Exploring Online Adoption Discourse in r/Adoption and r/Adopted

Ethan Kozlowski

SOCI 41033 Computational Content Analysis

March 8, 2024

James Evans

Word Count:

INTRODUCTION

Adoption is a global industry and affects the lives of millions of people around the world. Furthermore, Fisher (2003) notes that despite its clear relevance to many topics within the domain of sociology, little sociological research has been done on it. Adoption has the potential to challenge normative understandings of the nuclear family. Furthermore, for transracial adoptees—those who come from racial-ethnic/cultural backgrounds dissimilar to their adoptive parents (Silverman 1993)—there is large potential to explore the formation of racial-ethnic identities. In the ever-modernizing world, social interactions increasingly take place in online spaces. As a result, a large number of individuals, including adoptees use the Internet and social media platforms such as Reddit, Facebook, and X (formerly Twitter), to have discussions and interact with other users. From works on diaspora and social media studies like that of Yadlin-Segal (2020), social media seems to play a large role in connecting groups of people who would otherwise have no means to communicate with one another. As a result, adoptees and even more so transracial/intercountry adoptees—who are uprooted from their racial-ethnic and cultural backgrounds and often placed in environments with little to nurture interaction with their native culture—, have opportunities to explore and interact with others who find themselves in similar situations. I hope to elucidate some of the social interactions at play in the context of more niche topic subreddits based off of some facet of social identity: in my case, adoption.

RESEARCH QUESTIONS

How do the conversational dynamics and post sentiments within online adoption communities differ between general-purpose adoption subreddits like r/Adoption and adoptee-exclusive spaces like r/Adopted? Do adoptees and non-adoptees differ in their sentiment towards adoption within these different spaces? To help answer this main research question, I will also propose a handful of supplementary research questions. Are the topics discussed in r/Adoption and r/Adopted distinctly different from one another? I have a hunch that even though both subreddits were formed to discuss issues related to adoption, in reality the individuals populating these spaces discuss various aspects of it. r/Adoption, being a more general purpose, all-encompassing space for members of the adoption triad (birthmothers, adoptive parents, and adoptees), I reckon will have more diverse discussions and more topics. r/Adopted states in its about section that the subreddit is for “adoptees only” explicitly, whereas r/Adoption welcomes all “to discuss adoption-related news”.[[1]](#footnote-1) Since more adopters post in r/Adoption, there are more likely to be discussions about the process of adopting by prospective parents. Whereas in r/Adopted there might be more discussion about trauma. Furthermore, are the posts from adoptees distinctly different from posts not made by adoptees in r/Adopted and r/Adoption? This question is a relatively simple supervised ML classification task.

ABOUT THE DATA

*Data Selection and Collection*

I gathered data from two main sources. Firstly, a large portion of the posts and comments come from a Reddit archival site; r/Adoption is a sufficiently large subreddit, and was, therefore, archived on this site. One issue with this data was it was not up to date past December of 2021. Additionally, my other subreddit of interest—r/Adopted—was too small of a community to be archived on this site. Therefore, I developed a customized web scrapper to help me find Reddit posts from old.reddit.com, a simplified version of reddit.com that did not require the use of dynamic scraping packages like selenium. A decision was made early on in this paper to scrape this non-dynamic version of Reddit due to issues loading the selenium package for dynamic scraping; despite trouble shooting this issue for a few weeks, I went ahead with the somewhat flawed, but still useful data collection methods I had at my disposal to expedite the data collection process. The r/Adoption archived posts were augmented with more recent posts I found when scraping.

After designing my scraper, it took around three hours to scrape around a thousand Reddit threads (posts and their corresponding comments) for each subreddit. By the end of things, I had almost 330,000 posts and comments ready for analysis. Some of the important meta data I collected when scraping included the user’s name, the user’s flair, the post title and text, the number of comments, and the post score. The most important features of this data set were the user flair and the title/post text. The user flair is an optional feature within Reddit communities to display a short description of the user. These flairs are unique to each subreddit, and often act as a means to succinctly summarize a user’s identity within the space. For example, a user with the flair “TRA/ICA” is a shorthand way to tell others that they are a transracial and intercountry adoptee. However, because setting a user flair is optional, most users do not indicate whether or not they are adoptee. This is indicative of a machine learning classification task. For my analyses, I need to be able to identify adoptees from non-adoptees with some accuracy. I use the posts with user flairs as ground truth. I also sorted the user flairs into the binary of adoptee and non-adoptee. I manually looked at all the user flairs and made decisions on a case-by-case basis. For some tags, it was evident which side they belonged to. For example, flairs like “adoptee” or “TRA” were obviously adoptees and flairs like “birthmom,” or “future AP” were obviously non-adoptees. However, some flairs were much harder to discern; for example, many flairs overlapped in identity like “adoptee and birthmother.” I decided to encode any such flairs with double identities as adoptee. Some flairs I could not discern even when looking at the user posts themselves, and therefore I left as unlabeled despite their being a flair. I also assigned a handful of these labels (50) to my friends, and they reached an intercoder reliability of around .6422 (Cohen’s Kappa).

*Cleaning and Wrangling*

The textual data from these Reddit posts were generally clean, but still needed some attention. Firstly, I removed non-Latin alphabet characters because this might interfere with later analyses. I also attempted to remove hyperlinks using regular expressions (though upon running analyses later realized I had missed some). The archived data was much messier and required more work to remove duplicate posts, posts with no text, and posts that were later removed by subreddit moderators or the users themselves. Afterwards, I combined the post title and text into one document; comments did not have titles, so this step was ignored for them. These combined documents were then tokenized and normalized by word and by sentence.

ANALYSES

*Machine Learning*

An important task to aid in the analysis of this project is to answer this supplementary research question: Are the posts from adoptees distinctly different from posts not made by adoptees in r/Adopted and r/Adoption? Fundamentally, this involves a supervised Machine Learning classification task. As a result, I developed multiple ML models to attempt to classify the title combined with the post text of reddit threads into a binary classification: adoptee or non-adoptee. The performance results of these models are relatively poor, demonstrating the complex nature of these textual data. For example, I trained a dense neural network and a bidirectional long short term memory (biLSTM) recurrent neural network to classify these subreddits. The results of my findings are somewhat lackluster. I also tested other ML models including decision trees and logistic regression and tried incorporating ensemble methods like random forests and adaboosting. None of these models reached above 80% model accuracy or had satisfactory levels of recall. In terms of accuracy, the best model was an untuned logistic regression model at 77%. Because of these poor results, I decided not to go forward with labeling the unknown data and only conducted further analyses on posts known as adoptees and non-adoptees from the ground truth labels derived from the user flairs.[[2]](#footnote-2) Below are some of the validation metrics for the dense NN (left) and the untuned logistic regression (right). We can see that the precision, recall, and F1 score were relatively similar for both classes in the NN and the logistic regression (using TF-IDF vectorization). This was not always true for the other models A screenshot of a computer screen

Description automatically generatedA screenshot of a computer

Description automatically generatedwhich often had drastically different recall or precision rates for adoptees and non adoptees. For example, one of the tuned decision trees had an adoptee recall of around 84% and a non-adoptee recall of around 40%, which is more than halved. Though both the DNN and logistic regression have relatively high accuracy scores, they also seem to be overfitting the data still. We can see in the DNN’s validation graphics[[3]](#footnote-3) that the validation loss is increasing over time, suggesting that there is some overfitting on the training set of data. Even using logistic regression on embeddings instead of TF-IDF vectors performed poorly.

Thus, due to the somewhat lackluster performances of any of the machine learning models, this corresponding research question is somewhat fulfilled but still leaves some room for improvement. With some accuracy it is possible discern posts written by adoptees and those not written by adoptees. I choose to do this as binary classification, but future models could take into consideration the many other facets that are included within “non-adoptee”, purposefully vague and yet diverse group of people who discuss topics on adoption but are themselves not adoptees.

*Word Embedding*

*Topic Modeling*

For topic modeling, I first vectorized the high dimensional word space using TF-IDF vectors. After which I applied Latent Dirichlet Allocation (LDA) to discern the appropriate n topics. To choose the appropriate level of n for all subgroups I conducted a silhouette analysis for the first up to the first 7 topics. The number 7 was arbitrarily chosen after some testing revealed that the silhouette score was not increasing when increasing n. A screenshot of a graph

Description automatically generated

I found that n=3 had the highest average silhouette score at .069, though n=2 and n=4 were also pretty close (.065)[[4]](#footnote-4). Evidently, the complexity of the data is high due to the very low averages for silhouette scores for all values of n that were tested; values closer to 0 mean that these clusters overlap significantly.

We can take a peek into first few posts of adoptee, non-adoptee, r/Adopted, and r/Adoption to get a better sense of the topics being discussed in these contexts.[[5]](#footnote-5) Upon a deeper look in the top words for each topic[[6]](#footnote-6), the topics for r/Adoption are quite clear when drawing on knowledge of adoptions in context and from the literature. Topic0 seems to focus on around adoptees conducting birth searches. Looking at some of the corresponding posts with high levels of topic0 we see this is likely true. For example:

This is great info (I take an unplanned hiatus and look what happens). NICE WORK! My sister in law used ancestry.com's DNA feature and was able to find her cousins and via them, track down her father. Super neat.

or:

I’m not exactly sure how they get their information, but I suspect they do have access to certain Australian records. I did a search for “New South Wales public records” and came up with some promising sources -- other states are probably similar (or maybe public records are issued by the city/shire?). I imagine that the Search Squad volunteers would be able to apply what they know about searching in the US to an Australian search, even if they have no special access to subscription-only databases there. It’s absolutely worth a try.

Topic 1 and 2 for r/Adoption focus on traumatic experiences with adoption and criticisms/concerns of adoption and the foster care system respectively. For r/Adopted, the three topics somewhat meld into one. Topic 0 is reminiscent of reflecting on negative experiences around being adopted. Topic 1 relates to birth family search and the struggles and obstacles that come with that process. Topic 2 seems to relate to comments that show acts of confirmation and empathy. For example, the following post shows a lot of empathy towards the original poster (warning for explicit language):

“I completely hear you. It is horrible how, in this day and age, people just vanish without a trace and do not give a fuck how it affects you. It really hurts. I read a lot about the concept of impermanence in Buddhism and how everything changes, and try to appreciate people for the lessons they imparted upon me, and think of them like a river, like water slipping through my fingers in a river. Or some shit like that. Otherwise I might scream.”

These topics between subreddits are quite different from one another. My hunch that these subreddits would discuss different things seems to be proven true from this analysis. Especially exemplified by topic 2 for r/Adopted, these communities are meant to bring individuals with similar experiences together. There is some level of community level identity formation in the works. Adoptees in r/Adopted vocalize their support for others’ traumatizing experiences and provide sympathetic responses. Furthermore, in r/Adoption, adoptive parents and adoptees help provide perspective on the adoption process and general experiences for those who are looking into adoption for the very first time. The users in these communities, based off of the topics we have discovered, appear to generally focus on pooling experiences together and navigating this facet of identity together.

*Sentiment Analysis*

According to Hutto and Gilbert (2014), VADER (or Valance Aware Dictionary for sEntiment Reasoning) is a powerful tool for sentiment analysis, generally outperforming individual human raters. Because VADER is meant for use on social media data, I found the model particularly enticing. I incorporate this sentiment analysis tool to gather the compound sentiment from my cleaned but not tokenized data. Though I had originally considered training a fine-tuned LLM like Chat GPT to conduct a sentiment analysis, the costs of running these analyses seemed prohibitive and I instead opted for the simpler use of VADER. After conducting a 2-way analysis of variance test, the results of the statistical test are shown in the table below.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **sum\_sq** | **Degrees of Freedom** | **F-stat** | **PR(>F)** |
| is\_adoptee | 13.540146 | 1.0 | 36.756788 | 1.350233e-09 |
| subreddit | 121.939295 | 1.0 | 331.022779 | 1.144487e-73 |
| is\_adoptee \* subreddit | 0.000014 | 1.0 | 0.000038 | 9.950871e-01 |
| Residual | 14685.122317 | 39865.0 | NAN | NAN |

It is worth noting that the variables for is\_adoptee and for subreddit are both significant at the standard level of significance (α=.05). We see that the interaction effect between is\_adoptee and subreddit is not, in fact, significant. From these results we can see that being an adoptee and participating in discussion on one of the two subreddits both influence the sentiment of the user’s post. The average compound sentiment for all adoptees was around .40. The average compound sentiment for all non\_adoptees was around .45. For adoptees and non adoptees in r/Adopted, the average sentiment was .19 and .23 respectively. For users in r/Adopted, the average sentiment was .42 and .46 respectively. From the 2-way ANOVA, it is evident that the interaction between being an adoptee and participating in a specific subreddit is not statistically significant. From this information, we can conclude that user’s identity as an adoptee or non-adoptee and the subreddit they use are not multiplicative/dependent. In general, the r/Adopted subreddit posts are less positive than those of r/Adopted (evident from their means).

This result—partially being informed by cursory explorations into the social interactions of these subreddits—, is not surprising. That adoptees independent of their subreddit have generally more negative sentiment in their posts than non-adoptees is also expected. What is shocking to me is the average compound sentiment itself; I thought posts on r/Adopted would be much more negative or at least closer to 0, in general. However, these findings further the claims I made in the topic analysis section. These communities are meant to be spaces welcoming of their respective audiences: adoptees for r/Adopted and all people curious about adoption for r/Adoption. I believe that the average sentiments skew towards positive rather than what I had theorized (more negative) because of the support within the comment section of these posts that are venting about negative experiences they have faced, trauma they have endured. In these online social spaces, the community generally acts to provide support (positive sentiment) for their members, fostering a shared kinship and sense of community identity.

CONCLUSION

From the section on sentiment analysis covered above, we can properly answer the proposed research questions. Firstly, there are some distinct differences between the posts of r/Adoption and r/Adopted, and between adoptees and non adoptees more generally. From the topical analysis, we see that the posts written in r/Adoption share a topic related to adoptees currently in the process of birth searches or more generally DNA and bio-family histories. One distinct difference in topics is for r/Adopted,

*Improvements and Future Directions*

I recognize that this research paper has a few faults. One remedy for future works would be to adjust the intake of posts from both subreddits to be more equal. r/Adoption is a much larger and more active subreddit than r/Adopted, therefore I should have under sampled from this subreddit and oversampled from r/Adopted to keep the data close to being evenly distributed. Furthermore,

There are many possible avenues for future research.

APPENDIX:

Figure 1a: A screencap of the r/Adopted homepage. A screenshot of a computer

Description automatically generated

It is worth noting that the description of the subreddit is explicit with regards to its purpose as a space for “adoptees only.”

Figure 1b: A screencap of r/Adoption homepage.A screenshot of a computer

Description automatically generated

The description of the subreddit is antithetical of the previous subreddit. Explicitly, the community is welcoming of any individuals with an interesting discussing adoption-adjacant topics

Figure 2a: Silhouette Analysis for N=2

A green and black diagram

Description automatically generated

Figure 2b: Silhouette Analysis for N=4

A screenshot of a graph

Description automatically generated

A graph with blue and pink bars

Description automatically generatedFigure 3a: Top 3 topics for first 15 posts in r/Adoption (right) and r/Adopted (left)

A graph of blue and orange bars

Description automatically generated

A graph of a column with text

Description automatically generated with medium confidenceA graph of a number of people

Description automatically generated with medium confidenceFigure 3b: Top 3 topics for first 15 posts by adoptees (left) and non adoptees (right

Figure 3c: Top 10 words for N=3 topics for r/Adoption (left) and r/Adopted (right)

|  |  |  |
| --- | --- | --- |
| **Topic\_0** | **Topic\_1** | **Topic\_2** |
| parent | adoptee | child |
| child | adoption | adoption |
| family | people | adopt |
| know | trauma | foster |
| ant | experience | kid |
| adoption | post | care |
| birth | think | family |
| like | like | parent |
| think | feel | need |
| feel | comment | like |

|  |  |  |
| --- | --- | --- |
| **Topic\_0** | **Topic\_1** | **Topic\_2** |
| like | adoptee | people |
| feel | adoption | think |
| know | family | like |
| want | parent | adoption |
| adoptee | bio | family |
| adoption | know | know |
| family | like | adoptee |
| life | child | year |
| good | adopt | go |
| parent | want | get |

Figure 3d: Top 10 words for N=3 topics from posts by adoptees (left) and non-adoptees (right)

|  |  |  |
| --- | --- | --- |
| **Topic\_0** | **Topic\_1** | **Topic\_2** |
| adoption | parent | know |
| adoptee | child | family |
| people | family | want |
| child | adoptive | like |
| adopt | like | find |
| think | adopt | mom |
| want | think | birth |
| parent | adoptee | think |
| foster | kid | feel |
| like | feel | bio |

|  |  |  |
| --- | --- | --- |
| **Topic\_0** | **Topic\_1** | **Topic\_2** |
| alcohol | alternative | signature |
| wish | raise | wish |
| alternative | alcohol | alcohol |
| question | kidnap | life |
| narrative | ultimately | abuse |
| mom | adopt | question |
| abuse | wish | adoptive |
| sorry | complete | happen |
| adoptive | mom | ultimately |
| commenter | adoptive | bring |

Figure 4a: DNN validation graphs

A graph with blue dots

Description automatically generated A graph with blue dots

Description automatically generated

Figure 4b: biLSTM validation graphs

A graph with blue dots

Description automatically generated A graph with blue lines

Description automatically generated

Works Cited/Consulted

Fisher, Allen P. 2003. “Still ‘Not Quite as Good as Having Your Own’? Toward a Sociology of Adoption.” *Annual Review of Sociology* 29:335-361. doi:10.1146/annurev.soc.29.010202.100209.

Hutto, C.J. and Eric Gilbert. 2014. VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. *Association for the Advancement of Artificial Intelligence*. Retrieved March 8, 2024, (https://ojs.aaai.org/index.php/ICWSM/article/download/14550/14399/18068)

Davis, Jenny L., Tony P. Love, and Phoenicia Fares. 2019. “Collective Social Identity: Synthesizing Identity Theory and Social Identity Theory Using Digital Data.” *Social Psychology Quarterly* 82(3):254-273. doi:10.1177/0190272519851025.

Silverman, Arnold R. 1993. “Outcomes of Transracial Adoption.” *The Future of Children* 3(1):104-118. doi:10.2307/1602405.

Stuhler, Oscar. 2021. "What’s in a category? A new approach to Discourse Role Analysis." *Poetics* 88. doi:10.1016/j.poetic.2021.101568.

Waller, Isaac and Ashton Anderson. 2021. “Quantifying social organization and political polarization in online platforms.” *Nature* 600(7888):264-268. doi:10.1038/s41586-021-04167-x

Yadlin-Segal, Aya. 2020. “Articulating Persian identities between Iran and Israel: On nationality, diasporas, and lived ethnicities in online media.” *Nations and Nationalism* 27:347-362. doi:10.1111/nana.12639.

1. See Figures 1a and 1b in Appendix [↑](#footnote-ref-1)
2. For further discussions about model performances of the ML classification models please the p3\_evaluation notebook in GitHub [↑](#footnote-ref-2)
3. See Figure 4a [↑](#footnote-ref-3)
4. See Figure 2a and 2b [↑](#footnote-ref-4)
5. See Figure 3a and 3b [↑](#footnote-ref-5)
6. See Figure 3c and 3d [↑](#footnote-ref-6)