

The Neil deGrasse Tyson Problem: Methods for Exploring Base Memes in Web Archives

Amelia Acker

School of Information, University of
Texas at Austin, Texas, USA
aacker@ischool.utexas.edu

Anne C. Loos

School of Information, University of
Texas at Austin, Texas, USA
annecl@utexas.edu

Julia Sufrin

School of Information, University of
Texas at Austin, Texas, USA
julia.sufrin@utmail.utexas.edu

ABSTRACT

In this paper we introduce the concept of the “base meme” for characterizing unique information artifacts that are used to make derivative, new, and related memes. Base memes are antecedents to many versions of derivative memes that are published all across the web. While they can be created in meme template generator websites, their origins and diffusion can be difficult for researchers to verify. Despite the often ephemeral nature of memes that are shared via platforms, they can be fairly reliably found in web archive collections, such as the Internet Archive and the US Library of Congress’ Web Cultures Web Archive. In this paper, we first present the existing research on memes and discuss the challenges for researchers who study them (such as identification and language detection). We then describe the importance of web archives to social media research and building robust methods of inquiry for internet history. Using archived data from the Library of Congress’ Meme Generator Archive (N=57,652), we use descriptive analysis to calculate, measure, and describe this important public web archive of memes. Our results show that this collection has a variety of “base memes” that can be grouped with their related derivative memes (which we consider to be their related works). We use language detection software to identify a variety of languages present in the archived dataset of memes. We close by describing why approaching these metrics on “base meme” image macros alongside findings for derivative versions and the multiple languages present in web archives of social media allows researchers to study a diversity of voices, including linguistic diversity, distinctions in humor, and the variety of cultural expressions present in memes.

CCS CONCEPTS

• **Human-centered computing**; • **Collaborative and social computing**; • **Collaborative and social computing theory, concepts and paradigms**; • **Social networks**;

KEYWORDS

Base memes, language detection, Meme Generator, web archives

ACM Reference Format:

Amelia Acker, Anne C. Loos, and Julia Sufrin. 2020. The Neil deGrasse Tyson Problem: Methods for Exploring Base Memes in Web Archives. In *International Conference on Social Media and Society (SMSociety '20)*, July 22–24, 2020, Toronto, ON, Canada. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/3400806.3400836>

1 INTRODUCTION

Since 2014, the American Folklife Center (AFC) has been archiving the content on the World Wide Web. This vast collection documents “various digital vernaculars” and “emergent cultural traditions” of the web [1]. Located at the United States’ Library of Congress (LOC), the AFC’s Web Cultures Web Archive is contains items ranging from UrbanDictionary.com [2], a crowdsourced dictionary of slang words and phrases founded in 1999, to Emojipedia.org [3], the universal reference for standardized updates of emoji for Unicode. It also contains a collection of various of meme resources, including KnowYourMeme.com, the so-called “Internet Meme Database,” [4], YTMND.com, an online community centered on the creation of memetic web pages, and MemeGenerator.net [5], a website that allowed users to create, copy, and edit memes with templates. The AFC Web Cultures Web Archive considers memes and other internet artifacts, like emojis and GIFs, to be instances of 21st century folklore, and therefore helpful in documenting how cultures have changed and developed online [6]. This paper reports on a descriptive analysis of the derivative data collected from the LOC’s Meme Generator Web Archive, drawn from their web archive crawls of MemeGenerator.net.

Memes began circulating on Usenet Forums in the 1990s. As social media platforms like YouTube, Facebook, and Twitter gained prevalence, sharing and posting memes became increasingly popular, and memes gained status as internet artifacts. Since the early 2000s, websites emerged that allowed users to create and publish new memes using templates. One such website, MemeGenerator.net, was established in the early 2010s, and its homepage was first captured by the LOC Web Archives on July 25, 2012 [5]. The website allowed users to generate new memes by uploading their own images or by using existing templates. As of this writing, MemeGenerator.net is “down” on the internet; since as early as December 27, 2019 visitors to the website have been met with a “522 Error” message, which suggests that a connection error has occurred between the site’s content delivery network (CDN) and the original server. It is still unclear when, if ever, the site will be put back online.

The Meme Generator archive, and its derivative dataset of descriptive metadata about each meme in the collection, was created on May 5, 2018, from web crawls using the Heritrix External web crawler [7]. While there are many websites archived in the AFC Web Cultures Collection that feature internet memes, the LOC has

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
SMSociety '20, July 22–24, 2020, Toronto, ON, Canada
© 2020 Association for Computing Machinery.
ACM ISBN 978-1-4503-7688-4/20/07...\$15.00
<https://doi.org/10.1145/3400806.3400836>

made only two sets of derivative data from the collection available to public users: the Meme Generator archive and the GIPHY archive, populated with content harvested from MemeGenerator.net and GIPHY.com. The LOC has made these collections public as part of an experiment in advancing public user creation of unique “tools, art, applications, and visualizations” with its collection material [8]. The LOC Web Archive is already home to a multitude of digital humanities research projects and early internet website collections that represent the Library’s overall effort to enable more digital scholarship and internet history using its collections. Our study presents one of the first in depth looks at this important digital archive of memes and derived data, which includes descriptive metadata about each of the memes that were collected as part of archiving thousands of pages of the Meme Generator website and its many original memes.

As artifacts of digital folklore, memes are broadly referred to as image macros with text captions [6]. All of the memes archived in the Meme Generator dataset can be described in this way, because they are images with humorous captions or catchphrases digitally superimposed. These images are then widely used, published, shared, and referenced in various instances of online communication. Internet memes can also refer to moving images or illustrated GIFs that can also be embedded. For the purposes of this study, image macros will be referred to simply as “memes,” with their component parts described in later sections.

We begin by introducing some of the current literature in the field of memes in social media research, including the ways in which memes are defined and characterized as unique information artifacts with many sets of derivatives, variants, copies, versions or series. We then describe challenges in meme research, including social media data access and web archives, providing context for the importance of the Meme Generator archive and its significance for social media research.

In the third section we briefly describe the design and methodology of our investigation, including verification, and report on the limits of analysis with the available dataset. We then present two major findings from our analysis the Meme Generator dataset: 1) the existence of unique and canonical “base memes” used in many subsequent derivatives and antecedents within the meme archive, and 2) trends in languages detected across the memes that were archived. These lead to a third finding that results from detecting cross-language variants amongst original base memes, which we call “The Neil deGrasse Tyson Problem”. The paper ends with future questions for research on this data and the broader impact of these findings as they relate to the circulation of memes on internet platforms, methods for describing the contextual layers of memes as distinct information artifacts with antecedents, and the use of web archives in social media research to explore the diversity of language and expression online.

2 LITERATURE REVIEW

2.1 Internet Meme Studies

In 1976, Richard Dawkins published *The Selfish Gene* and used “meme” to describe the idea of a replicator, a unit of imitation, and a way for an idea to spread [9]. Since then, the word has been used to describe the phenomena of internet “memes”— ideas, jokes, satire,

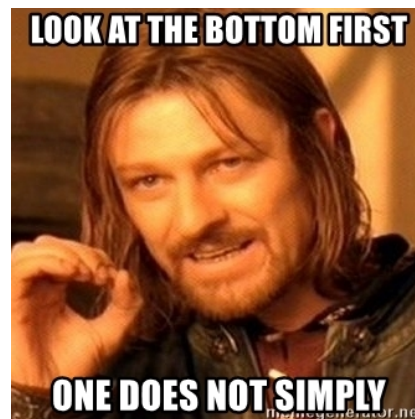


Figure 1: A meme illustrating two lines of text captions.

actions— that spread across subcultures of the internet. As stated above, the sharing of memes began on forums but now is a stable form of expression across the web, and social media platforms in particular. Online forums and internet resources such as Know Your Meme [10] and Reddit’s r/memes subreddit [11] offer popular definitions of memes, and even taxonomies with sub-genres [12]. Resources like Know Your Meme delve into the particular distinctions between different styles of memes and various meme subcultures and subgenres, from the straightforward image macros with overlaid text that are emblematic of the late 2000s and early 2010s, to the more surreal and artistically bizarre memes that are representative of the late 2010s, which derive their humor, in part, from their absurd style, which is sometimes directly compared to that of earlier popular memes.

The r/Memes subreddit defines memes very broadly, speaking to the vast network of artifacts and ideas that can be considered to be ‘memes’, from a video clip to a GIF. According to that subreddit, a meme is: “[a] way of describing cultural information being shared. An element of a culture or system of behavior that may be considered to be passed from one individual to another by non-genetic means, especially imitation” [11]. This emphasis on sharing, passing on, and imitation harkens back to Dawkins’ original definition for the term, which focused on transmission, movement, virality, and propagation. This r/memes subreddit definition includes not just image macro memes, but all forms of internet meme culture. According to this definition, moving images, GIFs, charts, web pages, screenshots, phrases, concepts, comparisons, characters, and events can all be considered ‘memes’ because of the cultural information they signify and the virality with which they spread. Still, there is no standardized set of terms for talking about specific, component parts of image-based memes with text captions. Early popular image-based memes were square images of pop-culture scenes or icons with text overlays on the top and bottom. Today, image-based memes can be multi- or single-paned, and ironic or surreal, effectively playing with how the object is consumed as a text-bearing image (such as Figure 1).

The challenge of defining the component parts of memes may be due to Dawkins’s original emphasis on propagation. Like many other internet artifacts, the sheer dynamism and rapid development

of the ‘social’ web, like social networking sites and mobile apps, remains a challenge for researchers, lawyers, archivists, and technologists. It is not that researchers have failed to describe memes, but rather memes themselves defy stasis and isolation as information artifacts. As Milner theorized, memes are participatory media—they are inherently dynamic and always in dialogue with one another [13]. Thus, most meme research, investigates memes’ spread across platforms as embedded artifacts within layers of social interaction and audience reception, uptake, or reuse because of their mutability and generative qualities.

Most research on the circulation and spread of memes [14, 16, 17] considers them to be unique cultural artifacts, and after identification, performs qualitative analysis on measurable any socio-cultural and political impacts of their reception [13, 15]. Few internet meme studies have compared the circulation of memes across languages. One recent cross-linguistic study [18] uses both quantitative and qualitative analyses to examine “global and local dimensions of mainstream meme culture” by tracing “the top 100 templates in meme generators in English, German, Spanish, and Chinese, using 10 examples to typify each ($N = 4000$)” [18]. Our study extends these findings by using a much larger data pool to identify languages in the meme archive.

Meme publication tools like Meme Generator have not been extensively used in contemporary studies of meme circulation, however a recent study by Dubey et al. experimented with datasets scraped from Meme Generator and Quick Meme to develop an algorithm that “maps image macros to the template image from which it was created, and then decouples the overlaid information from the base template” [17]. The authors write that the study was “the first of its kind in the domain of web content analysis that looks at the virality prediction problem through the lens of Image Macro structure” [17]. Other research on the detection and propagation of memes include Zannettou et al. [14], which studied the “popularity and diversity of meme images” across multiple fringe web communities (/pol/, Gab, and r/The_Donald subreddit), by clustering and then annotating memes using metadata obtained from Know Your Meme.

The complex structure of memes and their contributing components largely impact their uptake, spread, meaning and reception. Given this, describing and defining their component parts is essential to the accuracy and scholarship of their reception and significance, in addition to their identification as primary sources in web archives. Many researchers have relied upon online reference resources such as Know Your Meme to contextualize the meaning, reception, and early appearances of memes. Know Your Meme is an online database that uses wiki-software to provide crowdsourced context and history for popular memes, documenting their origins, spread, and any notable examples. It also features an editorial board that curates its content and provides quality-control for crowdsourced entries. Scholars who study memes’ significance have cited Know Your Meme as being “as close to an ‘authority’ on memes as there is” [14] and “the definitive meme database” [15] for describing and characterizing memes because of the contextual work that platform moderators do as arbiters of meme description and classification.

All internet studies of memes we have surveyed begin with the researcher or research team collecting and creating their own

research dataset. Instead of creating our own bespoke collection of memes, we have relied on a robust, public web archive on which to base our design, as we investigate the unique component parts of memes including the base meme image and unique lines of texts present in captions.

2.2 Web Archives for Social Media Research

Internet researchers face a number of challenges when researching digital media objects, particularly items that circulate on the web and across social media platforms. Digital media types such as photos, GIFs, geolocation check-ins, or embedded videos have two, interdependent layers of context that need to be archived if they are selected for long-term preservation. First, all digital media objects contain software code that formats them and allow them to be rendered, published, indexed and accessed on the web by machines. Second, digital objects are made up of semantic elements that people consume and make meaning of—sound, images, and texts that humans’ access, read, and share. So, web archives must preserve as much technical and semantic information from these two layers as possible in order to provide access to reliable and authentic sources. An additional challenge for researchers and archivists concerned with web-based digital media is the inherent and well-documented ephemerality [19, 20], due to their circulation, persistence, and cultural reception. Thus, web archives have become important sources for bootstrapping, enabling, and identifying the numerous access challenges that researchers face with information artifacts that are web-based [21].

Web archives are any form of deliberate and purposeful preservation of web material [22]. They can range from digital video collections, to image-sharing websites, to preserving whole platforms that have been shuttered, such as “This is my jam” or the Vine platform [23]. Brügger has characterized two main approaches to web archiving—the “macro” and the “micro” [22]. Macro web archiving approaches are typically carried out by large information institutions, such as museums, research universities, or cultural heritage organizations like national libraries or the Internet Archive, and usually rely on comprehensive web crawling techniques. Such macro web archiving efforts are concerned with capturing whole swaths of the web that can be largely representative; these efforts have reached broad audiences as an online resource. Micro web archiving efforts are instead undertaken by individuals or small groups that want to preserve slices of the web for more individualized or specific intentions, such as documenting the particular hashtag of social movement or short-term event.

Whether using a macro or a micro approach, web historians and digital archivists typically preserve the web using one or several in combination of the following methods: static screen shots; screen casts that capture video and/or audio components as well as interactive dynamism of the user interface like scrolling or clicking links; or web crawling web pages. Web crawling is the most time consuming and resource intensive because it aims to capture whole web pages and online resources by “crawling” each and every embedded hyperlink, capturing all components of a web resource [21].

While web archives of internet culture have existed for several decades, archives of social media or social media data extracted from platforms are severely inhibited because of the dynamic nature

of social media posts and the use of Application Programming Interfaces (APIs) for data extraction [24, 25]. Web archives of social media face some specific challenges that many archival scholars have investigated. First, social media is ephemeral and constantly streaming [19, 26]. Second, it is increasingly hard for researchers to access comprehensive data through APIs given new data privacy regulations, definitions of personally identifiable information (PII), and platforms' rollbacks on their terms of service [25]. Finally, embedded social media objects, such as memes or GIFs, that are framed within larger social media environments (for example, news feeds), or have been linked or re-posted from another source, pose a number of technical challenges for harvesting unbroken hyperlinks [19].

The few web archives of social media that do exist are not without their drawbacks [27, 28]. The LOC Meme Generator dataset archive is a unique example because the Library was permitted to make web crawls of the live website. Comprehensive web crawling can slow down live webpages and hosts will commonly discourage such comprehensive archiving efforts by blocking crawl requests [22]. Even though web archives can create new possibilities for researchers to investigate social media, there remain many questions that are unanswerable even with the datasets resulting from macro web archiving techniques. For example, web crawls can fail to capture all the elements on a webpage, or important layers of context that give meaning to the information intended to be preserved [29]. Often the data is too complex for the archiving group to even provide access to researchers, whether because of technical constraints or due to human subjects research ethics concerns. The Twitter public archive was touted for many years as an exciting and important public history project, but was widely panned as a failure after the LOC announced that it could not provide access to the data "gifted" from the platform [30]. These archiving and access challenges impact the historiography of social media, but they also impact the way that researchers can reproduce studies and ensure reliability of results [31]. So, where an institution with a mandate of providing public access (here, the LOC), creates a macro web archive using web crawling (here, the Meme Generator dataset), some of the challenges researchers typically may face are diminished. For our purposes, the Meme Generator dataset presents a generally reliable, preserved collection.

3 METHODS

For our study, we used the available Meme Generator dataset, which provided readily available data and descriptive metadata. Despite LOC's robust catalog of archived web pages and the institution's desire for researchers to interact with the data, we have not found any related research literature utilizing the LOC's Web Cultures Web Archive Meme Generator dataset for scholarly analysis. Thus, we believe we are the first research team to use the Meme Generator dataset for empirical data analysis. Furthermore, while memes are usually studied in context with other posts or as works that accrete layers of new meaning as they circulate amongst audiences, the LOC Web Archives collects memes together as a discrete collection of memes alone, allowing researchers to compare types and generate a taxonomy of classes amongst components from a popular meme creation website. The Meme Generator dataset web archive allows

us to examine trends in popular memes and derivative works as well as the cross-linguistic components of memes' captions by detecting the presence of different language instances within the dataset and comparing them to older or newer, derivative instances. In this section we describe how we accessed the dataset and identified its properties, how we verified the data, and the techniques we used for clustering and language detection.

The Meme Generator web archive and public dataset can be accessed by visiting the LOC Labs experiments page [8]. Two items can be downloaded for access—a 33.9 MB CSV File with a dataset of 57,652 unique memes instances, and a README.TXT File generated by the LOC Web Archiving Team that describes the contents, relevant definitions, and information about how the dataset was created [8]. The dataset was generated from a web crawl of the Meme Generator website in 2018 [6]. While the meme images themselves are not included in the data, the dataset does include links for accessing archived copies of each meme within the LOC's larger Web Cultures Web Archive. Each meme instance is represented by a row with seven columns, indicating up to seven data values for each instance. In addition to descriptive fields like Base Meme Name and Alternate Text (Fig. 2), each instance contains a unique Meme ID, an Archived URL, a Meme Page URL, an MD5 Hash for verification, and File Size in bytes. The catalog information and descriptive metadata of the dataset included in the README file defines a number of terms concerning memes and their components, in a section called "Anatomy of a Meme" [32].

When a user creates a meme in Meme Generator, they upload an image, and then enter captions in two separate text fields: one for the text on the top of the image and one for the text on the bottom. The original 2018 web crawl captured the text as a single field, and populated it in a field they called Alternate Text (or "Alt Text"). Generally, Alt text is an HTML attribute that provides a textual alternative to non-text content in web pages, allowing the content and function of the image to be accessible to those with visual or cognitive disabilities who rely on screen readers to relay visual information. The LOC Web Archiving team has since updated the Meme Generator dataset to capture the top and bottom lines as two separate text fields. Our analysis is based on the old "Alternative Text" data field which is single line of text from both the bottom and top of a Base Meme Image. Once we shared our initial results with the LOC Web Archiving Team, they responded to our recommendations by undertaking a new web crawl that captured top text and bottom text in separate metadata fields. The most current version of the Meme Generator dataset, which was updated May 17, 2019, collects the top text and bottom text as two separate fields, referred to as "Upper Text" and "Lower Text" in the updated README file of the archive's dataset [33].

After accessing the dataset, we began by confirming records and scanning for errors or missing data fields. The dataset includes 86,310 Base Meme Images that are represented in 57,652 unique Meme Instances. Initially when verifying the dataset, we noticed some errors such as duplicates of Meme Instances ($N=2,552$) and empty fields ($N = 266$). The sum of these errors represented less than 5% of records in the total records in the dataset. Errors were attributed to: issues with the initial web crawl, server complications from Meme Generator and instances where users created duplicates with templates of original Meme Instances [34]. Additionally, there


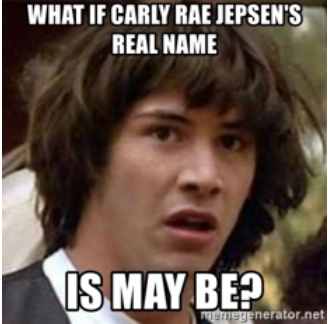
Value Field	Description	Format	Examples
Base Meme Image	Base Meme Images are the starting point for meme generation. They are the image without text applied.	JPEG	
Base Meme Name	Each Base Meme Image has a particular name associated with it, called the Base Meme Name.	text	Conspiracy Keanu
Alternate Text	User supplied text that is placed on top of the Base Meme Image.	text	If the Base Meme Image is Conspiracy Keanu the Alternate Text might be: "What If Carly Rae Jepsen's Real Name Is May Be?"
Meme Instance	The resulting combination of a Base Meme Image and user supplied Alternate Text.	JPEG	

Figure 2: Data value fields: definitions and examples for a single meme instance.

are some instances in which the initial web crawl captured up to five copies of the same meme instance. In these cases, the Meme ID, Archived URL, and Meme Page URL were unique, but the MD5 Hash, File Size, and Alternate Text were identical.

After confirming and verifying all the data provided by the web archive, we measured and calculated original Base Meme Names across the dataset. Using data analysis software OpenRefine we indexed and clustered the Base Meme Names with text facets to identify and calculate the most popular. OpenRefine uses Key Collision and Nearest Neighbor methods to reconcile similar text names [35, 36]. Clustering methods were ordered from strict to lax. By clustering the text data from the Base Meme Name data value field, we were able to narrow the total number of meme images represented in the dataset into further sets of more precise "Canonical Base Memes". We were then able to index and measure captions from the Alt Text field and identify the many different languages.

To increase confidence in OpenRefine's clustering recommendations, we cross-referenced the clustering results of varied but similar Base Meme Names with their entries in Know Your Meme [10]. Clustering results were also subject to manual analysis when textual data did not feature semantic closeness. For example, the

Base Meme Names "300," "Sparta," and "This is Sparta" are all variants of the "This is Sparta!" Base Meme Name [37]. Clustering recommended "Sparta" and "This is Sparta" instances to be clustered as the same image but not the numeric term "300". But with manual analysis of individual records in the dataset and cross-referencing sources with Know Your Meme entries confirmed that "300" should also be included in the "This is Sparta" cluster as a Canonical Base Meme class.

During our initial verification of the dataset we identified that there were many different languages represented in some of the data fields, specifically the Base Meme Name and more importantly, the Alt Text field that includes user-generated captions. Using Google Sheets, we uploaded the dataset and used the DETECTLANGUAGE function to identify the total number of different languages present in the dataset [38]. Leveraging the Google Translate API, the Sheets DETECTLANGUAGE function uses statistical machine translation and neural machine translation to identify language [39]. The function identified 89 languages in the dataset, where five languages (English, Portuguese, Russian, Spanish, and Ukrainian) make up 93% of the meme instances in the entire dataset. Based on the huge variety of languages present, the dataset has a wealth of opportunity for potentially meaningful analyses.

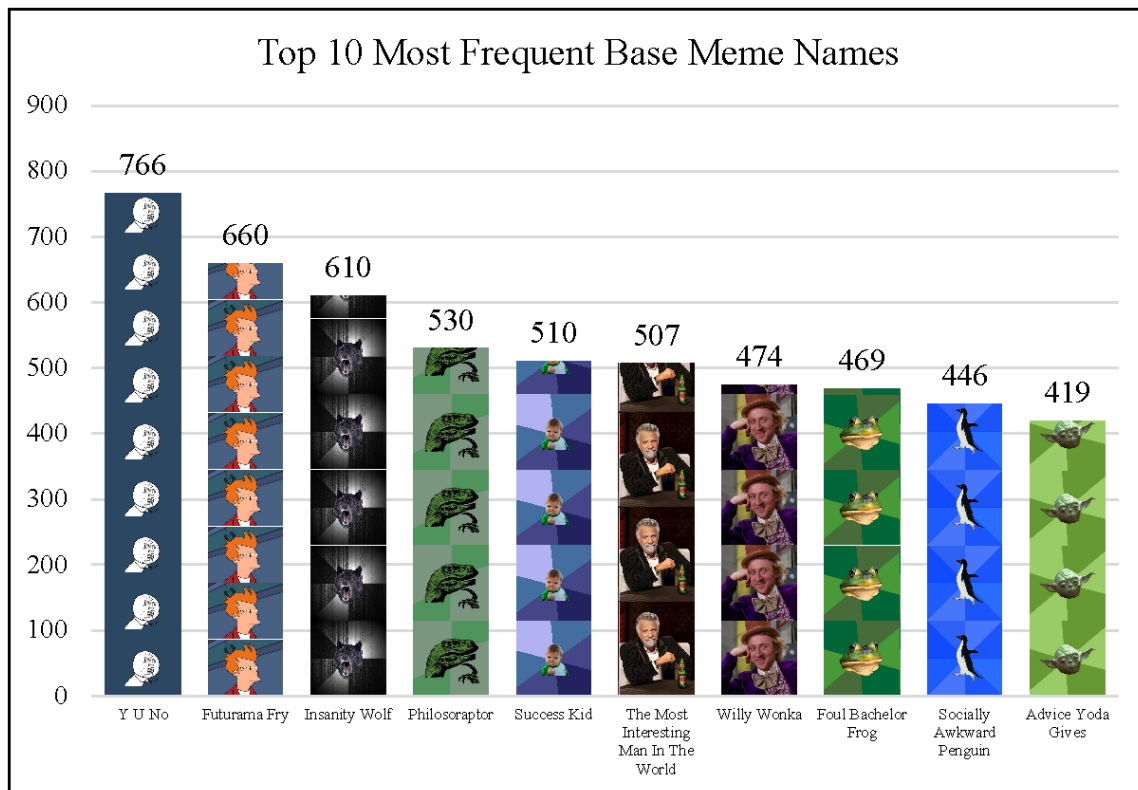


Figure 3: Top 10 Most Frequent Base Meme Names. Bar chart comparing the most frequently used Base Meme Images in the dataset, based on the data in the Base Meme Name field prior to applying clustering in OpenRefine. In order from most to least, they are: Y U No (766 instances), Futurama Fry (660), Insanity Wolf (610), Philosoraptor (530), Success Kid (510), The Most Interesting Man in the World (507), Willy Wonka (474), Foul Bachelor Frog (469), Socially Awkward Penguin (466), and Advice Yoda Gives (419).

4 RESULTS AND DISCUSSION

4.1 Clustering to Identify Canonical Base Memes

Indexing and sorting the data by unique Base Meme Names revealed that there are 1,913 unique Base Meme Names within the 57,652 total Meme Instances within the dataset. Earlier analysis from the LOC found the top 10 most frequently appearing in the dataset [6]. These can be seen in Figure 3

When reviewing the 1,913 Base Meme Names for trends, we saw opportunities for clustering similar names and merging many of the Base Meme Name entries for a new list of most frequently used names. For example, merging opportunities using Key Collision and Nearest Neighbor would recommend merging entries title “Joseph Ducreux” (414 meme instances) and “Joseph Ducreaux” (91 meme instances) into a new cluster value of 505 instances. In order to distinguish between the original, user-generated base meme name used in a meme generator template and more broadly used or “official” names according to online community resources, we developed a new term to identify and group identical image macros: the “Canonical Base Meme Name”. Drawing from official name entries according to Know Your Meme, Canonical Base Meme Names can

be applied to the new values of clustered sets found throughout the dataset.

The Canonical Base Meme Name is typically more recognizable to people than the original title used when the meme was generated in a template. For example, many internet users would know the canonical base meme name “One Does Not Simply Walk into Mordor” refers to the meme image from the 2001 *Lord of the Rings* film depicting the character Boromir with the phrasal template “One Does Not Simply. . .” [40]. This Canonical Base Meme Name is distinct from the original user-generated titles for memes in the dataset, which can include redundancies (e.g., “Lord Of The Rings Boromir One Does Not Simply Mordor” versus “Does not simply walk into mordor Boromir”), less obvious redundancies (e.g., “300” versus “This is Sparta!”), and formatting and spelling discrepancies (e.g., “Bad Luck Brian” versus “:badluckbrian:”). Clustering Base Meme Names created by users and merging them into similar groups of Canonical Base Meme variants defined by online communities affords us more precision in identifying variants and antecedents of popular memes. This method also provides new information for understanding more about frequently used Base Memes Names and Base Meme Images across the dataset. For example, by clustering “Joseph Ducreux” with “Joseph Ducreaux,”

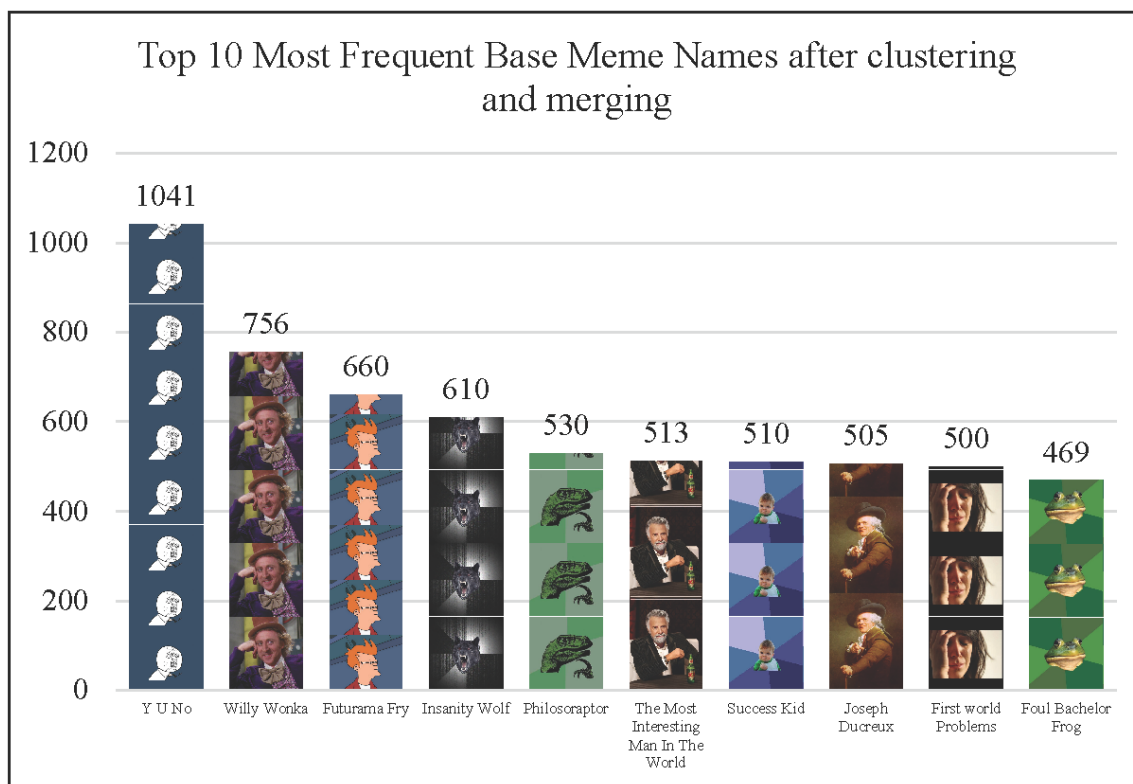


Figure 4: Top 10 Most Frequent Base Meme Names after clustering and merging. Bar chart comparing the most frequently used Base Meme Images in the dataset, based on the data in the Base Meme Name field after applying clustering in OpenRefine. In order from most to least, they are: Y U No (1041 instances), Willy Wonka (756), Futurama Fry (660), Insanity Wolf (610), Philosoraptor (530), The Most Interesting Man in the World (513), Success Kid (510), Joseph Ducreux (505), First world Problems (500), and Foul Bachelor Frog (469).

and merging the two sets to become the broader Canonical Base Meme Name, the “Joseph Ducreux” cluster goes from being the 11th most frequently used Base Meme in earlier analysis to the 8th most frequently used across the dataset (Fig. 4).

Through clustering, we were able to identify and merge the Base Meme Names from 1,913 to 1,514, effectively establishing the “top ten.” This effort was sufficient for our purposes, but future research may investigate further. This clustering process revealed that in many cases the distinguishing factors between base meme names and its corresponding base meme image are more complex than different capitalization or misspellings (e.g., “1st World Problems” versus “First World Problems”). As memes continue to be studied by social media researchers and internet historians, and the body of memes continues to grow, the distinction between the Base Meme Name and the Canonical Base Meme Name will become increasingly important for tracking the origins, popularity, mutation, and propagation of meme instances and their related variants.

4.2 Detecting Languages in Captions

After clustering to measure the most used Base Meme Names, we leveraged the Alt Text data field to detect the various languages represented in captions from the Meme Generator dataset. These

data can be used to find whether certain base memes are more frequently used in some languages than others. We isolated the text data and detected language to discover patterns related to languages, like the most popular base meme images used with Russian captions versus those with English captions. While 89 unique languages were identified, Spanish, English, Russian, and Portuguese, in that order, make up 91.8% of the dataset (Table 1). By determining which languages appear in the Meme Generator dataset, quantifying and ranking the different languages according to frequency, and then relating that to the base memes, we can draw meaningful connections between the frequency of images used across different languages.

While we were confident that the language detection algorithm identified meme instances from the most used languages, there were some flaws with detecting articulatory phonetic expressions, internet vernacular, slang, and memes that are purposefully misspelled. Because the DETECTLANGUAGE tool is based on translation functions, we can expect poor accuracy detection with some kinds of words. For example, meme instance #17271228 features lyrics from the theme song to Adam West’s Batman television show. The Alt Text field contains, “NA NA NANANA NA NA NANA BATMAN, BATMAN!” The tool identified this text as coming from the Igbo

Table 1: Top four languages represented in the Meme Generator dataset.

Language	Number of Meme Instances	Percentage
Spanish	19,656	34%
English	19,142	33%
Russian	13,572	23.5%
Portuguese	781	1.3%

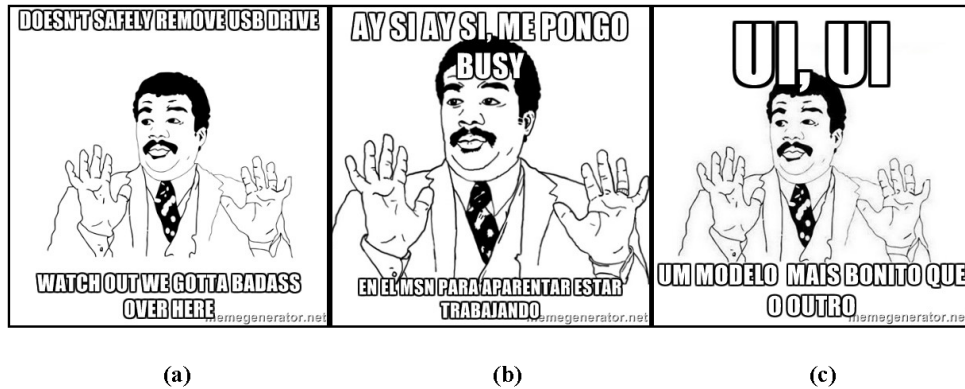


Figure 5: Sample of three meme instances that are categorized according to different Base Meme Images, (a) We Got Badass Over Here, (b) AY SI, and (c) UI, despite sharing the same Base Meme Image.

language, the native language spoken by the Igbo people of south-eastern Nigeria. When translated in Igbo, the phrase is “I love my Batman!” The detection algorithm had difficulty with detecting vernacular phrases and gibberish such as strings of characters “asd-fasdf,” which are letters placed next to each other on the QWERTY keyboard. While the tool tended to misidentify outliers like “NA NA NANANA”, the unexpected language recommendations it came up with did help guide us in quickly recognizing meme instances with phonetic expressions, slang, or internet vernacular such as “doge” instead of “dog”. By running natural language detection tools on the text captions data, we pursued an entirely new way of working with this multilingual dataset, affording us the opportunity to pursue further research questions in the future. We acknowledge the many imitations to automated translation and language detection that have been well documented, and suggest future thorough analysis of the 8% of memes where 85 languages were detected, in order to confirm the automated detection recommendations.

4.3 The Neil deGrasse Tyson Problem

As previously discussed, when clustering the Base Meme Names, we referred to Know Your Meme as the authority on meme provenance for instances of multiple Base Meme Names being used for identical Base Meme Images. In order to distinguish between the original, user-generated title, and the more broadly used, official name provided by Know Your Meme we used the term “Canonical Base Meme Name”. After identifying the issue of Canonical Base Meme Names, we wanted to verify and confirm that the Base Meme Images matched their names. Through our investigation, we discovered that while many records in the dataset shared the same Base Meme Image (the JPEG), they were in fact categorized

by different Base Meme Names in those cases where the textual captions were in different languages. This classification dilemma can be illustrated with what we came to call the “Neil deGrasse Tyson Problem”.

Through the clustering process we discovered several Base Meme Names referred to the same Base Meme Image of an outline drawing of famed astrophysicist Neil deGrasse Tyson raising both hands in the air emphatically. According to Know Your Meme, this Base Meme Image correlates to the Base Meme Name “Neil deGrasse Tyson Reaction,” which uses the phrasal template, “watch out, we got a badass over here”. Analysis of the Meme Generator dataset and Know Your Meme revealed that there were several other commonly used Base Meme Names for this same Base Meme Image, all of which had different contextual meanings across different languages. One common Base Meme Name was “Ay si ay si,” which is Spanish for “Oh yes oh yes” (Fig. 5b). In the Portuguese language, “Ui, Ui” (roughly “Oh wow, oh wow” in English) was frequently used (Fig. 5c).

Representation and identification dilemmas arise where there are antecedents and variants with different languages in the captions of a Base Meme Image in a meme corpus. The predicament in identifying this information artifact is whether to cluster and merge all Base Meme Names and Base Meme Images of Neil deGrasse Tyson under one Canonical Base Meme category. In other words, can a single Base Meme Image have multiple Base Meme Names, canonical or otherwise, if the meanings of captions, especially across different languages, are substantially different? This situation presents interesting questions about the idea of a collective meme canon, meme culture at large, and identification issues that cataloging librarians and archivists have always grappled with

when classifying and creating descriptive metadata information about unique information artifacts that may have many derivatives, copies, or versions. Drawing on research from Dubey et al. [17], which “propos[es] an algorithm based on sparse representations and deep learning to decouple various types of content in such images and produce a rich semantic embedding,” we can consider the importance of maintaining the semantic content when mapping meme instances to their original meme image template. Despite the sameness of the image macro or even identical user supplied titles, the semantic (and contextual) meanings from language to language may take on many different tones or intentions as the meme circulates online.

Through the establishment of a meme canon resource and corresponding “canonical” Base Meme Images, consideration of the origins and language variants in the Neil deGrasse Tyson image example could be more accurately cataloged in datasets from web archives like this one. Online resources like Know Your Meme feature contributors who attempt to describe memes and provide context into their development and reception. Internet researchers who investigate memes can use a number of analytical tools to study the origins and spread of memes across platforms. However, this derivative data shows that as users create and share memes from template generators, some original descriptive metadata such as user-generated names can be lost. The Canonical Base Meme Name and the Neil deGrasse Tyson Problem give researchers and meme archivists a framework for reconciling the messy reality of variant and duplicate memes, with inconsistent user-generated metadata and a multitude of languages. In the coming decades, as memes continue to be studied and catalogued as cultural artifacts, new terminology and descriptive concepts (as we have presented) will be needed to illuminate the complexities encountered when trying to historicize the ontogenic structure of internet meme culture.

5 CONCLUSIONS

In this study we used descriptive analysis, clustering and reconciliation methods, and language detection to investigate the major features of the Meme Generator dataset, including its format, structure, variants and trends as a comprehensive web archive of memes. To our knowledge, this is the first published research based on the Meme Generator dataset despite receiving much fanfare in mainstream technology coverage when the release of the archive was announced in 2018 [41] and in present-day coverage [42].

By using automated tools to parse the dataset and sort through its contents, we were able to identify a number of unique characteristics of the data that were previously unknown. For example, more accurate calculations of the top ten most popular Base Meme Names and identifying multiple languages represented in the web archive. The “Neil DeGrasse Tyson Problem” we identified is an example of the unique mutability of memes as a form of expression. As memes are generated by online communities, the same visual images can be re-appropriated for different uses, and may come to mean different things as they are used in different contexts and in different languages.

Social media scholars must take advantage of public web archives and automated tools, as well as online resources like Know Your Meme, in order to understand the varied contexts between differing

instances of memes and ultimately to keep track of the origins and meanings of Base Meme Images. This web archive from the LOC represents a broad, contemporary, and cultural expression from the internet that has emerged over the past two decades, and importantly, it is not limited to English-speaking online cultures or even to digital folklife from the United States. Based on our language detection findings, we argue that this web archive is an important resource for social media researchers to study a diversity of voices across linguistic boundaries, differences in humor, and a variety of cultural expressions present in memes from template generators as well as online cultures throughout the world.

ACKNOWLEDGMENTS

Here we describe the authors’ contributions to this article using the standard role taxonomy from “Publishing: Credit where credit is due,” published in *Nature*, 2014;508[7496]:312–313). J.S. and A.C.L. provided the study’s conception, performed the experiments, and wrote initial drafts. J.S. was responsible for the majority of visualizations and data presentation. A.C.L. was responsible for the majority of secondary source research, formatting and generating references. A.A. supervised the formal analysis, the core formulation of the research goals, and led the manuscript preparation and revisions.

REFERENCES

- [1] “About this Collection,” *Library of Congress*. [Online]. Available: <https://www.loc.gov/collections/web-cultures-web-archive/about-this-collection/>. [Accessed: 09-Jan-2020].
- [2] “Urban Dictionary: Define Your World,” *Library of Congress, Washington, D.C. 20540 USA*. [Online]. Available: <https://www.loc.gov/item/lcwaN0004130/>. [Accessed: 20-Jan-2020].
- [3] “Emojipedia - Emoji Meanings,” *Library of Congress, Washington, D.C. 20540 USA*. [Online]. Available: <https://www.loc.gov/item/lcwaN0010500/>. [Accessed: 20-Jan-2020].
- [4] “Internet Meme Database | Know Your Meme,” *Library of Congress, Washington, D.C. 20540 USA*. [Online]. Available: <https://www.loc.gov/item/lcwaN0009692/>. [Accessed: 16-Jan-2020].
- [5] “Meme Generator,” *Library of Congress, Washington, D.C. 20540 USA*. [Online]. Available: <https://www.loc.gov/item/lcwaN0010226/>. [Accessed: 09-Jan-2020].
- [6] T. Owens, “Data Mining Memes in the Digital Culture Web Archive,” *The Signal*, 11-Oct-2018. [Online]. Available: <https://blogs.loc.gov/thesignal/2018/10/data-mining-memes-in-the-digital-culture-web-archive/>. [Accessed: 09-Jan-2020].
- [7] “Frequently Asked Questions,” *Library of Congress*. [Online]. Available: <https://www.loc.gov/programs/web-archiving/about-this-program/frequently-asked-questions/>. [Accessed: 20-Jan-2020].
- [8] “Web Archive Datasets | Experiments | Welcome to Labs! | Library of Congress,” *Library of Congress, Washington, D.C. 20540 USA*. [Online]. Available: <https://labs.loc.gov/experiments/webarchive-datasets/>. [Accessed: 09-Jan-2020].
- [9] R. Dawkins, *The selfish gene*. New York: Oxford University Press, 1976.
- [10] “Know Your Meme,” *Know Your Meme*. [Online]. Available: <https://knowyourmeme.com/>. [Accessed: 20-Jan-2020].
- [11] “r/Memes the original since 2008,” *Reddit*. [Online]. Available: <https://www.reddit.com/r/memes/>. [Accessed: 24-Jan-2020].
- [12] “Memes Then, Memes Now | Know Your Meme,” *Know Your Meme*. [Online]. Available: <https://knowyourmeme.com/memes/memes-then-memes-now>. [Accessed: 24-Jan-2020].
- [13] R. M. Milner, *The World Made Meme: Public Conversations and Participatory Media*. Cambridge, Massachusetts: The MIT Press, 2016.
- [14] S. Zannettou et al., “On the Origins of Memes by Means of Fringe Web Communities,” in *Proceedings of the Internet Measurement Conference 2018*, Boston, MA, USA, 2018, pp. 188–202, doi: 10.1145/3278532.3278550.
- [15] L. Nooney and L. Portwood-Stacer, “One Does Not Simply: An Introduction to the Special Issue on Internet Memes,” *J. Vis. Cult.*, vol. 13, no. 3, pp. 248–252, Dec. 2014, doi: 10.1177/1470412914551351.
- [16] A. Dubey and S. Agarwal, “Modeling Image Virality with Pairwise Spatial Transformer Networks,” *ArXiv170907914 Cs*, Sep. 2017.
- [17] A. Dubey, E. Moro, M. Cebrian, and I. Rahwan, “MemeSequencer: Sparse Matching for Embedding Image Macros,” in *Proceedings of the 2018 World Wide Web Conference*, Republic and Canton of Geneva, Switzerland, 2018, pp. 1225–1235.

- doi: 10.1145/3178876.3186021.
- [18] A. Nissenbaum and L. Shifman, “Meme Templates as Expressive Repertoires in a Globalizing World: A Cross-Linguistic Study,” *J. Comput.-Mediat. Commun.*, vol. 23, no. 5, pp. 294–310, Sep. 2018, doi: 10.1093/jcmc/zmy016.
- [19] S. Walker, “The Complexity of Collecting Digital and Social Media Data in Ephemeral Contexts,” Thesis, 2017.
- [20] K. Kinder-Kurlanda, K. Weller, W. Zenk-Möltgen, J. Pfeffer, and F. Morstatter, “Archiving information from geotagged tweets to promote reproducibility and comparability in social media research,” *Big Data Soc.*, vol. 4, no. 2, p. 2053951717736336, Dec. 2017, doi: 10.1177/2053951717736336.
- [21] I. Milligan, “Lost in the Infinite Archive: The Promise and Pitfalls of Web Archives,” *Int. J. Humanit. Arts Comput.*, vol. 10, no. 1, pp. 78–94, Mar. 2016, doi: 10.3366/ijhac.2016.0161.
- [22] N. Brügger, “Web history and social media,” *Sage Handb. Soc. Media*, pp. 196–212, 2018.
- [23] K. Rogers, “Vine Is Closing Down, and the Internet Can’t Stand It,” *The New York Times*, 27-Oct-2016.
- [24] D. M. Boyd and N. B. Ellison, “Social network sites: Definition, history, and scholarship,” *J. Comput.-Mediat. Commun.*, vol. 13, no. 1, pp. 210–230, 2008.
- [25] A. Acker and A. Kriesberg, “Social media data archives in an API-driven world,” *Arch. Sci.*, Sep. 2019, doi: 10.1007/s10502-019-09325-9.
- [26] A. Acker and J. Donovan, “Data craft: a theory/methods package for critical internet studies,” *Inf. Commun. Soc.*, vol. 0, no. 0, pp. 1–20, Jul. 2019, doi: 10.1080/1369118X.2019.1645194.
- [27] A. Acker and A. Kriesberg, “Tweets may be archived: Civic engagement, digital preservation and obama white house social media data,” *Proc. Assoc. Inf. Sci. Technol.*, vol. 54, no. 1, pp. 1–9, Jan. 2017, doi: 10.1002/ptra2.2017.14505401001.
- [28] L. Hemphill, S. H. Leonard, and M. Hedstrom, “Developing a Social Media Archive at ICPSR,” in *Proceedings of Web Archiving and Digital Libraries (WADL’18)*, 2018.
- [29] A. Bruns, “Faster than the speed of print: Reconciling ‘big data’ social media analysis and academic scholarship,” *First Monday*, vol. 18, no. 10, Oct. 2013, doi: 10.5210/fm.v18i10.4879.
- [30] A. Bruns, “After the ‘APicalypse’: social media platforms and their fight against critical scholarly research,” *Inf. Commun. Soc.*, vol. 0, no. 0, pp. 1–23, Jul. 2019, doi: 10.1080/1369118X.2019.1637447.
- [31] D. Freelon, “Computational Research in the Post-API Age,” *Polit. Commun.*, vol. 35, no. 4, pp. 665–668, Oct. 2018, doi: 10.1080/10584609.2018.1477506.
- [32] C. Dooley, “[Data file] README.TXT - Anatomy of a Meme”. Library of Congress, 2018.
- [33] C. Dooley, “[Data file] README.TXT - Anatomy of a Meme”. Library of Congress, 2019.
- [34] C. Dooley, A. Grotkie, and G. Thomas, “Personal communication,” 03-May-2019.
- [35] O. Stephens, “Clustering In Depth: Methods and theory behind the clustering functionality in OpenRefine,” *GitHub*, 31-May-2019. [Online]. Available: <https://github.com/OpenRefine/OpenRefine>. [Accessed: 25-May-2019].
- [36] A. Delpeuch, “OpenRefine Reconciliation,” *GitHub*, 15-Jan-2020. [Online]. Available: <https://github.com/OpenRefine/OpenRefine>. [Accessed: 25-Jan-2020].
- [37] “This Is Sparta!,” *Know Your Meme*. [Online]. Available: <https://knowyourmeme.com/memes/this-is-sparta>. [Accessed: 20-Jan-2020].
- [38] Google Doc Editors, “DETECTLANGUAGE - Documentation”. [Online]. Available: <https://support.google.com/docs/answer/3093278?hl=en>. [Accessed: 25-Jan-2020].
- [39] B. Turovsky, “Found in translation: More accurate, fluent sentences in Google Translate,” *Google*, 15-Nov-2016. [Online]. Available: <https://blog.google/products/translate/translation-more-accurate-fluent-sentences-google-translate/>. [Accessed: 25-Jan-2020].
- [40] “One Does Not Simply Walk into Mordor | Know Your Meme,” *Know Your Meme*. [Online]. Available: <https://knowyourmeme.com/memes/one-does-not-simply-walk-into-mordor>. [Accessed: 24-Jan-2020].
- [41] “Why the Library of Congress Thinks Your Favorite Meme Is Worth Preserving | Smart News | Smithsonian Magazine”. [Online]. Available: <https://www.smithsonianmag.com/smart-news/library-of-congress-meme-preserve-180963705/>. [Accessed: 24-Jan-2020].
- [42] S. Kurutz, “Meet Your Meme Lords: A small team at the Library of Congress is archiving internet culture as fast as it can (now, from home).” [Online]. Available: <https://www.nytimes.com/2020/04/07/style/internet-archive-library-congress.html>. [Accessed: 7-Apr-2020].