The following figures thus show a dendrogram of the most frequent bi-grams (Figure 3) and a bootstrap consensus tree of the corpus' bigrams (Figure 4).

Figure 3. Dendrogram 300 Most Frequent 2-grams, 0-20% Culling, 5% increments

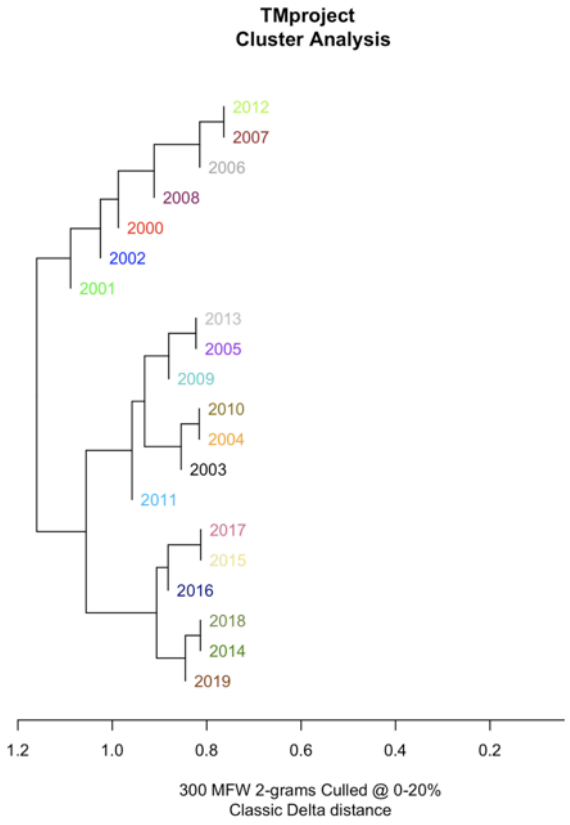
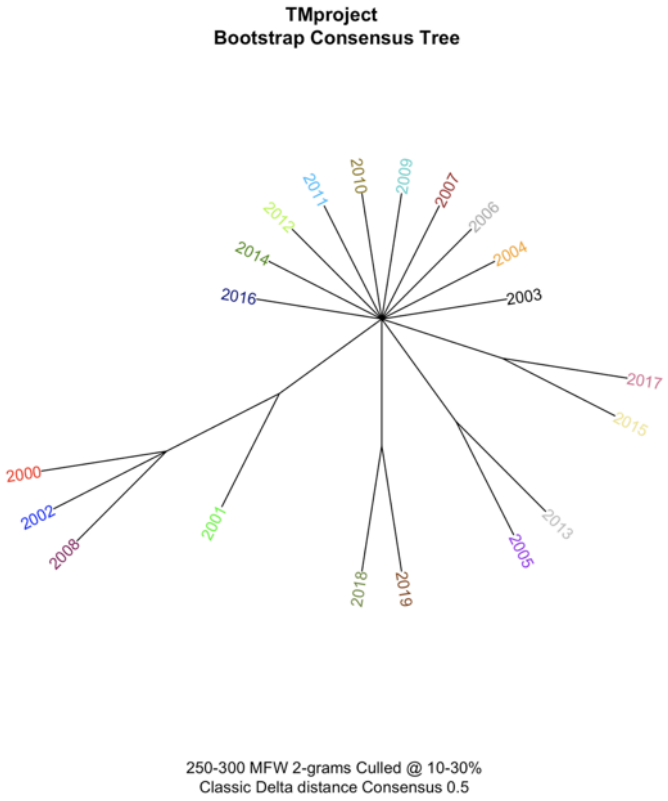


Figure 4. Bootstrap Consensus Tree 250-300 Most Frequent 2-grams, 10-30% Culling, 5% increments



In this dendrogram, there are three main stylometric clusters – (2000, 2001, 2002, 2006, 2007, 2008, 2012), (2003, 2004, 2005, 2009, 2010, 2011, 2013), and (2014-2019). Thus, music from the 2000s and the first half of the 2010s are split between the first two clusters while the third contains groups the latter half of the 2010s as part of the same stylometric "family" of sorts. This might indicate either a decrease in creativity within song writing or the establishment and promotion of what commercially successful, radio-friendly "sound" over music from other genres and subgenres, though neither supposition can be proven with only the tools on hand.

However, the first and second cluster also show “jumps” of what is popular between various years giving credence to the idea of trends phasing in and out of popularity as well as the introduction of using ‘samples’ from older songs as well as the prevalence of covers by other artists, remixes, and in the digital age – the concept of a song going “viral” years after being released, inspiring other artists.

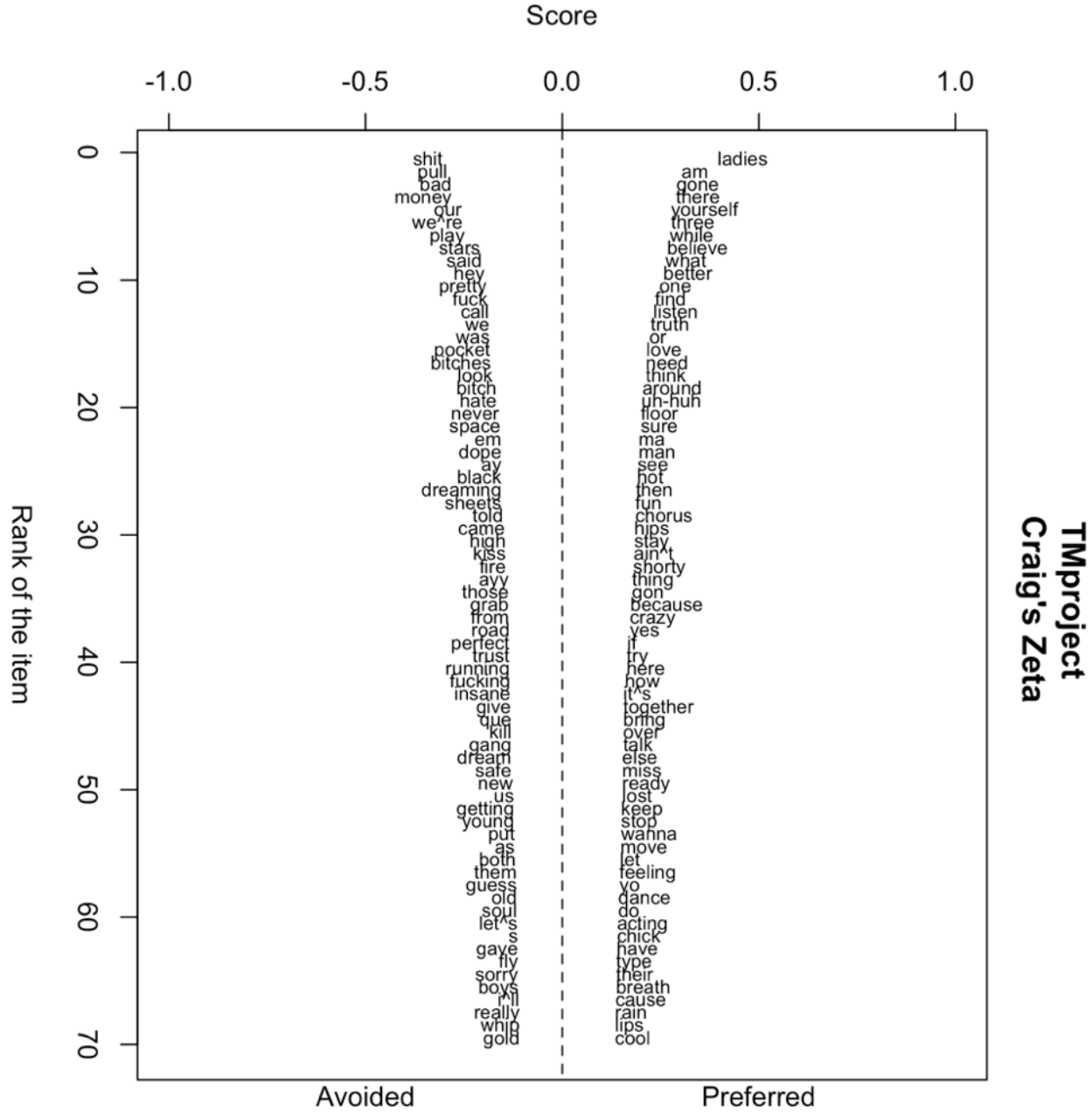
To account for the sheer repetition within the corpus with bi-grams like “oh yeah”, “yeah yeah”, or “I love” in the bootstrap consensus tree, the culling rate was set to 10%–30% rather than as previously done 0–20%. This allows Stylo to better filter out the frequent and less informative bi-grams. Despite this change in the visualisation between Figure 2 and Figure 4 due to the setting adjustment, the main difference is that the tree in the latter is more compact, less spread out, therefore allowing us to trace how the music really changed, and how much it stayed the same as the years went by.

While the cluster patterns within the above dendrograms identify and group years by the stylometric similarities of the most popular songs of each year, it does not indicate a shared songwriter or lyricist the way it would be if one were analysing a longer text. This is because lyrics are essentially a different work of art and therefore require different affordances due to their shorter form. In this case especially, it can only indicate similarities to music popular in other years and visualising clear outliers when it comes to musical or lyrical style. As the corpus consists of only a relatively small number of text files with shorter content than the average text, the dendrograms are good for indicating a starting point for where to begin close-reading and deeper analysis of the text or lyrics. As Stylo only provides visualisations without any specifics, one of the clearer ways to reveal lexical differences between songs popular in the 2000s versus the 2010s is by applying Craig’s Zeta Analysis (Figure 5). This allows us to look at the usage of the words themselves within popular music from each decade.

As the corpus was split into primary and secondary sets based on their decade, the Figure below reveals that the words on the right side of the graph are the ones ‘preferred’ by music from the primary set i.e. the Billboard top 10 from the 2000s. As such, the words on the left side of the graph are those words that are significant in songs from the secondary set i.e. Billboard top 10 from the 2010s. At first glance, the most obvious thing to note is the prevalence of the word “ladies” in 2000s music. On the other hand there is a stark difference when one compares the top 20 words from either decade’s 10 most popular songs.

In the 2000s corpus, the top 25 most used words in order of frequency were: ladies, am, gone, there, yourself, three, while, believe, what, better, one, find, listen, truth, or, love, need, think, around, uh-huh, floor, sure, ma, man, see. Whereas in the 2010s, the top 25 most frequent words were: shit, pull, bad, money, our, we^re, play, stars, said, hey, pretty, fuck, call, we, was, pocket, bitches, look, bitch, hate, never, space, em, dope, ay.

Figure 7. Craig's Zeta



This visualisation therefore indicates that the 2000s songs were, on average, more light hearted, referencing positive concepts like love, truth while the presence of numbers (one, three), “ladies”, and “floor” indicate the popularity of dance music in this decade. In contrast, the music from the 2010s include a startlingly larger proportion of swear words and insults and no concepts that could indicate a positive emotion other than “pretty” which from the context most likely refers to the amplifier “very” and not the pleasantness of one’s appearance. This word ambiguity is one of the issues of using Craig’s Zeta to make definitive statements regarding word usage within a text or, in this case, songs. Another criticism is that, much like the cluster analysis conducted earlier, Craig’s Zeta is most useful to indicate interesting patterns in subject matter, tone, and word usage for study rather than to use as evidence for a declarative statement regarding stylistic elements of a text.

5. Conclusion

In the context of this project, the dendrograms and bootstrap consensus tree visualisations revealed that the stylometric elements of popular music shift gradually as time passes, with the occasional outlier and the rise and fall of trends, though of course certain stylistic elements can come back into vogue due to unpredictable external factors (eg: album re-release, song/genre going viral on social media, etc.). Meanwhile the Craig’s Zeta score indicates a drastic change in the tone and use of language in popular music across the two decades.

As mentioned in the Data section, the corpus for this project was relatively small with only 20 text files, each containing the lyrics of a single year’s top 10 from the Billboard’s most popular songs. If we look at it through the lens of literature analysis, it can be equated to a corpus of short poems which are not often studied in this manner due to digital humanities methods being better equipped to analyse larger amounts of data. Though the tools employed revealed some trends to separate and compare the stylometric elements of popular music from the 2000s and the 2010s, the visualisations would have been far more robust if we had applied them to a much larger dataset, such as all 100 songs from each year. A larger sample size allows for a more comprehensive analysis of stylistic patterns and reduces the influence of outliers or idiosyncratic features. The overall reliability of stylometric analysis generally improves with longer text samples.

References

Davies, C., Page, B., Driesener, C., & Winzar, H. (2022). The power of nostalgia: Age and preference for popular music. Marketing Letters, 33(4), 681–692. https://doi.org/10.1007/s11002-022-09626-7

Eder, M., Rybicki, J., & Kestemont, M. (2016). Stylometry with R: A package for computational text analysis. The R Journal, 8(1), 107–121. https://journal.r-project.org/archive/2016/RJ-2016-007/index.html **[Stylo version 0.7.5**]

Hao, K. (2019). The brief history of music storytelling. Medium. Retrieved January 11, 2025, from https://medium.com/@kenhao/the-brief-history-of-music-storytelling-42ca9684bacf

Lagutina, K., Lagutina, N., Boychuk, E., Vorontsova, I., Shliakhtina, E., Belyaeva, O., Paramonov, I., & Demidov, P.G. (2013). A Survey on Stylometric Text Features. 2019 25th Conference of Open Innovations Association (FRUCT), 184–195, https://doi.org/10.23919/FRUCT48121.2019.8981504

López-Escobedo, F., Méndez-Cruz, C., Sierra, G. & Solórzano-Soto, J. (2013). Analysis of Stylometric Variables in Long and Short Texts. Procedia - Social and Behavioral Sciences, 95, 604–611. https://doi.org/10.1016/j.sbspro.2013.10.688

Massung, S. [Smassung]. (2012, September 7). function-words.txt [Data set]. Meta-Toolkit Github Repository. Retrieved January 2025, from https://github.com/meta-toolkit/meta/blob/master/data/function-words.txt

R Core Team (2024). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. https://www.R-project.org/ [R]

Appendix

Name	Contributions
Natalie	- Wrote draft of web scraping code, - Wrote draft of CSV metadata code, - Wrote Introduction, - Created initial format of report.
Liesa	- Finalised all code, - Extracted and prepared all data, - Performed analysis using Stylo, - Wrote draft for analysis, - Wrote Data + Methodology, - Created github repository.
Pragati	- Finalised analysis, - Wrote conclusion.

Appendix A: Division of labour