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[1]. 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 languages 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.

Other papers we pull from include “Censorship's Effect on Incidental Exposure to Information: Evidence From Wikipedia”, where this paper looks at how the censorship in China affected the Wikipedia use, where certain topics and their frequency of use changed because of the ban[2]. This should be noted since we are looking at Chinese Wikipedia data.

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 under 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	100

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 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 between Wikipedia articles can be astronomical) as well as pinpoint which articles are having the most change in sentiment and when.

3 Initial Analysis

In our initial exploratory data analysis, we look at the specific article “Same Sex Marriage” across all three languages. We chose this Wikipedia article because of its age (it was a part of the Wikipedia contents since near inception for all languages) and its relevancy to our topic. As it a singular page, we cannot make broad sweeping statements.

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.

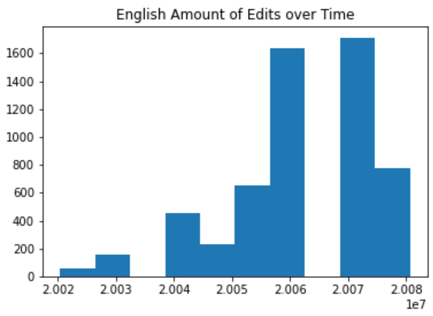
The data for the English files is incomplete, and there was no correlation for positive or negative associations over time. There is also an increasing amount of edits over time, as seen in the graphs above.

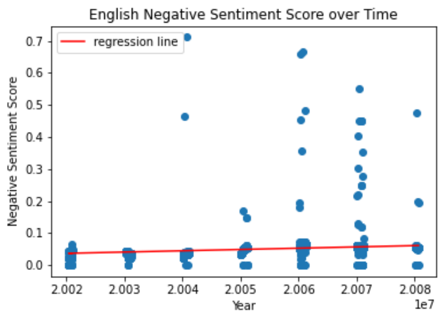
For the Chinese sentiment analyzer, the data implies positive correlation over time. There is also an increasing amount of edits over time, as seen in the graphs at the top of the next page.

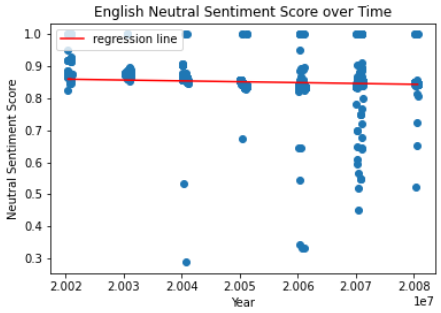
In the Spanish sentiment analysis, the data implies weak positive correlation over time. There is also an increasing amount of edits over time, as seen in the graphs at the top of the next page.

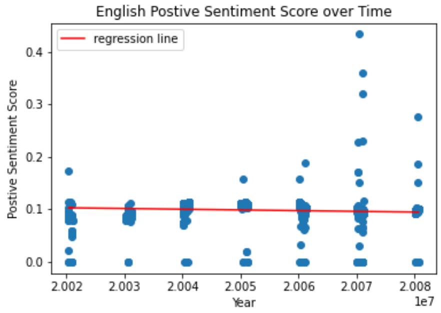
4 Results

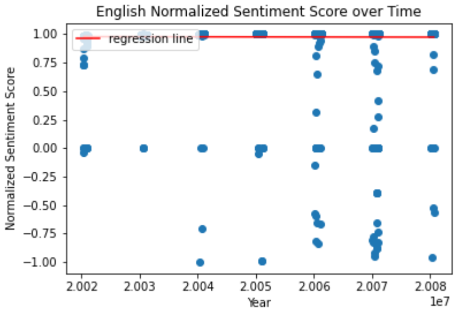
For issues that we encounter in our initial analysis, we find that our querying of article data sometimes leads to a limit in the amount of articles allowed, especially for articles with larger amounts of edits. This primarily presents itself as a problem for our English articles; in our EDA, we find that not only does it have the highest amount of edits (with 5576), but this only covers the edit history of that page from 2002 to 2009. In our pipeline, we need to make ad-

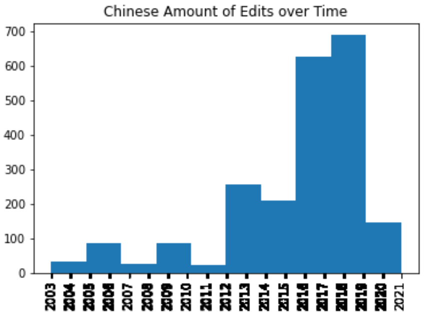


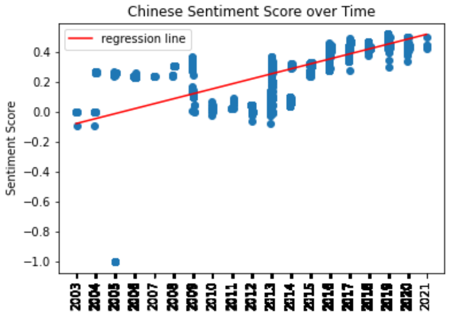












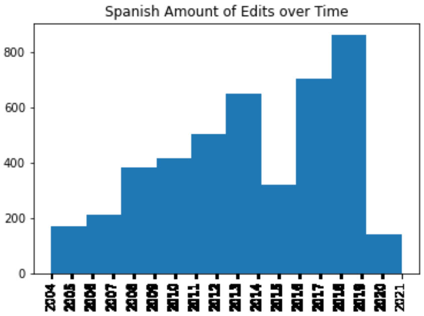
justments to allow for querying of such pages. We also have other problems with the English data; we are uncertain if our initial analysis reflects broader English trends, issues of our sentiment analyzer, or if it has to do with our querying problem.

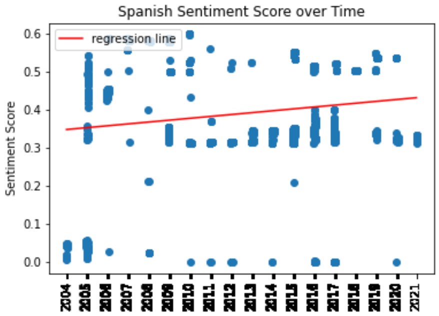
Additionally, we found that this time spent on the sentiment analysis and querying was excessive, leading to the possibility of scaling down our project.

We compute the sentiment scores for each edit of each article and apply fixed effect regression to the data.

5 Appendix

Project Proposal (Github repo) We want to examine the changes in public perception of issues concerning the LGBT community by analysing change in sentiment in Wikipedia data. This





data has been acquired before - the paper "Multilingual Contextual Affective Analysis of LGBT People Portrayals in Wikipedia" by Chan Young Park, Xinru Yan, Anjalie Field, and Yulia Tsvetkov performs the precise type of sentiment analysis we wish to perform. However, that analysis was performed on the pages as they were at the time - our goal is to examine the sentiment over time, especially around those times which feature prominent events which had strong effects on the LGBT community and its public perception. In other words: we know sentiment changes between cultures; we want to find out if sentiment changes with culture. We will perform this analysis as an academic paper. The techniques we use will follow the aforementioned paper: We will identify sentiment analyzers in NLPs which are best for each language and apply them to selected articles. As these pages are likely to be more contested given the nature of LGBT rights over the last two decades, there should be plenty of data for time series analysis. For our initial EDA, we plan on approaching it with the following 2 methods. First, we plan on looking at one of our chosen events across the 3 languages we are interested in (English, Chinese, and Spanish). Then, we plan at the page views and pages creation of LGBTQ+ before and after that event. In addition, we want to identify one event and Wikipedia page and perform sentiment analysis. We count the number of articles with some certain keywords (LGBT, lesbian, gay and etc.) in the title to check if there're enough edits for research. The results are listed below:

Keyword	**Count**
LGBT	62210
Gay	94396
Lesbian	14254
Bisexual	5842
Transgender	4517

Table 2: Keyword Edit Count Results

We notice that many articles with Gay in the title are actually not related to LGBT but some people named as Gay. The 5 keywords we choose are just a part of all the articles related to LGBT community. The total number of edits is roughly enough for sentiment analysis or word embedding.

References

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

[2] Jennifer Pan and Margaret E. Roberts. “Censorship’s Effect on Incidental Exposure to Information: Evidence From Wikipedia”. In: 2020.

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