A Study of LGBTQ+ Wikipedia Articles Sentiment over Time

Henry Lozada
A15127559
hlozada@ucsd.edu

Parth Patel
A14410868
pmp006@ucsd.edu

Emma Logomasini
A14125382
elogomas@ucsd.edu

Yuanbo Shi
A14892544
yus263@ucsd.edu

Abstract

We detail a specific method that determines how, if at all, sentiment changes over time for a category of Wikipedia articles, which, in our study, are articles categorized by Wikipedia as LGBT articles. This method uses three different sentiment analyzers, one for each of the three different language editions of Wikipedia we are analyzing, to calculate the sentiment of a Wikipedia article, doing so for all edits in the article's revision history and for all articles in each language's LGBT category. This enables us to calculate a fixed effects regression for each language's sentiment scores, allowing us to determine whether or not time has a positive effect on the articles' sentiment scores, as well as to compare these trends across languages.

1 Introduction

With its 20 year history, Wikipedia is not merely a resource of crowd-sourced information; it is a reflection of changing times and attitudes across the world, as the definition of what we consider "public knowledge" changes with the times. The significance of the last 20 years for social and political developments in LGBTQ+ representation also cannot be understated. Within that time, 29 countries around the world have legally recognized and performed same-sex marriages[2]. Given such advances, we aim to explore their impact on Wikipedia data.

Other papers also seek to understand Wikipedia data in relation to LGBTQ+ representation. "Multilingual Contextual Affective Analysis of LGBT People Portrayals in Wikipedia" performs contextual affective analysis to examine the Wikipedia pages for LGBTQ+ individuals across different languages: Russian, English, and Spanish. This paper primarily seeks to identify the nuanced language individuals with LGBTQ+ identities or ties are discussed, and performs this by measuring the connotation of common verbs in the fields of agency, power, and sentiment in said articles. Using the connotation frames, "lexicons of verbs annotated to elicit implications," for study helps frame a language's unconscious bias[3]; however, our team primarily concerns ourselves with focus on sentiment. This somewhat narrows our scope of nuanced analysis, but allows us to identify broader trends in the data. In addition, unlike this paper, we want to look at changes in these trends.

In our analysis, we attempt to measure and quantify this change for articles related to LGBTQ+ issues. Specifically, we wish to understand these changes in different languages, so we can cross-compare trends. The languages we choose to focus on are English, Spanish, and Chinese.

2 Methodology

The first step in our analysis is identifying the pages on which to perform analysis. Wikipedia has article categories which help us to identify which pages are relevant to our analysis. However, after investigating the articles marked un-

der the "LGBT" Wikipedia category, we found that the sheer volume of pages which are marked "LGBT", as well as the fact that not all of them are strongly related to what we're looking for, means that a slightly more manual approach is necessary. We chose to manually select a number of sub-categories to analyze based on category size and article relevancy. This serves the purpose of being manual enough for us to tweak our selection to be as relevant and small as we need it to be, while also being automatic enough so that we don't have to pick through articles one by one. After this, we query Wikipedia for the edit data for each page. Holding all this data in RAM at once is impossible, so each page's edit history is saved locally for sentiment analysis. Each page's edit history is stored as a separate JSON file. We chose this approach over a single large JSON file as it makes it easier to implement parallelization in both this and the following step.

Language	Article Count
Spanish	1606
Chinese	992
English	913

Table 1: Number of Articles for Each Language

Once the page histories are stored, we can begin analyzing the sentiment for each edit, for each page. This is as simple as loading each edit history and looping through the edits, running the relevant language's sentiment analyzer. To keep this code as clean as possible, a wrapper class is used for sentiment analysis. This has the added benefit of making the pipeline easily extensible, as all this analysis can be done for any other language once an analyzer for that language is added to the sentiment analysis wrapper class. The results for all pages are stored in a single json file after this step, as opposed to multiple files. This is done for two reasons: A single file is easier to transfer between machines; and we no longer need to worry about parallelization after this step. The single float which symbolizes sentiment is much smaller than the strings which represent article text, so RAM isn't a problem anymore either.

Finally, the data is moved into a Pandas DataFrame for fixed effects regression. This helps us control for the level of heterogeneity between articles (the difference in views, and therefore edits, between Wikipedia articles can be astronomical) as well as pinpoint which articles are having the most change in sentiment and when. The timestamps from the sentiment json file are converted to years in order for us to calculate the average and median sentiments for an article per year, as we are aiming to calculate trends over years rather than over seconds, which is the granularity of timestamp data. In order to complete the fixed effects regression, we create dummy variables for each article in the data, making sure to drop the first column (which represents an article) to avoid the issue of multicollinearity as best as possible. The dependent variable of the regression is the previously calculated average sentiment score and the independent variables are the year column and all the dummy variable columns. We additionally calculate a second fixed effects regression, with the only change being the dependent variable is now the median sentiment score. We do this to see if the effect of outliers in the sentiment score data is significant enough to change the result of the regression.

3 Analysis

Our analysis primarily deals with sentiment calculated in our JSON Data. To get this data, first, we collect the article's edit history with the the time and date these edits took place. Next, we pass the text of the article through a sentiment analyzer, which outputs a float value ranging from 0 to 1. A value of 0 indicates negative sentiment, while a value of 1 indicates positive sentiment. We do this for every revision of the article so that all edits from its creation up until its most recent edit are covered, and we do this for all three languages, each with their own, unique sentiment analyzer.

For every year in 2003-2021 (Wikipedia's inception to more recent edits), we calculate the average sentiment in each language for edits. Our results do not match our expectations of a positive and linear sentiment relationship.

The data for the English files is the most extensive. As shown in Figure 1, there is actually a decrease in sentiment over time for this data with a very weak negative correlation. The Spanish data (Figure 3) actually shows a greater decrease in sentiment. Comparatively, the Chinese data does not have a simple linear relationship. Before 2013, the sentiment of these articles increases, and after 2013, it decreases over time.

Our next steps then deal with trying explain these phenomenon.

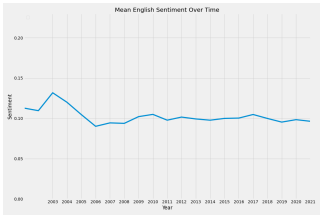

Figure 1: English Mean Sentiment Edits Over Time

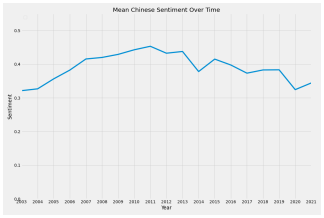

Figure 2: Chinese Mean Sentiment Edits Over Time

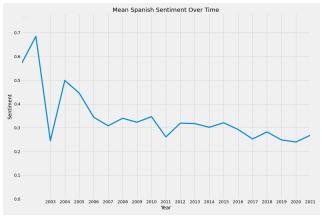

Figure 3: Spanish Mean Sentiment Edits Over Time

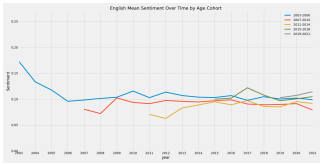

Figure 4: English Mean Sentiment Over Time With Age

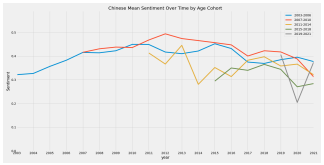

Figure 5: Chinese Mean Sentiment Over Time With Age

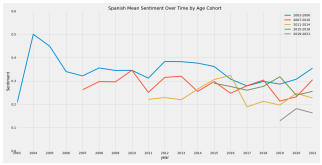

Figure 6: Spanish Mean Sentiment Over Time WIth Age

4 Results

In our results, we attempt to explain the sentiment results that we discovered. In particular, we look at the age of the articles. After looking at different features of the articles, we find that the age of the articles as a major contributing factor to the mean sentiments that we saw, although this is more true for certain languages than others.

For both the Chinese and the Spanish data, the means are dragged down by more recent articles and their sentiment because they act as lower outliers (Figure 5, Figure 6). This does not occur in the English sentiment data. There is not a significant difference between newer article sentiment and older sentiment; however, this makes sense as after 2006, the sentiment of English articles has no correlation with the year (and even considering all the years, the found correlation is not significant).

For the full analysis of the aforementioned category, we compute the sentiment scores for each edit of each article and apply fixed effect regression to the data, as described in the methodology section. We use both mean and median values to make sure that outliers are not going to change our result significantly. As seen in Figures 7-12, across all languages and using both methods of fixed effects regression explained previously, time has a weak negative relationship with sentiment over time. The sentiment scores are distributed without any obvious pattern across different languages. The mean and median graphs are not very different from each other, which means that the outliers in the sentiment data are not very significant. However, in Figure 12, the regression line for the Median Spanish Sentiment Score by Year appears to be closer to 0 than the regression line for the Mean Spanish Sentiment Score by Year in Figure 9, indicating that the outliers in the Spanish sentiment data skewed more in the positive direction. This is an indication that the majority of Spanish sentiment scores were negative, with fewer receiving positive scores, all hallmarks of a right-skewed distribution.

5 Discussion

Since we did not analyze every article in the LGBT category but only one main subcategory, much work remains to be done for all other LGBTQ+ related articles. As shown in Table 2, lots of the articles related to the LGBT community have not been taken into account due to computing and time constraints. Though our regression shows that time has a weak negative relationship with sentiment over time, it is possible that other parts of the data may have a different pattern that a full study over all the articles in LGBT category would uncover, as a trend in one subcategory does not necessarily indicate the same trend across all other subcategories. As it will take a very long time to process all the articles (approximately weeks), more efficient algorithms are quite necessary to delve into more widespread analysis.

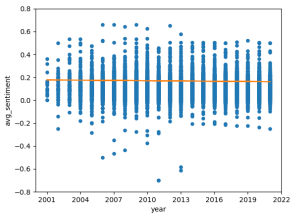

Figure 7: Mean English Sentiment Score by Year

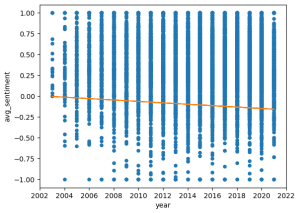

Figure 8: Mean Chinese Sentiment Score by Year

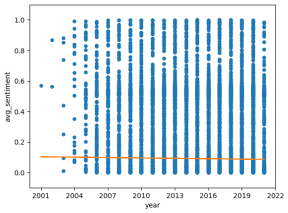

Figure 9: Mean Spanish Sentiment Score by Year

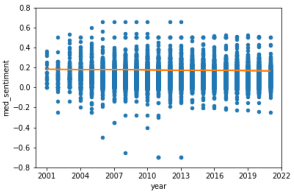

Figure 10: Median English Sentiment Score by Year

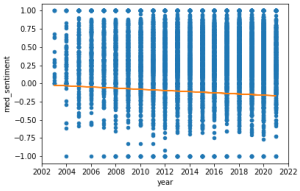

Figure 11: Median Chinese Sentiment Score by Year

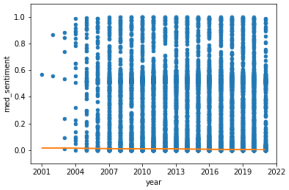

Figure 12: Median Spanish Sentiment Score by Year

Subcategory	Article Count
Transgender	61612
LGBT culture	27986
LGBT by region	27665
LGBT history	35652
LGBT people	19472
LGBT studies	5616

Table 2: Top 6 subcategories of English Wikipedia under category LGBT

Another important part that needs improvement is the cleaning of the texts. There are many different types of non-text contents that should be removed before sentiment analysis to improve the scores. We are currently using WikiExtractor[1], but there are still many elements remaining, including links and reference marks (something like "[1]").

Finally, though we dropped the first column of our dummy variable dataframe in an attempt to avoid the dummy variable trap, the sheer amount of dummy variables in the dataframe seems to cause multicollinearity to persist. If this same analysis were to be done on a larger subcategory, the problem would only be magnified. Future approaches to analyzing collections of Wikipedia articles for sentiment changes over time would benefit greatly from using methods that would reduce multicollinearity as much as possible so that the trend of change over time could be captured more efficiently.

References

[1] Giusepppe Attardi. “WikiExtractor”. In: GitHub, 2015.

[2] *Marriage Equality Around the World*. URL: https://www.hrc.org/resources/marriage-equality-around-the-world.

[3] Chan Park, Xinru Yan, Anjalie Field, and Yulia Tsvetkov. “Multilingual Contextual Affective Analysis of LGBT People Portrayals in Wikipedia”. In: Oct. 2020.