Journal of Open Humanities Data

A Mixed Method Twitter Methodology and Anonymous Corpus

DATA PAPER

TIM RIBARIC

ubiquity press

ABSTRACT

This dataset represents the first 5 years of posts made by the anonymous twitter bot @lis_grievances combined with a series of custom and pre-established metrics. The bot is a platform for workers in libraries and affiliated fields to make unattributable pronouncements. A simple vetting process ensured no defamatory or explicit posts were made. Anonymity is assured as no evidence of who made the submission is retained, even to the operator of the bot. This dataset represents a collection of thoughts, fears, and cutting remarks made by information workers about their field and the places where they work.

CORRESPONDING AUTHOR:

Tim Ribaric

Library, Brock University, St. Catharines, Canada

tribaric@brocku.ca

KEYWORDS:

Twitter; online disinhibition effect; sentiment analysis

TO CITE THIS ARTICLE:

Ribaric, T. (2023). A Mixed Method Twitter Methodology and Anonymous Corpus. *Journal of Open Humanities Data*, 9: 8, pp. 1–6. DOI: https://doi.org/10.5334/johd.109

(1) OVERVIEW

REPOSITORY LOCATION

https://doi.org/10.5683/SP3/PHWSVM.

CONTEXT

The @lis_grievances bot was first activated February 26, 2016, and continues to operate to this day. The messages that it tweets are all anonymously harvested and thus allow workers (presumably) in libraries and related fields to 'air their grievances' (@lis_grievances bot, n.d.). There has been an enthusiastic discussion in social media and within Library and Information Science (LIS) literature on the utility of the bot and both its harm and benefit to the profession (Skyrme & Levesque, 2019).

The five-year archive was created as the basis for a chapter (Ribaric, 2022b) within a monograph that investigated the hypothesis that Libraries are dysfunctional workplaces (Acadia, 2022). This research conducted an analysis that partitioned the tweets into various categories in order to understand themes found in the corpus. It also introduced a novel metric called the *Grief Index* (GI), which gave a quantitative ratio of how many submissions made to the bot were not posted. Every submission made to the bot is checked by a moderator before it is posted, this is to ensure that no one is specifically mentioned in a tweet and that discriminatory language is not used. This difference in submissions to the bot versus actual posted messages is the basis of GI. This value provides a proxy for understanding the amount of material submitted that is not suitable for posting, which also avoids the need to share actual inappropriate posts.

(2) METHOD

STEPS

The actual dataset is an aggregation of tweets made by the @lis_grievances account (n = 4096) and retrieved from the Twitter API using the Tweepy platform (Roesslein, 2020). Some metadata of the tweets is retained and augmented with a custom metric called the *Grief Index* as well as the three components of the VADER sentiment score for the full text of the tweet (Hutto & Gilbert, 2014). The complete software used to create the bot is hosted on GitHub (Ribaric, 2022a). A key component of this software is that it contains a mechanism that retrieves the direct messages sent to the bot through a process that ensures anonymity of the sender from the operator of the bot. The Twitter archive of the account was requested on February 27, 2021 and as such, any favourite or retweet counts is current to that day. The etymology of 'bot' is preferred to describe this account since the posting and retrieval of messages is mediated through an API interface in conjunction with custom software that ensures anonymity of posts. While all posts are submitted by humans, no submission is posted without a comprehensive mediated quality control process.

The basis of the analysis was the creation of a metric dubbed Engagement Score (ES) which was the sum of the retweets and favourites a tweet received in the 5-year period. By combining this quantitative scoring with a close reading of the tweets a mixed method was conducted. Tweets with a high ES were examined in an attempt to uncover themes present in the full corpus.

COLUMN DESCRIPTION

Description of all columns retained from the Twitter archive export and additional data added as part of the analysis can be found in Table 1.

This dataset is a combination of data exported directly from Twitter and enriched with additional analysis specific description of the provenance of each column, as described in Table 2.

ID	GLOBAL ID OF TWEET AS STRING
favorite_count	integer count of how many times tweet was favourited
retweet_count	integer count of how many times tweet was retweeted
created_at	timestamp of when tweet was made
full_text	full text of tweet
entities.hashtags	list of hashtags found in tweet
entities.symbols	list of symbols found in tweet
entities.user_mentions	list of users mentioned in the tweet
entities.urls	list of URLs found in the tweet
possibly_sensitive	flag autogenerated to indicate possibility of sensitive content
entities.media	python list of media found in tweet, e.g. images
full_text_norm	Normalized full text of tweet
vscore_pos	VADER positive dimension score for the tweet full-text, 0 to 1 inclusive
vscore_neg	VADER negative dimension score for the tweet full-text, 0 to 1 inclusive
vscore_neu	VADER neutral dimension score for the tweet full-text, 0 to 1 inclusive
vscore_compound	VADER composite score for the tweet full-text, –1 to 1 inclusive
swears	flag if tweet contains a swear word
engaged	flag if tweet was either favourited or retweeted
total_engagement	Engagement score, ie. combined count of number of retweets and favourites
hashtags	flag if tweet contains a hashtag
questions	flag if tweet contains a question (full text includes a question mark)
media	flag if tweet contains image
fav_quant	what quantile tweet is in based on favourite count, if applicable
g_index	the grief index value for the month that the tweet was made

Table 1 All columns found in the dataset.

PROVENANCE	COLUMNS
Twitter Export	favourite_count retweet_count created_at full_text entities.hastags entities.symbols entities.user_mentions entities.urls possibly_sensitive entities.media
Derived	full_text_norm vscore_pos vscore_neg vscore_neu vscore_compound swears engaged total_engagement hashtags questions media fav_quant g_index

Table 2 Description of data origin, either direct from Twitter export or result of analysis.

(3) DATASET DESCRIPTION

Object name – LIS_G_5_YEAR_ARCHIVE

Format names and versions – .CSV

Creation dates – 2016-02-26 to 2021-02-27

Dataset creators – Tim Ribaric

Language – English

License – CC0 1.0

Repository name – Borealis

Publication date – 2022-10-11

STATISTICS AND CONTENTS

As mentioned, the investigation focused on ES of the corpus, but it contrasted this score against other dynamics of the tweets. Box plots of the different facets used to partition the tweets are seen in Figure 1. Here we see that the inclusion of swears, for example, lead to a higher mean score compared to other facets.

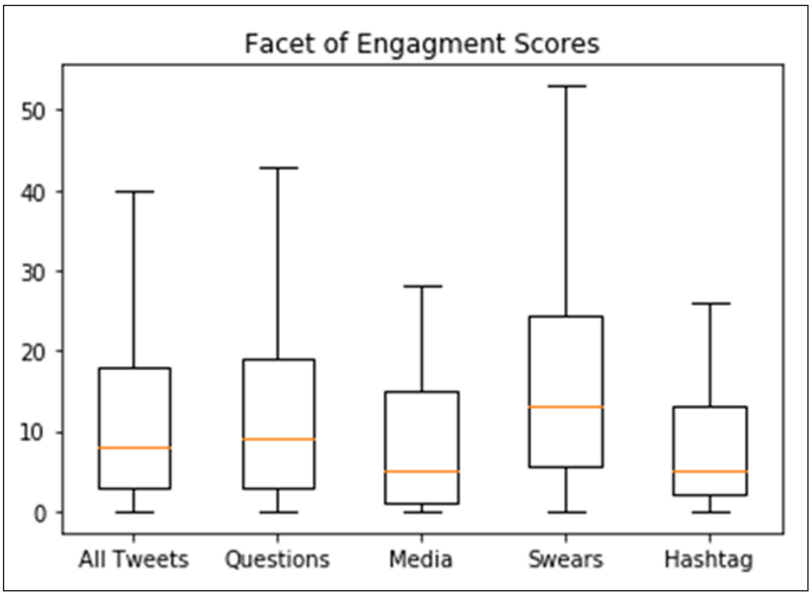

Figure 1 Engagement score box plots for tweets with different characteristics.

Figure 2 shows the general distribution of ES across all tweets in the corpus.

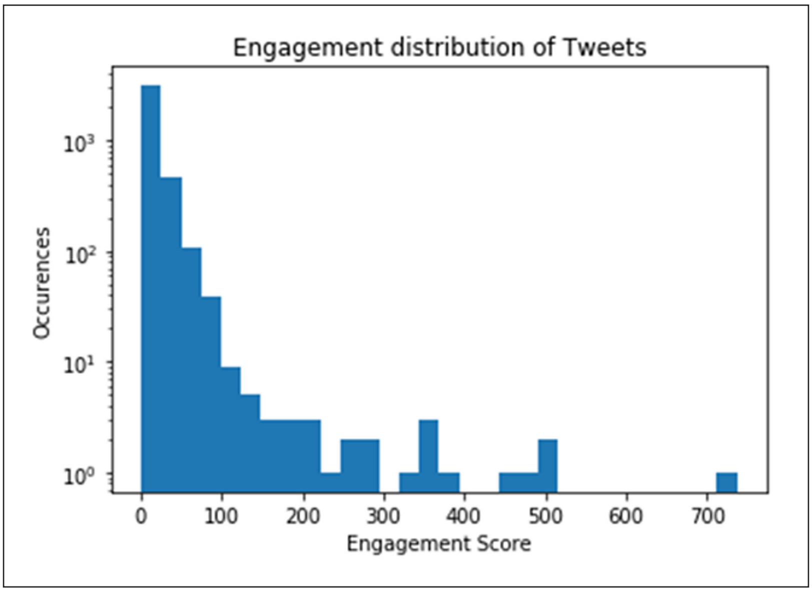

Figure 2 Engagement score distribution of all tweets.

This provides us with a quick view of the distribution of scores along with some evidence that outliers were also present. To further shed light on the corpus, VADER sentiment scores were also calculated. Figure 3 shows an example of sentiment score composition for the swear word facet.

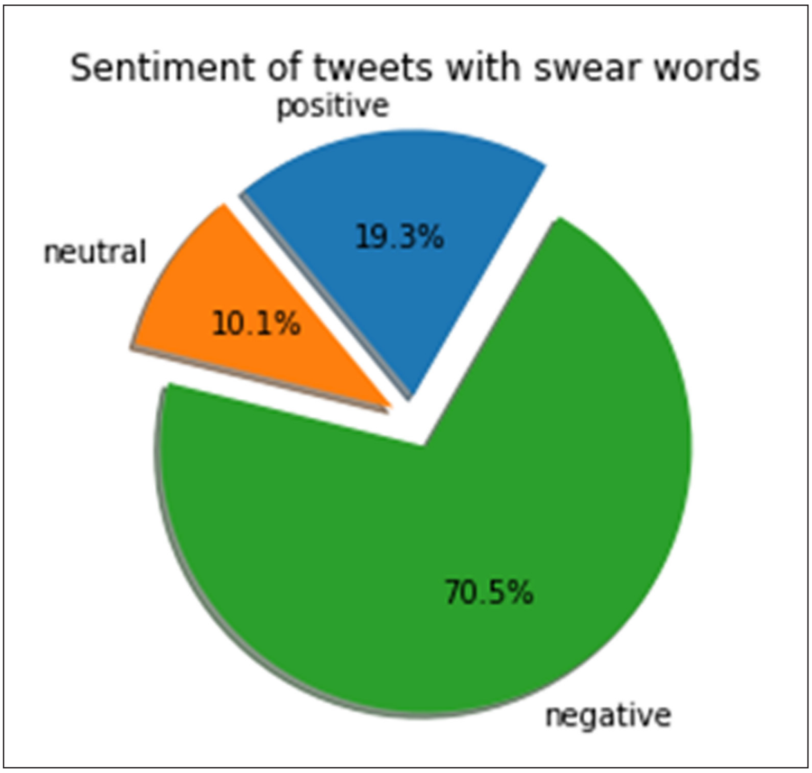

Figure 3 VADER sentiment score breakdown of all tweets in the archive.

Lastly, to provide a general sense of what is in the corpus, a word cloud is presented in Figure 4. It appears that Librarians enjoyed talking about themselves and the places in which they work.

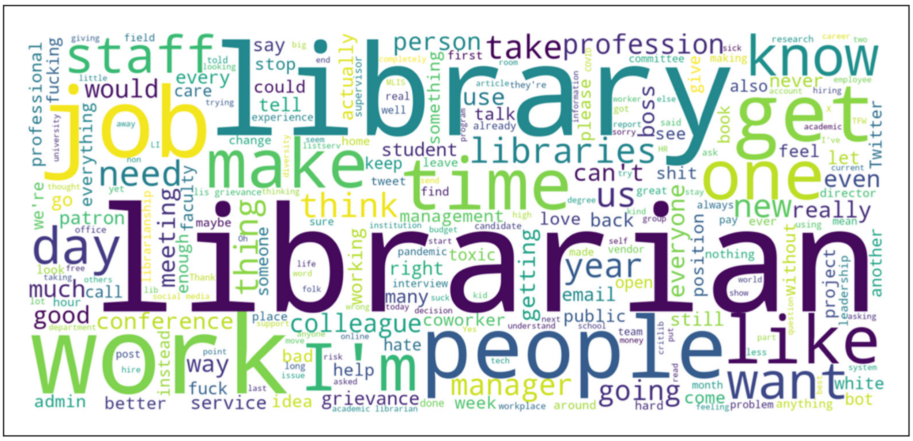

Figure 4 Word cloud of all the tweets in the archive.

(4) REUSE POTENTIAL

The primary goal of the research was to propose a mixed-method approach to analysing the corpus in order to derive insights into its contents without the need of having a researcher examine each tweet and hand-code for themes; however, many other uses of the corpus can be devised. This archive of tweets has potential to inform investigations in many different areas. For example, it can be used to assess the perceived accuracy of the VADER sentiment analysis scoring system. It can also be used to study the online disinhibition effect (ODE). ODE is the supposition that when given anonymity people will express themselves in stronger ways than if their speech is attributed.

Lastly, the dataset can be used for sociological or LIS inquiry, such as to investigate a profession's self-image. Within the LIS field, romanticisation of professional self-identity is known as vocational awe (Ettarh, 2018). One tenant of vocational awe is that librarians in the course of their work will put up with outrageous workplace deficiencies simply because of the importance of the job. This candid archive of librarian self-reflection could very well prove useful in an examination of this phenomenon.

COMPETING INTERESTS

The author has no competing interests to declare.

AUTHOR CONTRIBUTIONS

Tim Ribaric: Conceptualization, Formal Analysis, Investigation, Methodology, Software.

AUTHOR AFFILIATION

Tim Ribaric orcid.org/0000-0001-9229-8569
Library, Brock University, St. Catharines, Canada

REFERENCES

Acadia, S. (Ed.) (2022). *Libraries as dysfunctional organizations and workplaces*. New York: Routledge. DOI: https://doi.org/10.4324/9781003159155

Ettarh, F. (2018). *Vocational Awe and Librarianship: The Lies We Tell Ourselves – In the Library with the Lead Pipe*. Retrieved from https://www.inthelibrarywiththeleadpipe.org/2018/vocational-awe/ (last accessed: April 28, 2023).

Hutto, C. J., & **Gilbert, E.** (2014, May 16). VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text [Full Paper]. *Eighth International AAAI Conference on Weblogs and Social Media*, Ann Arbor, Michigan, US. DOI: https://doi.org/10.1609/icwsm.v8i1.14550

@lis_grievances bot. (n.d.). *LIS Grievances—About*. Retrieved from https://lisgrievances.com/about.html (last accessed: April 17, 2023).

Ribaric, T. (2022a). *LIS Grievances*. Retrieved from https://github.com/elibtronic/lis_grievances (last accessed: April 28, 2023).

Ribaric, T. (2022b). “Put the fucking salary in the job ad!” In S. Acadia (Ed.), *Libraries as Dysfunctional Organizations and Workplaces* (1st ed., pp. 167–192). New York: Routledge. DOI: https://doi.org/10.4324/9781003159155-8

Roesslein, J. (2020). *Tweepy: Twitter for Python*. Retrieved from https://Github.Com/Tweepy/Tweepy (last accessed: April 28, 2023).

Skyrme, A. E., & **Levesque, L.** (2019). New Librarians and the Practice of Everyday Life. *Canadian Journal of Academic Librarianship*, *5*, 1–24. DOI: https://doi.org/10.33137/cjal-rcbu.v5.29652

TO CITE THIS ARTICLE:
Ribaric, T. (2023). A Mixed Method Twitter Methodology and Anonymous Corpus. *Journal of Open Humanities Data*, 9: 8, pp. 1–6. DOI: https://doi.org/10.5334/johd.109

Submitted: 27 April 2023
Accepted: 14 June 2023
Published: 29 June 2023

Journal of Open Humanities Data is a peer-reviewed open access journal published by Ubiquity Press.