Helen Wang

Project Report: Examining SHSAT Sentiments

Introduction

Every year, students across New York City (NYC) take the Specialized High School Admission Test (SHSAT) to gain admission to one of NYC's prestigious Specialized High Schools. Historically considered as pathways to elite colleges and financial success, admission to these Specialized High Schools has only been possible through the SHSAT, which was established in 1971 through the Hecht-Calandra Act (Bradford, 2015; Taylor, 2019).

However, the SHSAT has become increasingly controversial over the years. Proponents have favored it as an objective and meritocratic admissions metric, while opponents have argued that admission should be based on more holistic measures, such as middle school grades and extracurricular performance.

In 2013, the National Association for the Advancement of Colored People's (NAACP) Legal Defense Fund (LDF) filed a complaint, highlighting the racial discrepancies observed in SHSAT admissions (Hirsch, 2020). The complaint argued that Black and Hispanic/Latinx students were disproportionately less likely to be admitted, and so were students who identified as low-income and/or female. In contrast, high-income male students who identified as White or Asian were disproportionately more likely to be accepted.

Partially in response to those allegations, in 2018, Mayor Bill de Blasio proposed eliminating the SHSAT altogether and instead basing admissions on middle school performance and scores on statewide standardized tests (Segers, 2018). He also proposed reserving 20% of seats at each Specialized High School for disadvantaged students through a discovery program. This prompted fierce debate and while the proposal was ultimately shut down, in 2020, the discovery program was launched at Specialized High Schools (Zimmerman, 2023). Since then, the issue of the SHSAT exam has continuously emerged in local NYC policy discussions, especially in the wake of recent mayoral elections.

Objectives

The goal of this project is thus two-fold:

(1) Examine how semantic associations to the SHSAT have changed over time. Specifically, has it become more associated with DEI or more associated with meritocracy as dialogue around the exam shifted over time?

(2) Evaluate whether stances towards the SHSAT have shifted across time. Has public opinion become more negative or positive with the debate?

Data & Methods

To answer these questions, I scraped NYC-based news outlets for articles related to the SHSAT to create a corpus. To obtain the list of news outlets, I relied on domain expertise and also examined the news outlets that appeared consistently in the Google News Search Engine when "SHSAT" was inputted as a search term. My dataset thus consists of articles across ten different news outlets including: the New York Times, the New York Post, the New York Daily News, the New York Amsterdam News, Gotham Gazette, City Journal, Chalkbeat, Brooklyn Daily Eagle, Spectrum NY, Queens Chronicle, and the amNY (see Figure 1 for a breakdown of the number of articles per news outlet post-preprocessing).

I was able to query the New York Times API for articles, and similarly, for the Gotham Gazette, I was able to query Google's Custom Search API. For all other news outlets, I designed separate webscraping pipelines, either static or dynamic, depending on the website format. For the New York Times, I used the following search term: "SHSAT OR Specialized High Schools Admissions Test OR Specialized High Schools Admissions Exam." For all others, I used "SHSAT" as the search query. This was done as other news outlets did not allow the same level of querying as the New York Times API. All data was collected as of November 18th, 2025, and a total of 1016 articles were successfully scraped.

From each website, I extracted the headline, author, date published, and the article abstract/snippet when possible. All sites had abstracts/snippets of the articles with the exception of amNY. I extracted the year of publication from each of the articles and, as part of preprocessing, dropped all articles that were either missing a headline and/or year of publication.

For text preprocessing, I removed numbers, punctuation, and symbols and lowercased all text. I largely tokenized everything as unigrams but manually tokenized some n-grams, such as 'New York City' and 'Specialized High Schools', along with some names such as 'Bill de Blasio.' After preprocessing, roughly 870 articles were retained.

For the temporal analysis, I elected to compare change across two time periods: Pre- and Post-2018. This is largely due to the fact that visual examination of the distribution of year of publication suggests that the bulk of the articles were published after 2018, or after Mayor Bill de Blasio's controversial proposal, suggesting this may be a defining event or period for which there may have been a semantic or stance-related shift. 142 articles were in the Pre-2018 period and roughly 728 were in the Post-2018 period (see Figure 2 for a breakdown of articles per year of publication).

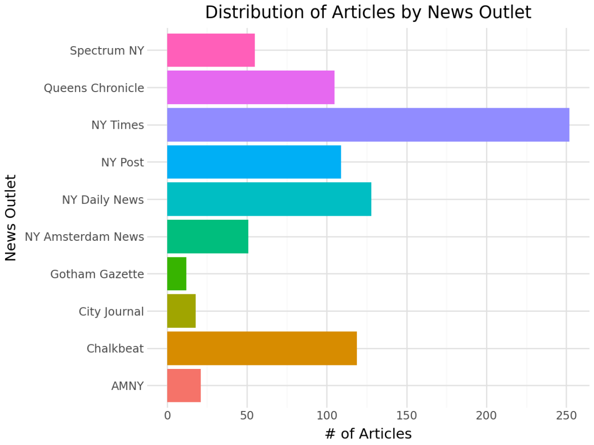


Figure 1. Article Count by News Outlet

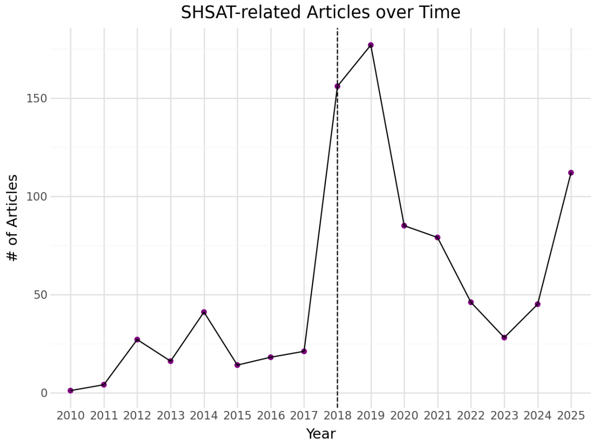


Figure 2. Article Count per Year

Analysis

Diachronic Semantic Analysis

To assess changes in semantic associations with the SHSAT over time, I explored two modelling approaches to diachronic semantic analysis, discussed by Rodman (2020). I explore the Aligned Time Series Model as well as the Chronologically Trained Model. The former requires training word embeddings for each time slice and aligning them post-hoc using orthogonal procrustes matrix alignment. In this case, I used the Pre-2018 as the time slice to “anchor” alignment across periods. While this approach addresses spatial nonconformity, the vocabulary is limited to vocabulary that exists across both time slices.

The Chronologically Trained Model, on the other hand, is similar to traditional transfer learning or pretrained word embeddings. In effect, this model requires that each time slice be initialized using the word vectors from the previous time slice. In this case, for example, the word embeddings for Post-2018 were initialized using the word vectors from Pre-2018. The Pre-2018 word embeddings, however, due to a lack of a “previous” time slice, were initialized on the full corpus. In theory, this approach allows for semantic linkages across time slices without directly addressing the issue of spatial nonconformity.

As the corpus is fairly small by machine learning and text analysis standards, I followed the recommendations of Rodman (2020) and also conducted bootstrapped resampling. As such, for each time slice, I randomly sampled with replacement *n* documents and trained word embeddings using those documents, with *n* being the total number of documents in that era. This was done for 100 iterations before word embeddings were averaged to get stable embeddings for each era. When computing semantic change for pairs of words, cosine similarity was computed for pairs for each iteration and then averaged. In theory, bootstrap resampling in this way (1) prevents single documents from affecting analysis, (2) stabilizes outputs (i.e., cosine similarity scores), and (3) allows for the generation of confidence intervals (Rodman, 2020).

As a sanity check and also to compare modelling approach performance, I computed the cosine similarity of a set of SHSAT, DEI, and Merit-related words across time periods (see Table 1). That is to say, I assessed, for example, how similar the word “SHSAT” was in Pre-2018 to “SHSAT” in Post-2018. To visualize semantic drift, I used t-Distributed Stochastic Neighbor Embedding (t-SNE) to reduce word vectors into a two-dimensional space.

I also assessed how the cosine similarity of word pairs has changed over time. For example, I computed how the cosine similarity of words such as “SHSAT” and “race” has changed between Pre-2018 and Post-2018 to examine if, semantically, the words have become closer or farther apart.

Stance Detection

To complete my second objective of examining how stances towards the SHSAT have changed over time, I employed an off-the-shelf classifier from Hugging Face specialized in stance detection through natural language inference (Horne, Dolinsky, & Huber, 2024). While specialized and fine-tuned specifically for political text, the classifier takes in a “target word” and computes the probability of a given text having a positive, negative, or neutral stance towards the target.

As there were numerous synonyms for my target, I opted to have the model compute these probabilities for a series of target words for each document. My target words in this case were: SHSAT, Test, Exam, Specialized High School, and Specialized High Schools. I then took the maximum probabilities across the target words for negative, positive, and neutral and classified the documents based on which averaged probability was highest. I considered averaging probabilities across the target words for negative, positive, and neutral, but the classifier lacked the sensitivity to pick up stances, based on manual spot checking of some test cases.

To evaluate if stances towards the SHSAT had changed over time, I ran two logit models, using the classification of positive, negative, and neutral documents as predictors of time period. For one model, I included no covariates, and for the other, I controlled for news source, given that some news outlets may have party/political leanings. For both, I computed odds ratios.

To validate my findings from the stance detection model, I recruited two other human coders in addition to myself. I randomly selected 200 documents from my corpus and had all coders manually code the documents as either positive, negative, or neutral towards the SHSAT. I then computed Fleiss Kappa to evaluate inter-coder agreement, averaged our coded responses, and compared the model performance to human coders. To do so, I computed overall accuracy and a confusion matrix.

Results

Diachronic Semantic Analysis

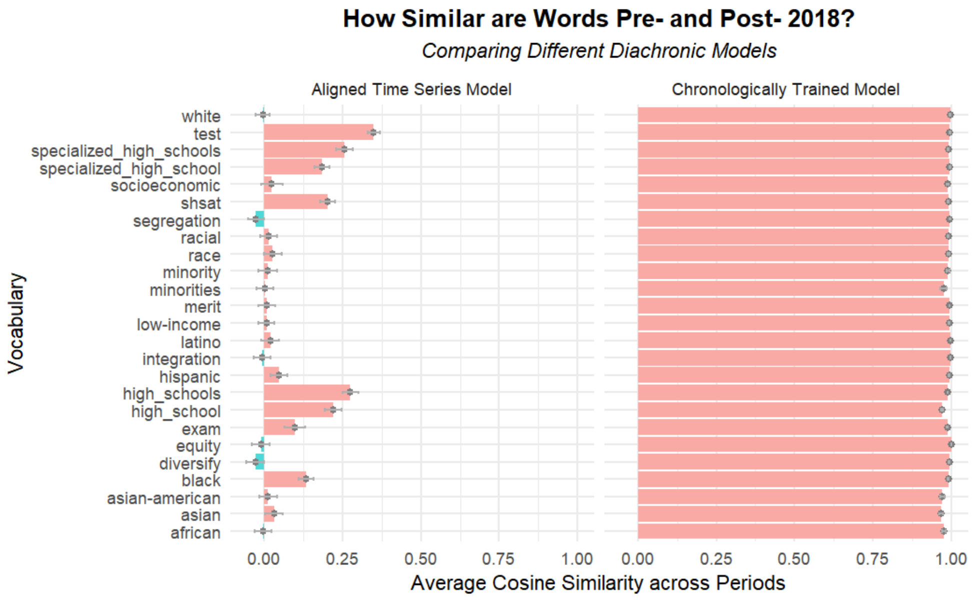


Figure 3. Temporal Cosine Similarity of Individual Words, comparing modelling approaches

In Figure 3, I compared the cosine similarity of individual words across time using two different models: Aligned Time Series and the Chronologically Trained Model. Both produced fairly different results, with words remaining very similar across time in the Chronologically Trained model (M = 0.988, SE = 0.001) but varying significantly in the Aligned Time Series Model (M = 0.075, SE = 0.014). While I expected perhaps some semantic drift in individual words, the result for the

Aligned Time Series Model seemed too drastic and unstable compared to the Chronologically Trained Model. More importantly, past work had previously suggested that the Chronologically Trained Model had superior performance and stability compared to other models (Rodman, 2020). As such, I utilized the Chronologically Trained Model for all future semantic analyses.

In Figure 4, I map semantic drift across time using the same words of interest as before. Examination of the plot suggests that Pre-2018, all words were clustered fairly close to each other and were closely aligned in semantic association. In Post-2018, I observe some semantic drift, with words related to the SHSAT (notably test, exam, shsat, specialized high school, etc.) being farther away from words related to race (notably latino, black, hispanic, white, asian, etc.). This suggests that over time, words related to the SHSAT are less likely to co-occur with words related to race.

Figure 5 reflects a faceted version of Figure 4, and here, three distinct groups of words formed in Post-2018 are observable, while in Pre-2018, all words are clustered together. Notably, in Post 2018, one group of words was directly linked to the admissions test, another group capturing DEI-related words (i.e., diversify, segregation, socioeconomic), and a final group capturing words related to traditionally disadvantaged racial/ethnic groups (i.e., black, latino, hispanic). Notably, DEI-related words are closely related but distinct from both test-related and race-related words, while test-related and race-related words are the farthest apart semantically.

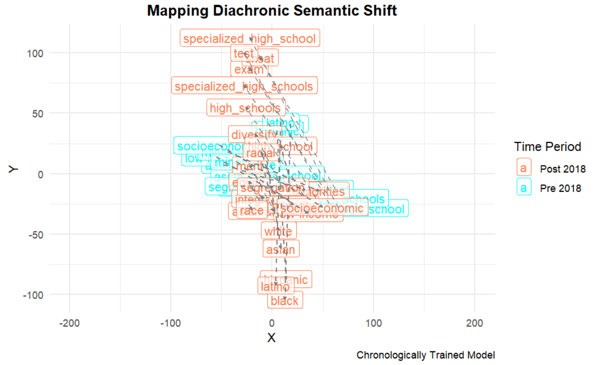


Figure 4. Semantic Shift across Time

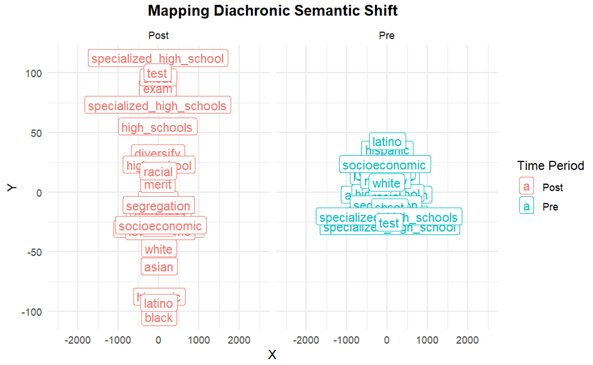


Figure 5. Semantic Shift across Time, facetted

When examining word pairs, as in Figure 6-9, we observe evidence of semantic drift, with word pairs generally becoming less semantic associated over time, with the exception of test-based words. Examining the temporal cosine similarity of word pairs containing words such as SHSAT, Merit, Equity, and Diversify, for example, we generally observe lowered cosine similarity for all pairs in the post-2018 period with some minor exceptions.

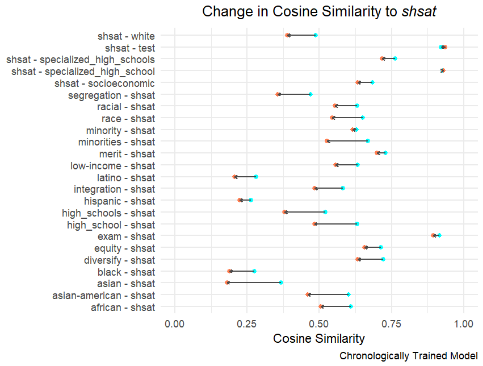


Figure 6. Cosine Similarity for Word Pairs with SHSAT

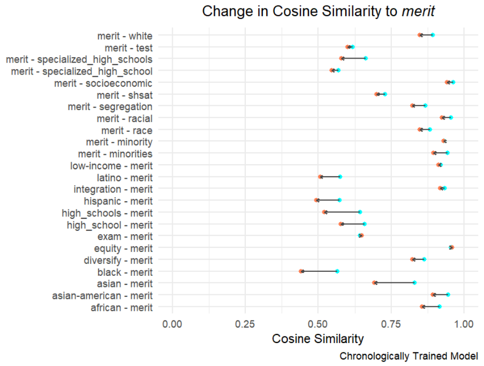


Figure 7. Cosine Similarity for Word Pairs with Merit

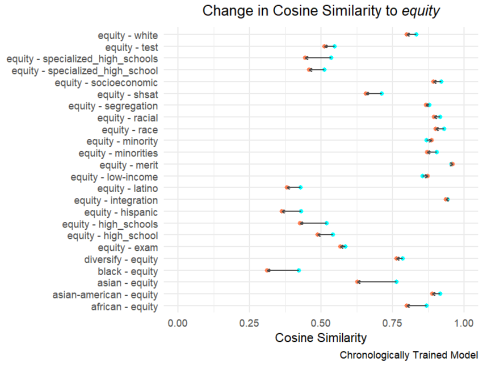


Figure 8. Cosine Similarity for Word Pairs with Equity

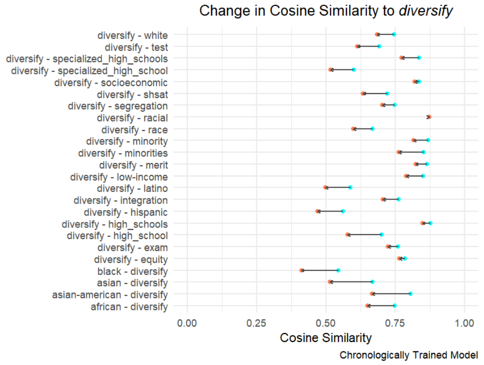
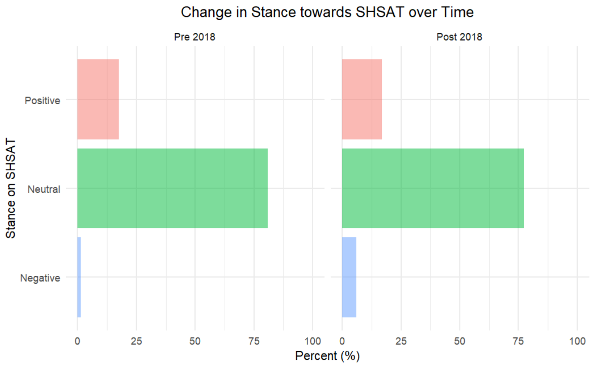


Figure 9. Cosine Similarity for Word Pairs with Diversify

Stance Detection

In terms of stance detection, Figure 10 shows the distribution of positive, negative, and neutral articles across periods. Visually, there appears to be minimal change across time, outside of a marginal increase in negative or anti-SHSAT articles. In Figure 11, I present results from two logistic regression models, predicting time period with the stance classifications as independent variables, with one model controlling for news outlet and the other having no covariates.

When the models do not control for news outlet source, negative articles are linked to significantly higher odds of the date of publication being post 2018 (OR = 4.502, p = 0.039). This effect disappears, however, when models control for news outlet source (OR = 2.320, p = 0.265).



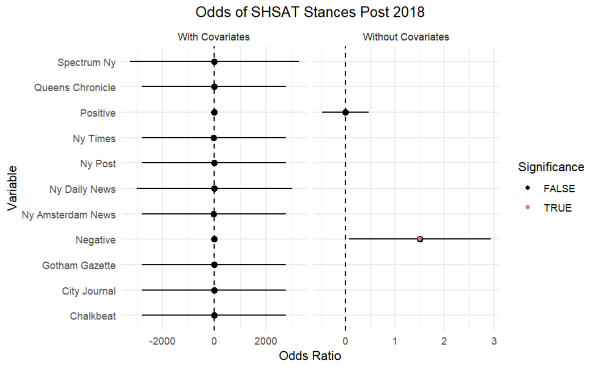
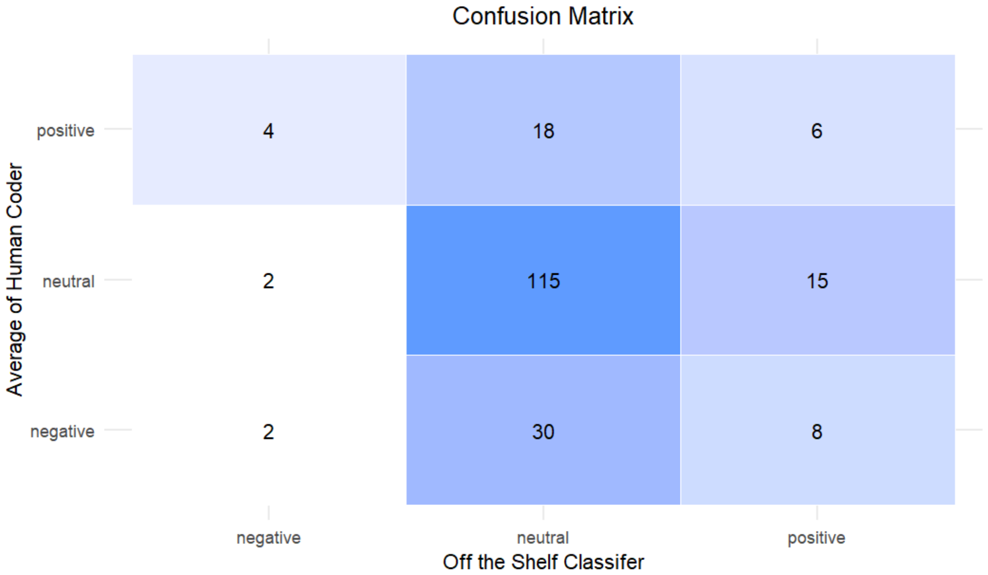


Figure 10. Stance Distribution by Period

Figure 11. Comparing Model Results

For the validation, inter-rater reliability was poor amongst the human coders, with coders only reaching moderate agreement at a Fleiss Kappa of 0.465. After averaging, I compared the model classification to the human classification, and achieved an accuracy a little better than chance at 0.615. Examination of the confusion matrix suggests generally good performance at detecting neutral articles but poor performance at detecting positive or negative articles (see Figure 12).



Note: Validation was Conducted with a Random Sample of 200 Documents and 3 Individual Coders

Figure 12. Comparing Off-the-Shelf Classifier performance to Human Coders

Discussion

My project focused on two main objectives: (1) examining changes in semantic association with the SHSAT, specifically in regards to DEI and merit and (2) assessing if stances towards the SHSAT have changed over time. To that end, I observe semantic drift across time, with DEI- and race-related terms becoming further decoupled from the SHSAT, especially race. Overall, findings suggest evidence of decreased semantic association for most terms, with the exception of some test-specific terminology, which is expected. Stance detection analysis suggests that stances, at least across news coverage, have not changed significantly over time; however, these findings are presented with the caveat that both model performance and human coder consensus were generally poor, likely due to both class imbalance and the tendency for news coverage to veer on the side of objectivity.

Examining Semantic Associations

In this case, findings suggest that over time, the SHSAT has become less and less coupled with DEI and race. The fact that test-related word pairs have near identical cosine similarity across time suggests that the results are valid, and I hypothesize that we see this pattern due to three reasons:

(1) Coverage of the racial discrepancies in specialized high school admissions occurred in 2013, so it is likely that co-occurrence of DEI, race, and test-related terminology was higher during this time period compared to post-2018

(2) 2020 saw the implementation of the discovery program, which reserved seats for disadvantaged students. As such, it is likely that newspaper framing was focused more on the desegregation of specialized high schools rather than the way that the SHSAT served as an institutional barrier.

(3) In 2018, while Mayor Bill de Blasio ran a campaign to eliminate the SHSAT, likely in response to reports of systemic racism, news coverage likely focused on the specifics of his proposal rather than the relation between racism and the test itself.

Stances towards the SHSAT

My findings overall suggested that stances have not changed significantly over time. However, these findings are weakened by poor model performance and lack of human coder consensus. The limitations for this analysis revolve around two main issues:

(1) The majority of the articles scrapped were largely neutral. While this suggests that news coverage in the NY metropolitan area has remained largely objective over time, it presents the conundrum of class imbalance for this

analysis, as there aren't enough pro-SHSAT and anti-SHSAT cases for the model to learn the patterns and properly classify them.

(2) Even with human coders, the consensus was that it was difficult to identify which articles were neutral, positive, or neutral. Agreement between coders were poor. This can be largely contributed to the fact that most of the articles in themselves were objective and stances were difficult to identify.

Limitations & Future Directions

By far, the biggest limitation of this project was the size of the corpus. While the issue was addressed somewhat with bootstrapped resampling, this analysis would have benefited from expanded scraping (perhaps to national or more local news outlets). Having an expanded dataset would also allow us to perhaps have a more granular analysis, allowing us to train word embeddings for every year, for example.

Furthermore, alternative modeling approaches could have been explored for the diachronic semantic analysis. Temporal Word Embeddings with a Compass (TWEC), for example, have been successfully used with small corpora and offer alternative methods to deal with spatial nonconformity (Bianchi, 2019).

In terms of the stance detection analysis, I could have perhaps not only employed alternative off-the-shelf classifiers, but also fine-tuned the transformers with some labelled data to increase performance. In addition, I could have refined this analysis by creating a more systematic way of identifying and dealing with target words synonyms. In this case, I took the maximum across target words but averaging probabilities would have likely been more robust, despite poorer model sensitivity.

Acknowledgments

Validation of model performance would not have been possible without our volunteer coders, Esmeralda Gloria and Mehak Kazi. Thank you for the service!

References

Bianchi, F. (2019, October 2). Aligning Temporal Word Embeddings with a Compass. Medium. https://fede-bianchi.medium.com/aligning-temporal-diachronic-word-embeddings-with-a-compass-732ab7427955

Bradford, D. (2015, February 2). In defense of New York City's selective high schools. Fordham Institute.

Corcoran, S. P., & Baker-Smith, E. C. (2018). Pathways to an Elite Education: Application, Admission, and Matriculation to New York City's Specialized High Schools. Association for Education Finance and Policy.

Hirsch, C. (2020). Constitutional Diversity in New York's Specialized High Schools: The SHSAT, the Discovery Program, and the Fourteenth Amendment | Cardozo Law Review. Cardozo Law Review, 41(1627). https://cardozolawreview.com/constitutional-diversity-in-new-yorks-specialized-high-schools-the-shsat-the-discovery-program-and-the-fourteenth-amendment/

Horne, W., Dolinsky, A. O., & Huber, L. M. (2024). mDeBERTa Stance Detection Model for Political Group Appeals [Computer software]. https://huggingface.co/rwillh11/mdeberta_NLI_stance_NoContext

Rodman, E. (2020). A Timely Intervention: Tracking the Changing Meanings of Political Concepts with Word Vectors. Political Analysis, 28(1), 87–111. https://doi.org/10.1017/pan.2019.23

Segers, G. (2018, June 7). A guide to the controversy around NYC's specialized high schools. City & State NY. https://www.cityandstateny.com/policy/2018/06/a-guide-to-the-controversy-around-nycs-specialized-high-schools/178407/

Shakarian, K. (n.d.). The History of New York City's Special High Schools. Gotham Gazette. Retrieved December 9, 2025, from https://www.gothamgazette.com/government/5392-the-history-of-new-york-citys-special-high-schools-timeline

Taylor, J. (2019). FAIRNESS TO GIFTED GIRLS: ADMISSIONS TO NEW YORK CITY'S ELITE PUBLIC HIGH SCHOOLS. Journal of Women and Minorities in Science and Engineering, 25(1), 75–91. https://doi.org/10.1615/JWomenMinorScienEng.2019026894

Zimmer. (2023, June 14). Black and Latino enrollment lags in specialized high school integration program. Chalkbeat. https://www.chalkbeat.org/newyork/2023/6/14/23759303/nyc-specialized-high-schools-discovery-program-integration-diversity/