Self-Censoring on Social Media Sites

Talia A. Feshbach

Bryn Mawr College

Abstract:

Many users of social media sites self-censor by using character replacements or euphemisms to obscure the language they are using. The reasons for doing so vary from trying to avoid other users finding their posts in searches to evading censorship from site administration to adapting to site culture. This paper seeks to answer whether or not there is an association between reasons for self-censoring and tactics used for self-censoring. It also examines three sites where self-censoring is in different forms and amounts - Tumblr, Twitter, and TikTok - and how the tactics and reasons for self-censoring appear on those sites. To do so, I first investigated the rules and communities of the social media sites, taboos, terms and taxonomies for self-censoring practices, and the use of self-censoring in online communication and memes. Next, I conducted a survey and used that survey data to perform a chi-square test using reasons for self-censoring and tactics used for self-censoring as variables. I also conducted percentage observations on the data using the sites and tactics as variables, and then the sites and reasonings as variables. While my results for the chi-squared test were flawed, I found no correlation between self-censoring tactics and reasons for self-censoring. However, I did observe different patterns of tactics used and reasons for self-censoring on the different sites.

1 Introduction

If you have been using the internet in the past few years, you may have come across a new word - ‘unalive.’ Commonly used in captions on TikTok, a video-hosting platform for short videos, this word functions as a euphemism for words like ‘kill’ or ‘suicide.’ This is a form of self-censoring, where a person censors themselves or uses a euphemism for a word. Online, this can take multiple forms, including new euphemisms like ‘unalive’, existing euphemisms like ‘screwing’ for ‘sex’, replacing letters with asterisks, symbols or numbers like ‘k!ll’ for ‘kill’, or using words or phrases that sound similar like ‘slip and slide’ for ‘suicide.’ These forms of self-censoring, referred to as tactics, exist on various different social medias.

The reasons that people self-censor are as numerous as the tactics they use. Some people do not want some of their posts to be found in searches by other users. Conversely, others don’t want their content to be censored by site administration or shadowbanned - having their content hidden in searches. Others simply want to belong in their community and participate in site culture by adopting the terms and practices of the people around them.

Self-censoring occurs on many social media sites, but for this paper I focused on three websites - Tumblr, TikTok, and Tumblr. These websites all have different levels of moderation and different communities. As such, it is possible that the tactics used for self-censoring and the reasons for self-censoring vary between the sites.

These three factors - tactics, reasonings, and sites - are the main variables for this paper. However, I am most interested in whether or not the function of self-censoring is linked to the form. As such, my main research question is ‘Is there an association between self-censoring reasonings and self-censoring tactics?’ I also have two auxiliary questions: ‘What self censoring tactics are used most on which sites?’ and ‘What are the most common stated reasons for self-censorship on each site?’

To answer these questions, I conducted an internet survey to gather data about these three variables, as well as related minor variables and demographic data. I then performed a chi-square test - a statistical test for calculating correlation - on the data for my main question, and percentage observations on the data for my auxiliary questions. In doing so, I found that while different sites do display different patterns of tactics used and reasons for self-censoring, there is no correlation between self-censoring tactics and reasons for self-censoring. As such, the answer to my main research question is ‘There is no association between self-censoring reasonings and self-censoring tactics.’

In section 2, Motivation, I explain why I chose this topic and what questions guided me there. In section 3, Background, I detail some background information and existing research on this topic. Specifically, 3.1 contains information on the terms of service, algorithms, and site cultures of TikTok, Tumblr, and Twitter, 3.2 dives into taboos and their connection to self-censoring, 3.3 defines the terms, taxonomies and typologies I researched or created, and 3.4 details the usage of self-censoring in communication and memes. Section 4, Methodology, outlines the research methods I used and data analysis that is performed in section 5, Research. This section includes the data and preliminary results for each of my three questions. Section 6, Analysis, Limitations and Future Work examines what the results mean, how they are limited, what could have been done better, and what possible work remains. Finally, in section 7, Conclusion, I summarize my work so far and what I have learned.

I should note that this paper includes discussion of the terms used to describe suicide, as well as other sensitive language like curse words. Reader discretion is advised.

2 Motivation

I was initially drawn to this general topic due to a love of internet linguistics, and the ways that language shifts and changes based on online influences. In particular, Gretchen McCulloch’s book *Because Internet: Understanding the New Rules of Language* influenced me strongly. Other influences included observing language use on my primary social media, Tumblr, and how this usage differed from the language usage on other social media I would catch glimpses of in screenshots and links. Some of these posts would comment on how the language use was different. Many of these posts would be complaints or warnings - Tumblr users noting that due to how the filtering system works, censoring sensitive topics in the way users do on other sites would be rude. Others noted that such self-censoring makes reading posts difficult for screen-readers, which cannot correctly parse the text. I first planned to figure out if I could create a program to allow a screen reader to correctly parse a text with symbol replacements, but I was soon drawn down another avenue.

I had begun to notice a trend - Tumblr posts rarely had censored words, and when they did, they were often the names of people or media, or they were for comedic effect. Some tweets, however, had words for sensitive topics or names censored out with asterisks or symbols.

Conversely, the captions of some TikToks used symbol replacements for sensitive words, but many also used strange misspellings or euphemisms like ‘seggs’ or ‘unalive.’ I began to wonder if my observations were indicative of a larger trend for each of these platforms. If so, why do people censor different words on different platforms? Does it have something to do with the algorithmic pressures of that website? Also, why do users self-censor using different techniques? Is there any correlation between their reasonings and their practices, or is it some other factor? Finally, if there is some correlation between reasoning and practices, what might this mean for the future of language in a world where communication is increasingly digital? This thesis grew out of those questions.

3 Background

3.1 Terms of Service, Algorithms, and Site Culture

In this paper, I am primarily investigating the social media sites Twitter, TikTok, and Tumblr. Twitter is a microblogging and social networking site, on which users can make ‘tweets,’ short posts up to 280 characters long. These tweets can include photos or videos as well, but the site is primarily text-based. TikTok is a video-hosting platform for short videos, and video is the primary medium, but these videos may have captions or descriptions. Tumblr is a multimedia microblogging and social networking site, where users can post videos, images, and long sections of text. Each of these sites has different content moderation policies, meaning that users may feel they need to self-censor differently on each site. The sites also differ in demographics, functionality, and site culture.

Twitter traditionally prohibits the use of threatening language against individuals and groups but allows hyperbolic speech that clearly indicates no violent intent, although due to recent changes in ownership these rules are currently in flux. It also prohibits the promotion of suicide or self-harm but allows users to discuss their experiences with such matters if they don’t share detailed information about strategies and methods. However, many Twitter users have had concerns about whether they would be suspended or banned for using specific terms. In Smith (2018), an article on a pop culture site, Smith displays multiple tweets that had were flagged as inappropriate or tweets that users believed had gotten them temporarily suspended, as well as tweets discussing users’ thoughts on the apparent recent surge in suspensions and deletions. Many of these tweets are now inaccessible. One tweet from @hotlinekream gives a list of “trigger words” that may cause suspension, including “k*l* m*s*lf”, “s*ic*de”, and “p*nch.”

These were deciphered by Smith as “kill myself”, “suicide”, and “punch”. @hotlinekream clearly believed that censoring their words with asterisks would circumvent the suspension - as the tweet itself is now inaccessible and their Twitter account appears to no longer exists at the time of writing, it is unknown if this truly worked, or if it was even necessary in the first place.

Twitter feeds are primarily composed of content from people a user follows, but will also feature content from the users those people follow, content from topics that user is interested in, tweets from topics that Twitter is suggesting to the user, and ads (Twitter Help 2023). Twitter users can also deliberately search for topics or hashtags they are interested in or browse popular tweets in trending topics. As such, a significant amount of content a Twitter user sees will be from people they do not follow, whose tweets are recommended in one way or another by the Twitter algorithm.

TikTok prohibits many forms of content it deems harmful or inflammatory, including any material that would endanger the safety of minors, content depicting or promoting dangerous acts, content that depicts or may encourage suicide or self-harm, nudity, sexually explicit content, bullying, harassment, threats of violence, content supporting or promoting hateful ideology or violent extremism, spam or impersonation, depiction of criminal activities, violent or graphic content, or copyright infringements (TikTok 2022). TikTok claims it does not restrict content due to political sensitivities, but users have noted that some content not explicitly restricted by their community guidelines has been removed or shadowbanned (Ryan et.al. 2020, 6). Shadowbanning is when content tagged with a given hashtag is “suppressed and often totally hidden from public view; posts are made much more difficult to find on the platform though they’re not necessarily deleted” (Ryan et.al. 2020, 6). Research shows that “hashtags related to LGBTQ+ issues are suppressed on the platform in at least 8 languages", and TikTok has consistently suppressed content related to political issues like the Black Lives Matter protests or anti-monarchy protests in Thailand as well (Ryan et.al. 2020, 4). While TikTok claims that it does not shadowban or restrict politically sensitive content, a former content moderator for the site told the *New York Times* in November 2019 that “managers in the US had instructed moderators to hide videos that included any political messages or themes, not just those related to China”, and that they were to “allow such political posts to remain on users’ profile pages but to prevent them from being shared more widely in TikTok’s main video feed,” (Ryan et.al. 2020, 10).

TikTok users are aware of the fact that discussion of sensitive topics may result in content being shadowbanned, removed, or restricted (Delkic 2022). Videos are both scanned for violations and can be reported by other users, and from there may be either automatically removed or referred for review by a human moderator (TikTok 2022). As such, many users on the site have begun to adopt substitutes for restricted words that may trigger automatic moderation (Delkic 2022). These include “panoramic” for pandemic, “leg booty” for LGBT, “cornucopia” for homophobia, “seggs” for sex, and “le$bian” for lesbian. While there is no public list of sensitive words, “some things are consistent enough that creators know to avoid them, and many share lists of words that have triggered the system,” (Delkic 2022). Some creators have noted that TikTok is less likely to flag videos talking about sensitive topics if the topics in question are incorporated into popular music and sounds. Black creators in particular have noted how “censoring the sharing of political or culturally relevant content while supporting more lighthearted content is similar in method to tone policing,” (Day 2021). Creators must be careful if they wish to discuss sensitive or banned topics, and many choose to either communicate in coded ways or through music that the platform deems acceptable.

One reason TikTok shadowbanning has such a strong effect is the way TikTok feeds work. TikTok has two main home pages - the Following page and the For You page, both of which are nonoptional. The following page includes videos from accounts you have chosen to follow, similar to many other social media sites. The For You page includes videos recommended to you by the TikTok algorithm, based on the people you follow, the videos you interact with, the content you make, and demographic data like age, gender, and location. TikTok’s algorithm personalizes each user’s For You page to show videos that it determines the user would like the most, and draws from popular videos in the topics that user would presumably be interested in. Shadowbanning, however, means that a video will not show up on anyone’s For You page, dramatically reducing the number of people it can reach. Users can also search for videos, but shadowbanned videos, or videos with shadowbanned hashtags, will not show up in searches. TikTok does not delete the shadowbanned content, but the only way to see it is to go directly to a creator’s page. Since site interactions revolve around the For You page, if a creator wants their content to be seen it is within their best interests to keep their content from being shadowbanned.

Unlike TikTok and Twitter, Tumblr does not rely on an algorithm to recommend content.

The main Following page of the dashboard is chronological, and while there is a For You page and an optional Best First feature, these are recent additions and not always used. After the implementation of the Best First feature, posts circulated informing people how to turn it off, as many users wished to avoid that option. Without an algorithm, posts will not circulate without reblogs, making reblogs essential to the site ecosystem. Likes – a button users can click on a post they see to express that they 'like' it - are useful for expressing support but will not affect the likelihood of a post to circulate. Tags, partially hidden comments that a user can choose to add to a post when they create or reblog it, can be used as organizational tools. However, they are also used as a place to put personal comments that a user does not deem important enough to put in the text of a reblog or doesn't want people besides their followers to see, since once a post is reblogged the previous user's tags are erased.

Culturally speaking, Tumblr has a long site-wide memory - a given meme or reference may come and go quickly, but some will remain in use for years on end. Most bloggers use pseudonyms, encouraging a certain level of disconnect from the "real world". While some celebrities use Tumblr, the lack of mandatory 'For You' algorithms make influencers and brand accounts rare. Additionally, there is no public follower count. There are advertisements by some companies, and users can 'Blaze' their posts to turn them into ads, but these ads are not targeted – they appear for a random selection of users.

Tumblr's rules are also functionally more relaxed than the other sites. The community guidelines discourage hate speech, the promotion of terrorism, child sexual abuse material, the promotion of self harm, violent content and threats, gore and mutilation, fraudulent links, spam, copyright infringement, impersonation, harassment, privacy violations, election interference, sexually explicit material, and other content (Tumblr, 2022). These guidelines are generally enforced with the use of algorithmic content moderation or reports. However, due to Tumblr's small size and limited staff, some of these guidelines are firmer than others, and enforcement is lax.

Due to its primarily anonymous nature and lax rule enforcement, Tumblr can be a chaotic site, and site culture can be combative as well - many users refer to it as a "hellsite" due to its generally esoteric nature and hostile user base (Mashable SEA 2022). When Twitter was bought by Elon Musk in October of 2022, some Twitter users who had previously used Tumblr expressed a desire to return to the site (Fishbein 2022). On Tumblr, however, most users expressed distaste with the concept of new or returning users, as these users could bring an influx of "the infighting and upheaval that users experienced during earlier iterations of the platform" (Fishbein 2022).

Tumblr is also notoriously dysfunctional. Many users have noted that it is easier to search for a specific Tumblr post on Google than with Tumblr's own search function (etakeh 2022). Tumblr's lack of 'For You' algorithms could also be seen as a sign of limited functionality. However, many users do not mind this aspect of the site, or consider it a positive. Many Tumblr users find the lack of algorithm reassuring, as it indicates the site is less likely to be collecting personal data than other sites. However, while the search function is limited, it is present, and users can go search for a specific term to find posts that mention that term or use it as a tag. This connects users that do not follow each other, and such connections can be unwanted in some situations (as discussed at the end of section 3.4).

Overall, Tumblr is regarded as a place for niche interests, subcultures and ingroups. The anonymity allows for a disconnect from public life - a user could badmouth their employer without fear of retribution - and the lax enforcement of rules gives users a freer reign over possible content - a user could describe the harm they wish they could inflict on a celebrity without fear of censorship.

All three sites employ some form of algorithmic content moderation, as large social media sites with millions of users cannot rely solely on human moderation to evaluate if content violates the Terms of Service. However, algorithmic content moderation faces many technical and political challenges, and even the best algorithms could exacerbate existing problems with content moderation policies. Gorwa et. al (2020) defines algorithmic moderation as "systems that classify user generated content based on either matching or prediction, leading to a decision and governance outcome (e.g., removal, geoblocking, account takedown)" (Gorwa et. al 2020, 3). These decisions may be done automatically by the algorithm, or could flag the content and send it to a human for review. Content moderation is necessary for many sites, as it is "one of the core commodities provided by a platform – enabling it to serve advertiser, as well as user needs, and therefore be a viable business" (5). YouTube, for instance, de-monetizes videos its algorithm deems 'toxic' or 'vulgar' to prevent advertisers from having their content paired with something that could damage their brand (Gorwa et. al 2020, 9-10). Moderation is also used to create a healthier site culture, thus attracting users, and to remove any potentially illegal or dangerous content.

There are multiple problems with algorithmic moderation, however. These include lack of transparency, perpetuations of injustice, and the obfuscation of the inherently political nature of moderation. Platforms are notably "cagey about the details of how they conduct algorithmic moderation," and users are often left unsure as to what exactly could cause a takedown of an account or content (Gorwa et. al 2020, 1). In this case, this means that when users are not precisely sure if the use of a term would cause punitive action, they may over-censor to avoid potential negative consequences. Platforms also may accidentally end up banning the use of terms related to protected groups, as algorithms used to identify hate speech may flag neutral language related to that group because that language may be used in a pejorative manner, even when used in a neutral or positive context (Dias Oliva et. Al. 2021, 729). Additionally, "language is incredibly complicated, personal and context dependent: even words that are widely accepted to be slurs may be used by members of a group to reclaim certain terms" (Gorwa et. al 2020, 10).

For instance, Dias Oliva et. Al. (2021) investigate how 'Perspective', a technology used to measure 'toxicity', is more likely to deem tweets by drag queens as 'toxic' than tweets by white nationalists. They note that in the drag queen community, 'mock impoliteness' and the reclaimed slurs are often used in positive contexts, but due to their pejorative usage in most other contexts, tweets using such language are often deemed highly 'toxic' by Perspective. On the other hand, white nationalist tweets comparing homosexuality to cannibalism or asserting the superiority of western culture received low toxicity ratings, possibly due to their lack of any words explicitly flagged as offensive. The paper found that "the most probable explanation for the findings described in the sections above is that machine learning techniques find correlation between input features (words) and target classification (toxicity)" (Dias Oliva et. al 2021, 729). Algorithms like this fail to fully understand context and appear to rely on the presence or absence of specific terms to determine what content is deemed offensive. Whether this is true for a specific website or algorithm is irrelevant to this paper - what matters is how users perceive the algorithm, and what they believe could circumvent it.

3.2 Taboos

While self-censoring on social media sites is a phenomenon inherently tied to the Internet, taboo avoidance practices have been around as long as taboos have and have been widely studied. Allan and Burridge (2006) note that taboos are inherently tied to culture, and stem from social constraints on behavior that could cause “discomfort, harm or injury”, including metaphysical or physical risk (1). These taboos then lead to language restraints, as referring to sensitive topics may cause harm. They differentiate between censorship and censoring, defining censorship to be “the suppression or prohibition of speech or writing that is condemned as subversive of the common good” by an institutional force, while censoring can be done by anyone, powerful or otherwise (13). Censorship is therefore a subcategory of censoring.

When speakers wish to avoid discussing a taboo topic, they may use an X-phemism, a term Allan and Burridge created to encompass euphemisms, dysphemisms, and orthophemisms. All X-phemisms are ways of referring to denotata, which are basic neutral expressions with little to no additional connotation. In most social situations there exists a possible denotatum, but what qualifies as a denotatum differs based on culture and context - ‘dog’ is a neutral term when referring to the animal, but a pejorative when referring to a person. Allan and Burridge define a dysphemism to be “a word or phrase with connotations that are offensive either about the denotatum and/or to people addressed or overhearing the utterance” (31). Dysphemisms are usually impolite, dispreferred language that people use to talk about things that annoy others, things they disapprove of and want to degrade, or curses, name-calling, and other derogatory content used to let off steam.

Polite behavior is non-dysphemistic and avoids the use of dispreferred language. Both euphemisms and orthophemisms are used as alternatives to dispreferred or taboo words or expressions and help avoid social loss of face. However, orthophemisms (a term created by Allan and Burridge) are more formal than euphemisms. In some cases, orthophemisms are technical language that are overly precise or esoteric. In others, they are synonymous with the denotata, but because in casual language people perceive the denotata as overly formal, they are categorized as orthophemisms. X-phemisms for the same denotata have cross-varietal synonymy, as they “have the same meaning as other words used in different contexts” (29). Some examples of cross-varietal X-phemisms referring to the same denotata are shown in Table 1.

Denotata	**Orthophemism**	**Euphemism**	**Dysphemism**

poop	*feces*	*number two*	*shit*
having sex	*having intercourse*	*making love*	*fucking*
die	*perish*	*pass away*	*kick the bucket*
prostitute	*courtesan*	*lady of the night*	*whore*
penis	*phallus*	*member*	*cock*

Table 1: Examples of X-phemisms for the same denotata

In a separate paper, Burridge (2012) identifies three main linguistic strategies for X-phemism formation - analogy, distortion, and borrowing. Analogy is the “generalization of forms to new situations", distortion is “modification of forms,” and borrowing is “incorporation of forms from elsewhere,” (Burridge, 21). Analogies take expressions from other parts of the language and use them in new situations through the use of metaphor, understatement, hyperbole, and other ways of re-expressing a denotatum with different language. Distortions modify the existing expressions in some way, through the use of shortening, acronyms, initialisms, ellipsis, circumlocution, phonological remodeling, affixation, blending, alliteration, and rhyming. Borrowing involves the substitution of other words from jargon and slang within the same language or words from other languages. Dysphemisms are more likely to borrow from slang, while euphemisms are more likely to use vague or general language, but all X-phemisms can be formed from any of these linguistic strategies.

3.3 Terms, Typologies and Taxonomies

Another word for self-censoring is ‘Voldemorting’, a term referencing the Harry Potter series that was first codified and defined by an internet blogger known as Eugene, who defines Voldemorting as “when you deprive someone terrible of power by refusing to speak their name,” (Eugene 2013). Nagel (2018) expands upon this tactic in an investigation into user’s interference with algorithmic connections on social media. Algorithmic connections are the links social media sites make between “otherwise disparate data points” and are often used to increase “the number of users and the time they actively spend on the platform.” If users dislike the prospect of a platform creating these connections, they may seek to thwart those strategies by avoiding algorithmic connections via tactics. One form of algorithmic connection is the search algorithm, which will bring up any posts using the words in the search. Voldemorting is a tactic used to avoid this particular connection, as it is often employed to prevent posts from being found by other users through searches. This study investigates avoidance of attention from other users as one of many possible reasons for self-censoring.

Cho and Kim (2021) investigate intentionally noisy user-generated text and create a typology of avoidance strategies and a taxonomy of noisy texts. They categorize three kinds of stakeholders for avoidance strategies: these are the author, those who should understand the text (peers), and those who should not understand the text (others). When both peers and others exist, the noisy text is a ‘trick’ meant to make the text understood by peers and overlooked by others. When only peers exist, the text is a ‘meme’ meant to be understood by peers but does not intentionally exclude others. When only others exist, it is a ‘filler’ meant as personal expression that avoids others understanding it. When neither exists, it is a ‘code’ meant to be only comprehensible to the author. Tricks and fillers are similar, as both often seek to avoid censorship - however, tricks are used when the author is trying to communicate with a specific audience, while fillers are used when the author does not have a set audience of peers. Tricks and fillers are commonly found in self-censoring, as many users employ self-censoring and taboo avoidance strategies to avoid censorship or to avoid algorithmic connection created by an algorithmic ‘other’ that would connect them to an undesired audience. However, some users may simply be adopting the self-censoring practices they see around them to participate in site culture, and therefore the same tactics used as fillers or tricks in the hands of some users may be memes in the hands of others.

Further, Cho and Kim create a taxonomy of noisy texts, and identify morphological, morpho-phonological, optical, semantic, and other strategies as ways to create noisy text. I have created my own categorization system, with four main categories of tactics used in self-censoring - asterisk replacements, like ‘k*ll’, symbol or number replacements, like ‘k!ll’ or ‘k1ll’, phonetic X-phemisms or misspellings, like ‘krill’, and non-phonetic X-phemisms, like ‘unalive’ (Table 2). Symbol and number replacement are optical strategies that rely upon using visually similar characters - ‘i’ visually resembles ‘!’ and ‘1’. Asterisk replacements are optical as well, but do not rely on visual similarity, hence their separation into a different category. Phonetic X-phemisms are what Cho and Kim would define as morpho-phonological, as they replace a word with a word that sounds similar - ‘krill’ sounds like ‘kill’ and ‘super slide’ sounds like ‘suicide’. Non-phonetic X-phemisms, like ‘unalive’ for ‘kill’ or ‘dead’ or ‘Orange Cheeto’ for ‘Trump’ would be semantic, as they are intended to refer to the same thing as the censored word while using completely different sounds, letters, and lemmas.[1]

Denotata	kill
Asterisk Replacements	k*ll
Symbol or Number replacements	k!ll or k1ll
Phonetic X-phemisms or misspellings	krill
Non-phonetic X-phemisms	unalive

Table 2: Categories of self-censoring tactics and examples of their implementation

The four main self-censoring tactics detailed above are divided into two main categories based on Burridge’s strategies for X-phemism formation, and from there divided based on Cho and Kim’s taxonomy. Asterisk replacements, symbol or number replacements, and phonetic X-phemisms are all distortions that rely on the audience knowing the sound or form of the underlying word and deciphering the distortion to uncover it. Non-phonetic X-phemisms are analogies or borrowings that use other words or metaphors to convey a similar semantic meaning without relying on the sound or form of the underlying word. The distortions can be divided into optical and phonetic, with asterisk replacements and symbol or number replacements being optical and phonetic X-phemisms being phonetic. However, it should be noted that another categorization schema may be useful as well - character composition. Asterisk replacements and

[1] At this point, I must make a terminology clarification. Nagel (2018) uses ‘tactics’ to refer to the practices users employ to avoid algorithmic connections and ‘strategies’ to refer to the methods social media sites use to create connections and increase engagement. Cho and Kim (2021), however, use ‘strategies’ to refer to the practices users employ to create noisy text, such as morphological or optical strategies , while Burridge (2012) uses ‘strategies’ to refer to manners of word formation. I will be using ‘tactics’ and ‘strategies’ interchangeably to refer to the taboo avoidance practices users employ, and ‘reasons’, ‘motivations’, and other similar words to refer to the reasons why users employ these strategies.

symbol or number replacements inherently include non-alphabetic symbols, while both phonetic and non-phonetic X-phemisms do not. These categories will allow for a more specific analysis, as trends along category lines would indicate potential causes for using one category over another. A representation of this division can be seen below, in Fig. 1.

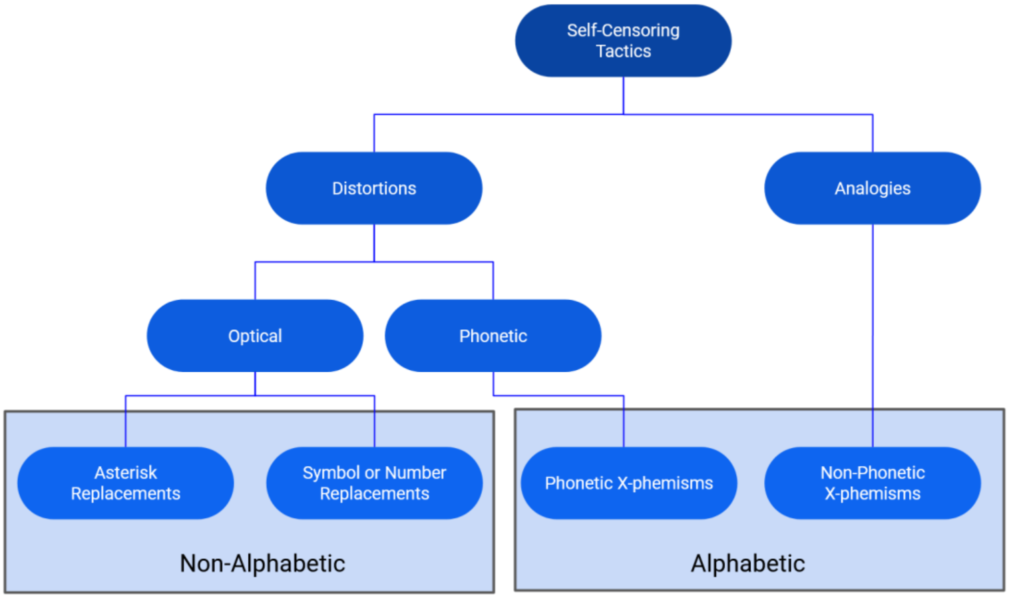

Fig. 1: A tree diagram of a categorization scheme for the different self-censorship tactics

3.4 Communication and Memes

Yus (2005) assesses the communicative usefulness of noisy user-generated text (which he terms textual deformations) like repetition of letters and punctuation marks in Spanish chatrooms. He found that while readers may not agree on the quality of the sender's emotions, and do not assign intensity based on quantity of text, they do play a part in the way users interpret text. These deformations can lead readers to the underlying message of the text and "underlying propositional attitudes, affective attitudes and emotions attached to the message when it is typed on the computer keyboard" (148). Facial expressions and gestures can function as paralinguistic codes that convey the attitudes, feelings and emotions of the speaker. In text-only environments like chat rooms, textual deformations can work in a similar manner, making up for the lack of verbal or visual information. Yus found that these textual deformations are "the outcome of an intentional verbal strategy used by chat users when they are willing to connote their messages with attitudes and emotions," but that textual deformation is not good at conveying "subtle varieties of attitudes and emotions" (168).

For some users, self-censoring may convey attitudes and emotions in a similar way as other forms of textual deformation. For example, if a user sees others in their community censoring people's names when disparaging them, they may associate such censoring with a negative connotation. As such, they would parse the sentence "J*hn Sm*th was performing today" as carrying the connotation that the speaker is annoyed at the presence of John Smith, while "John Smith was performing today" would not carry such a connotation. Some forms of self-censoring like dysphemisms inherently carry a negative connotation, but symbol replacements may not be parsed as pejorative by someone outside of the community where they are used. As such, distortions may function as tricks and memes more often than analogies, as both require an ingroup that can understand a hidden meaning.

Yus (2018) addresses meme communication as it relates to identity and posits that "every single stage of meme communication entails a greater or lesser impact on the user's self-concept, self-awareness and overall identity" (1). Some possible effects on identity are feelings of connectedness, reduced loneliness, feeling noticed, being more willing to self-disclosure, the generation of social capital, and feelings of well-being through emotional display. Yus identifies the stages of meme communication as decoding, inferring, sharing, strengthening, and spreading. During decoding, the user's ability to identify the genre of a meme and its meaning makes them aware of their place in a community - "the identification of the discursive qualities of the meme signals appropriateness and, ultimately, group membership" (5). During inferring, users draw on context clues and knowledge of cultural texts to correctly decipher how a meme is meant to be interpreted. For instance, a meme about food would have different connotations in a cooking group versus a dieting group, even if the meme itself is the same. By identifying the intended context and understanding any references, users can place themselves in the same ingroup as the creator of the meme.

During sharing, "there are different levels of impact on identity depending on whether the "addressee user" feels that he/she is part of a "mass circulation" of the meme, or feels that the "sender user" has deliberately chosen him/her as selective recipient of the meme" (9). Meme sharing is a social phenomena, and so addressees' awareness of how the meme was shared with them affects their interpretation of their relationship with the sender. During strengthening, the previous stages affect social standing and belonging in a group. Having ‘meme literacy’ - being able to properly decode and interpret a meme - indicates commitment to the community and may affect user’s standing in those communities. Lastly, during spreading, since memes “both reflect norms and constitute a central practice in their formation,” only memes that are suited to the community environment will spread, indicating that users will only share memes that complement that specific environment (10).

For some self-censoring tactics, users may see the self-censoring as a meme that conveys information in a coded manner. In this case, using a self-censoring tactic would convey both the literal information and an awareness of community culture. Even if some users self-censor for different reasons, others may see their tactics as a form of slang. Not understanding or using that slang would indicate a lack of commitment to the community, so users may adopt it to participate in site culture.

However, users may also champion a lack of self-censoring as an indication of community belonging. Many Tumblr users view excessive self-censoring as both unnecessary and unkind. As one Tumblr user says in a post addressing new arrivals from other social medias (bundibird, 2022):

> “Hi all – newbies in particular, and newbies from tiktok extra in particular…Please be aware that if you are tagging sensitive topics for the sake of other people’s blacklists, you NEED TO SPELL THE WORD CORRECTLY. Lets use the word “yellow” as an example. Let’s say the word yellow means something horrible, and many people might want to filter it so that they don’t see posts that mention or discuss anything to do with the word yellow. If you tag the potentially triggering post as “y3llow” or “ye11ow” or “yell0w,” then youre actually circumventing most peoples block lists. Youre making your post highly difficult to avoid by anyone who doesn’t want to see that word. People have to add every possible spelling combination of the word “Y3770w” to their block lists, and even then, some that they haven’t thought of will slip through their blocklist net and will give them an unpleasant surprise.”

Other posts in a similar vein mention that self-censoring in this way makes it difficult for screen readers to read a post or tags. Others, like user pockopeas, simply take pride in not needing to self-censor (pockopeas, 2022):

> “its so funny to me that people on twitter n tiktok are like “ok but porns still banned on tumblr so at least we’re better then them” as if they dont have to typ3 w0rd$ I1k3 th!$ to get around their censors”

It should be noted that Tumblr users do self-censor in some situations, both for avoiding algorithmic connections and participation in site culture. In a response to other users discussing how some Tumblr users self-censor to ‘be mean’, user astraltrickster notes that originally, Tumblr users would self censor to avoid being unkind to other users. They note that there was “an etiquette guideline - do not post your negativity in the public tags…Because people go into the public tags to find content about things they LIKE” (astraltrickster, 2022). As such, self-censoring something in a post complaining about it is done “to keep it from showing up when people are searching for what they love and to prevent pointless drama.” Eventually, however, it became “just part of our site culture, for both peacekeeping reasons and petty glee” (astraltrickster, 2022).

The perception of self-censoring on Tumblr is evidence that self-censoring practices can not only take many forms or mean different things to different users, but can evolve over time. In this case, self-censoring was originally implemented as a tactic to avoid algorithmic connections that would unnecessarily aggravate other users. As this practice continued, self-censoring gained the connotation of negativity, and users began to parse it as textual deformation that indicated the underlying feelings of the poster. From there, it evolved into a meme, as the practice of censoring a name to make it a pejorative became both humorous and a way to indicate allegiance to site culture.

With this information about each of the variables - rules and cultures on social media sites, taxonomies for the tactics of self-censoring, and the memes and taboos that may be reasons for self-censoring - we can move into investigating how these variables interact with the research questions.

4 Methodology

4.1 Research Questions

My main research question with this data is whether there is an association between the reason given for self-censoring and the self-censoring tactics used. Since the reasons for self-censoring may differ based on social media and site culture, I separated my data between those sites. To get a full view on the question at hand, I investigated three variables: social media site, self-censoring tactic, and reason for self-censoring. I first used basic percentage calculations to investigate two auxiliary questions: 'What self censoring tactics are used most on which sites?' and 'What are the most common stated reasons for self-censorship on each site?' I then used a chi-squared test to answer my main question, 'Is there an association between self-censoring reasonings and self-censoring tactics?'

4.2 Research Process

My research primarily involved surveying social media users about their use and observation of self-censoring. Part way through the survey process, I made minor edits to the survey to allow for more options and greater comprehension of questions. Most notably, I added 'No censoring' as an option to many questions that previously only had 'Other.' Since many respondents wished to answer 'No censoring' and before the update could only indicate so in 'Other', this was a necessary change to make. As such, I have included both the original and edited survey in the appendix.

The survey was distributed in three ways – word of mouth through family and friends, messages in three private Discord servers, and a post on my Tumblr. During the month that the survey way open, I received an unexpectedly large number of responses from Tumblr users, as the survey was reblogged by a popular blog and spread rapidly from there. As such, I was expecting 100 responses total, and received 3,390, the majority of which are from Tumblr; 84.95% of respondents reported Tumblr as one of their primary social medias, with Twitter at 8.24%, TikTok at 3.09%, and other social media at 10.86%. As such, the sample size for Tumblr users will be much larger than the other social media sites.

To simplify the analysis responses, I chose to only analyze responses from the updated survey, as the clearer division between 'No censoring' and 'Other' allows for more accurate analysis. These criteria narrows down the total analyzed responses from 3,390 to 2,090. Further reducing the survey responses to only those who said yes to the consent form and correctly answered the consent form comprehension question (16) yields 1,380 analyzed responses. Of those people, 1,116 are primarily Tumblr users, ninety-six are Twitter users, and 40 are TikTok users. I will be analyzing each separately below. Refer to the appendix for the full text of each question mentioned.

In the survey, I asked three usage questions (7-9), that asked respondents how often they used Tumblr, TikTok and Twitter on a scale of 1 to 5. I also asked what the respondent's primary social media was (in question 10). To identify which social media a respondent primarily uses, I qualified any response of three or above on the usage questions (7-9) and the selection of that social media as a primary choice on question 10 as an indication of use. Both factors must be present for the user to count in that category. To exclude users who use another social media, there must be a response of below three on the usage questions (7-9) and they must not mark the social media as a primary choice on question 10. If a user fulfills one requirement but not the other for the other social media, they are included.

I am dividing my data into three main sections - solely Tumblr users, solely Twitter users, and solely TikTok users. While I had intended to investigate the intersection of these sites, according to my primary use qualifications, only 8 people reported being both primary users of Tumblr and TikTok, with only 16 for Tumblr and Twitter, one person for Twitter and TikTok, and none for the intersection of all three. I will not be analyzing people who listed their only primary social media as none of these options, but those who listed both Tumblr and Facebook, for example, will be counted as Tumblr users if they listed their usage of Tumblr as three (uses Tumblr occasionally) or above on question seven. I will also sort these by social media. Other demographic factors may affect responses but were not standardized for this research. I will state relevant demographic for each social media category in section 6.4, and the full demographic data of the 1,380 analyzed responses in section A.2.

5 Results

5.1 Tactics and Sites

The question to answer here is 'What self-censoring tactics are used most on which sites?' This is a straightforward question, but the answer is more complicated. The sites I analyzed were Tumblr, TikTok and Twitter, using the categorization of primary media described above. The possible strategies were asterisk replacement, symbol or number replacement, similar sounding euphemisms, different sounding euphemism, no censoring, or other[2]. The strategy used was determined by the answer to question 11, “Do you/other users of your primary site censor words/phrases, e.g., kill, by asterisk replacement (like k*ll), symbol or number replacement (like k!ll or k1ll), euphemisms/misspellings that sound similar (like krill), euphemisms that sound different (like unalive), or other? As a side note, euphemisms in this case are any words or phrases used to refer to a concept without stating it outright.”

Since respondents could select multiple responses for this question, I decided to look at not only the raw data for each response, but also the separated data for each combination of tactics. For the raw data, I first found the total counts of responses and then calculated the percentage of respondents for each site who selected a response. For the separated data, I made each combination of responses a different category - ‘No censoring+Asterisk replacement’ would be a separate category from only ‘No censoring’ or only ‘Asterisk replacement.’ As such, none of the categories overlap.

For simplicity, in the section below ‘No censoring’ is represented as ‘No’, ‘Asterisk replacement’ as ‘Asterisk’, ‘Symbol or number replacement’ as ‘Symbol’, ‘Similar sounding euphemism’ as ‘SS Euph’, ‘Different sounding euphemism’ as ‘Diff Euph’, and ‘Other’ as ‘Other’.

5.1.1 Raw Data

Below is a table representing the counts of responses for each tactic and site. While the raw data contains overlaps, it can provide an interesting overview to the question.

	Tumblr	TikTok	Twitter
Asterisk	178	14	38
Symbol	58	20	30
SS Euph	39	21	28

[2] The original categories of ‘Phonetic X-phemisms or misspellings’ and ‘Non-phonetic X-phemisms’ were simplified for the survey, as ‘phonetic’ and ‘X-phemism’ are terms that many survey takers may not have been familiar with.

Diff Euph	78	30	31
Other	59	1	6
No	954	4	35

Table 3: A table counting each response for a given tactic and site - responses can overlap

As the number of respondents varies so dramatically between sites, the graphs below better display the distribution of tactics across each site. I chose to use a column chart rather than a pie chart because the percentages for each site tactic add up to more than 100%, due to overlap. There were 1116 total respondents from Tumblr, 40 from TikTok, and 96 from Twitter.

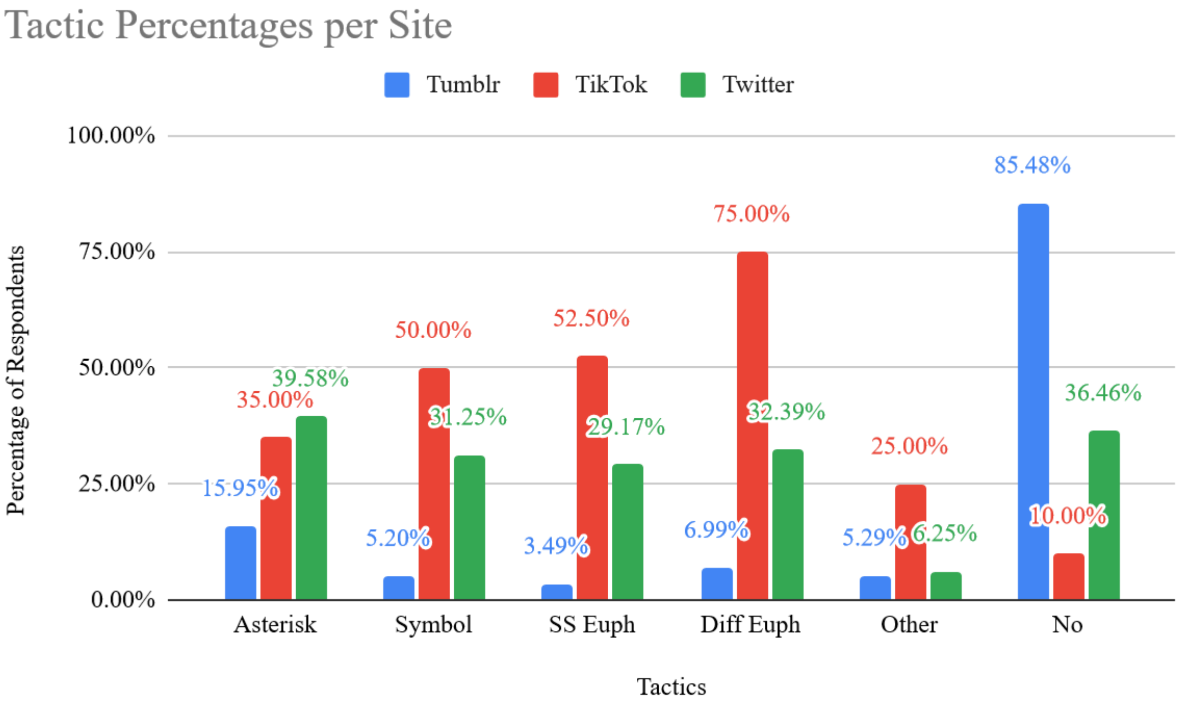

Fig 2: A column chart recording the percentage of responses for each tactic on each site

Here, we can clearly see that the majority of Tumblr respondents - 85.48% - selected ‘No,’ indicating they either do not censor, do not see censoring, or both. This contrasts with Twitter, with 36.46% selecting ‘No,’ and TikTok, with only 10%. Besides ‘No,’ all choices were low, with ‘Asterisk’ being the highest of them, and the only one above 10%, at 15.95%.

The highest reported tactic for Twitter was ‘Asterisk’ at 39.58%, but Twitter respondents selected a fairly even spread of responses, with all choices besides ‘Other’ between 29% and 40%. ‘Other,’ however, was low, at only 6.25%.

For TikTok, the majority of respondents -75% - selected ‘Diff Euph’, contrasting with 6.99% from Tumblr and 32.39% from Twitter. However, unlike Tumblr, the highest choice does not overshadow all the others. ‘SS Euph’ and ‘Symbol’ received 52.5% and 50% respectively, with ‘Asterisk’ at 35% and ‘Other’ at 25%. The only low response was ‘No,’ at 10%.

This data displays some trends among the sites concerning what self-censoring tactics users mostly see or use, but since it does not show overlaps or combinations, it does not show the full picture.

5.1.2 Separated Data

To examine the data more accurately, I created 63 different categories, one for each possible combination of tactics. I then found the percent of each category for each site by dividing the count for each category by the total number of responses. Since “respondents who selected ‘No’” is different than “respondents who selected only ‘No’”, the percentages for each category differ from those seen above. Below are a series of pie charts depicting the results.

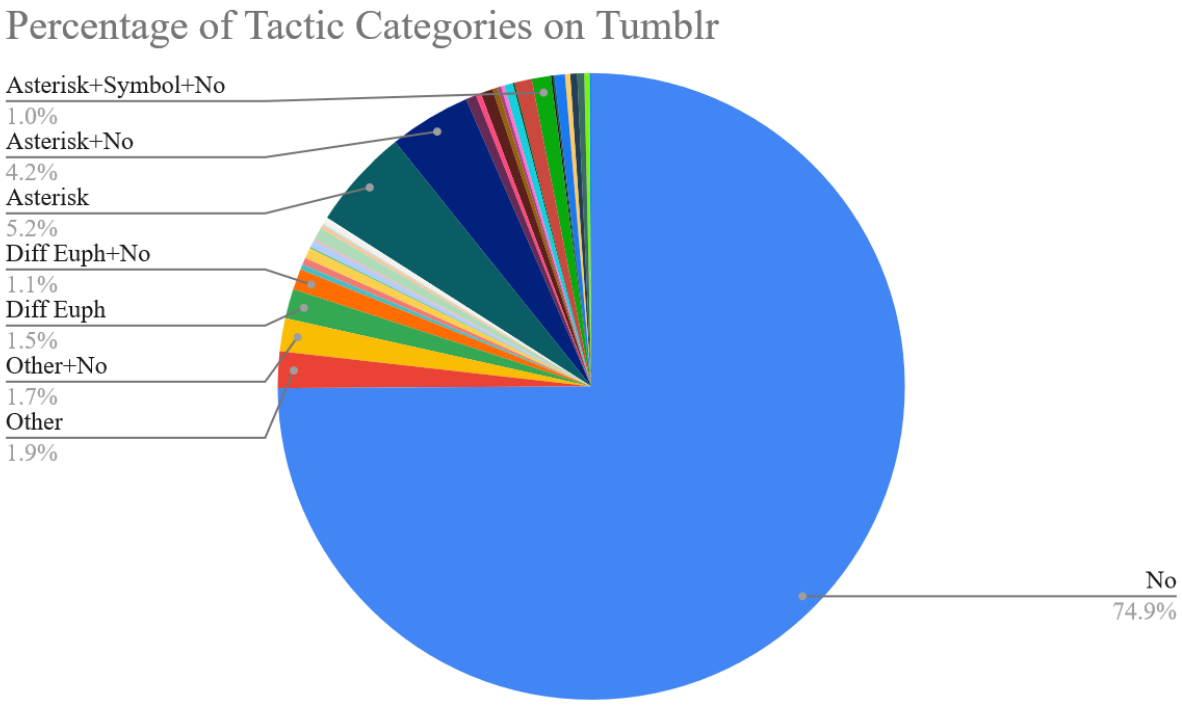

Fig. 3: A pie chart representing the results received for each tactic category on Tumblr

For Tumblr, a full 74.9% of respondents selected only 'No.' The only other categories for Tumblr above 2% were only 'Asterisk', at 5.2%, and 'No+Asterisk', at 4.21%. The only others above 1% were only 'Other' at 1.88%, 'Other+No' at 1.7%, only 'Diff Euph' at 1.52%, and 'Diff Euph+No' at 1.08%.

Here, we see that the many of the respondents not only selected 'No,' but selected only 'No', indicating that most Tumblr respondents do not see or use self-censoring tactics at all. Some, however, reported 'No' alongside other responses, or did not select 'No' at all and chose other tactics. Of these other tactics, 'Asterisk' is the most prominent, appearing both by itself and alongside 'No.' Since 74.9% of respondents selected only 'No,' only 25.1% of Tumblr respondents selected any tactics at all.

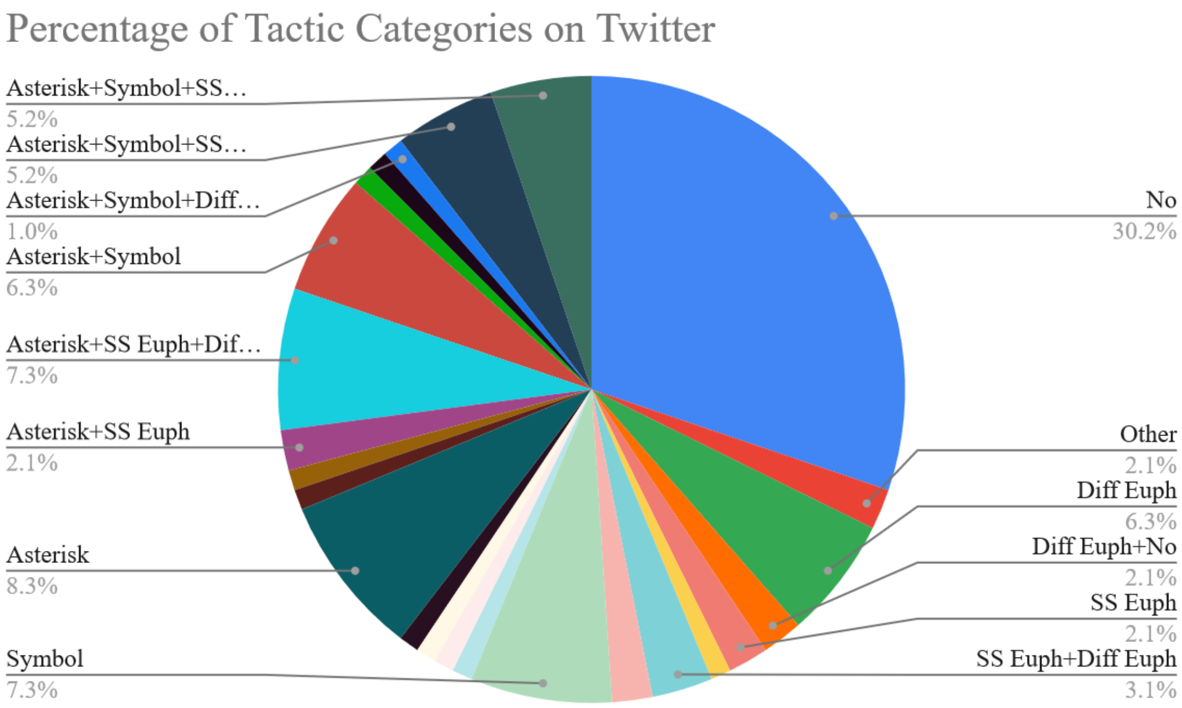

Fig. 4: A pie chart representing the results received for each tactic category on Twitter

30.2% of Twitter respondents selected only 'No.' No other categories came above 10%, but quite a few others were below 10% but above 2%, these being only 'Other', only 'Diff Euph', 'Diff Euph+No', only 'SS Euph', 'SS Euph+Diff Euph', only 'Symbol', only 'Asterisk', 'Asterisk+SS Euph', 'Asterisk+SS Euph+Diff Euph', 'Asterisk+Symbol', 'Asterisk+Symbol+SS Euph', and 'Asterisk+Symbol+SS Euph+Diff Euph'.

The largest single category for Twitter respondents was only ‘No’ at 30.2%, close to the total number of respondents who selected ‘No,’ 36.46%. For the self-censoring tactics, however, there was a great deal of overlap. Only ‘Asterisk’ was at 8.3% and only ‘Symbol’ was at 7.3%, but Asterisk+Symbol was at 6.3%, not far behind, and ‘Asterisk+SS Euph+Diff Euph’ was at 7.3% alongside only ‘Symbol.’ As such, while no single tactic alone was more than 8.3%, the percentage of respondents who selected at least one tactic (and not only ‘No’) was 69.8%.

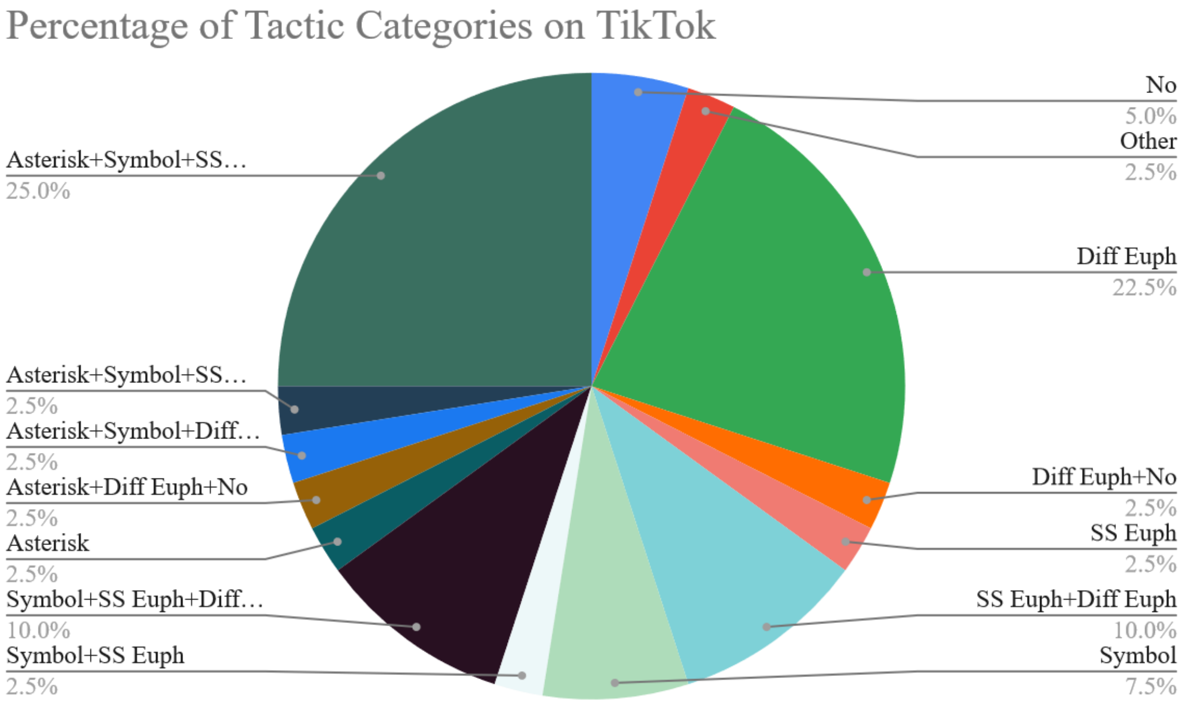

Fig. 5: A pie chart representing the results received for each tactic category on TikTok

For TikTok, 25% of respondents selected ‘Asterisk+SS Euph+Diff Euph+Symbol’, which is the response that selects all strategies besides ‘Other’ and ‘No’. 22.5% of TikTok respondents selected only ‘Diff Euph.’ ‘Symbol+SS Euph+Diff Euph’ received 10%, as did ‘SS Euph+Diff Euph’, while only ‘No’ received 5%. 7.5% of people selected only ‘Symbol.’ Several categories received 2.5%, these being only ‘Other,’ ‘Diff Euph+No’, only ‘SS Euph’, ‘Symbol+SS Euph’, only ‘Asterisk’, ‘Asterisk+Diff Euph+No’, ‘Asterisk+Symbol+Diff Euph’, and ‘Asterisk+Symbol+SS Euph’.

TikTok differs from Tumblr and Twitter in that its largest category is a combination of tactics rather than only ‘No.’ Here, only ‘No’ is only 5%, meaning 95% of respondents selected at least one tactic. Like Twitter, TikTok displays a large amount of overlap between tactics, as seen by the highest category being a combination of all of them besides ‘Other’, but unlike Twitter there is a clear standout among the tactics. 22.5% of respondents selected ‘Diff Euph’ alone, and it appears in both categories that received 10%, ‘Symbol+SS Euph+Diff Euph’ and ‘SS Euph+Diff Euph’.

With both the raw data and the separated data, we can begin to parse exactly where the usage and observation of each tactic (or lack of tactic) falls for each site. We have straightforward evidence of what tactics respondents reported using or seeing on which sites and how the sites differ in this manner. The answer to my question ‘What self censoring tactics are used most on which sites?’ is clear. Tumblr users mostly use or observe no self-censoring tactics, but many of the few who do use asterisk replacements. Some Twitter users use or observe no self-censoring tactics, but those who do are around equally likely to use some combination of asterisk replacement, symbol or number replacement, similar sounding euphemisms, or different sounding euphemisms. The vast majority of TikTok users do use or observe self-censoring, and many of those who do use different sounding euphemisms, but asterisk replacement, symbol or number replacement, and similar sounding euphemisms are not rare.

Before we dive into the implications of these results, however, we should first examine the rest of the questions.

5.2 Reasonings and Sites

Our next question is ‘What are the most commonly stated reasons for self-censorship on each site?’ Similar to the above, I investigated TikTok, Twitter, and Tumblr. The possible reasonings were “I am avoiding shadowbanning/censorship from site administration”, “I am trying to avoid attention from other users/detection from searches”, “I am adopting the terms I see around me to participate in site culture”, “I do not self censor”, ‘Humor’, and ‘Other’. The first two reasons were directly influenced by my research, which indicated that site policies and algorithmic connections cause some users to self-censor. The rest are based on personal observation of user behavior on various sites and casual conversations with friends and acquaintances about their self-censoring practices.

The reasoning listed was determined by the answer to question 20, “If you self-censor on your website, why do you do so?” While ‘Humor’ was not an option listed for that question, a sizable percentage of respondents who selected ‘Other’ gave a reasoning related to the use of self-censoring to communicate irony, comedy, or parody. As such, I separated out ‘Humor’ from the rest of the ‘Other’ responses for analysis.

Again, since respondents could select multiple responses for question 20, I looked at not both the raw data for each response and the separated data for each possible combination of tactics. For the raw data, I first found the total counts of responses and then calculated the percentage of respondents for each site who selected a response. For the separated data, I made each combination of responses a different category - ‘Humor+I do not self censor’ would be a separate category from only ‘Humor’ or only ‘I do not censor.’ As such, none of the categories overlap. However, it should be noted that there was no way to completely separate ‘Humor’ from ‘Other’, as in the responses ‘Humor’ is a subset of ‘Other.’ As such, those who selected only ‘Humor’ were categorized under ‘Humor+Other’, and ‘Other’ included both ‘Humor’ responses and ‘Other’ responses.

For simplicity, in the following section, “I am avoiding shadowbanning/censorship from site administration” is represented as ‘Censorship’, “I am trying to avoid attention from other users/detection from searches” is represented by ‘Connection’, “I am adopting the terms I see around me to participate in site culture” is represented by ‘Culture, “I do not self censor” is represented by ‘No’, and ‘Humor’ and ‘Other’ are the same.

5.2.1 Raw Data

	Tumblr	TikTok	Twitter
Censorship	62	19	27
Connection	185	3	34
Culture	92	12	19
Humor	77	2	5
Other	39	3	7
No	860	16	42

Table 4: A table counting each response for a given reason and site - responses can overlap

As the number of respondents varies so dramatically between sites, the graphs below better display the distribution of reasonings across each site. I chose to use a column chart rather than a pie chart because the percentages for each reasoning tactic add up to more than 100%, due to overlap. There were 1116 total respondents from Tumblr, forty from TikTok, and 96 from Twitter.

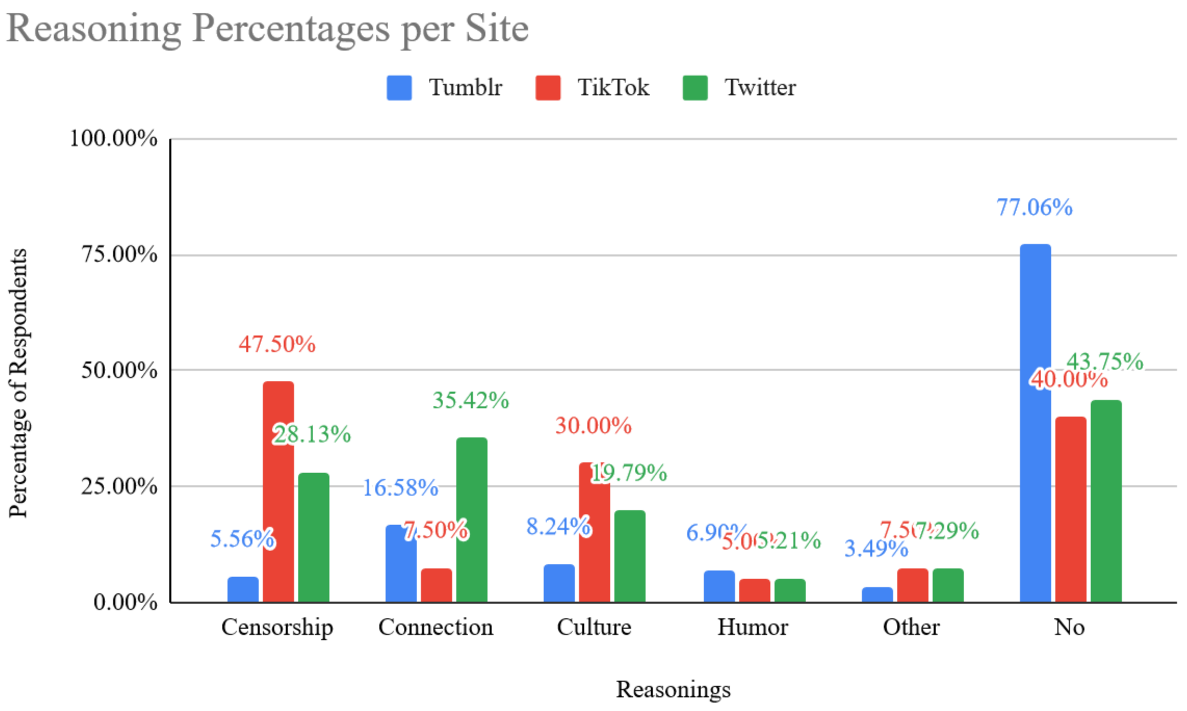

Fig. 6: A column chart recording the percentage of responses for each reason on each site

Tumblr users once again leaned most heavily towards no censoring, as 77.06% of respondents selected 'No.' The rest of the choices skewed low again too, with most being below 10%. The only reasoning above 10% was 'Connection', at 16.58%.

Unlike for the tactics, Twitter respondents did not select an even spread of choices. 43.75% selected 'No,' with 'Connection' next with 35.42% and 'Censorship' trailing further with 28.13% and Culture far behind with 19.79%. 'Humor' and 'Other,' though, were below 10%, much lower than any of the others. I should note, however, that these categories did not exceed 10% for any site.

The highest choice for TikTok respondents was 'Censorship' at 47.5%, but this choice was closely followed by 'No' at 40%. The only other choice above 10% was 'Culture' at 30%, with 'Connection', 'Other', and 'No' all low.

This data shows us some trends among the sites concerning what reasons users have for self-censoring, but we must once again examine the overlaps and combinations of these reasons to see the full picture.

5.2.2 Separated Data

To analyze the data more accurately, I created sixty-three different categories, one for each possible combination of reasonings (and excluding responses that selected no reasonings). After creating the categories and counts, I found the percent of each category for each site by dividing the count for each category by the total number of responses. Below are a series of pie charts depicting the results.

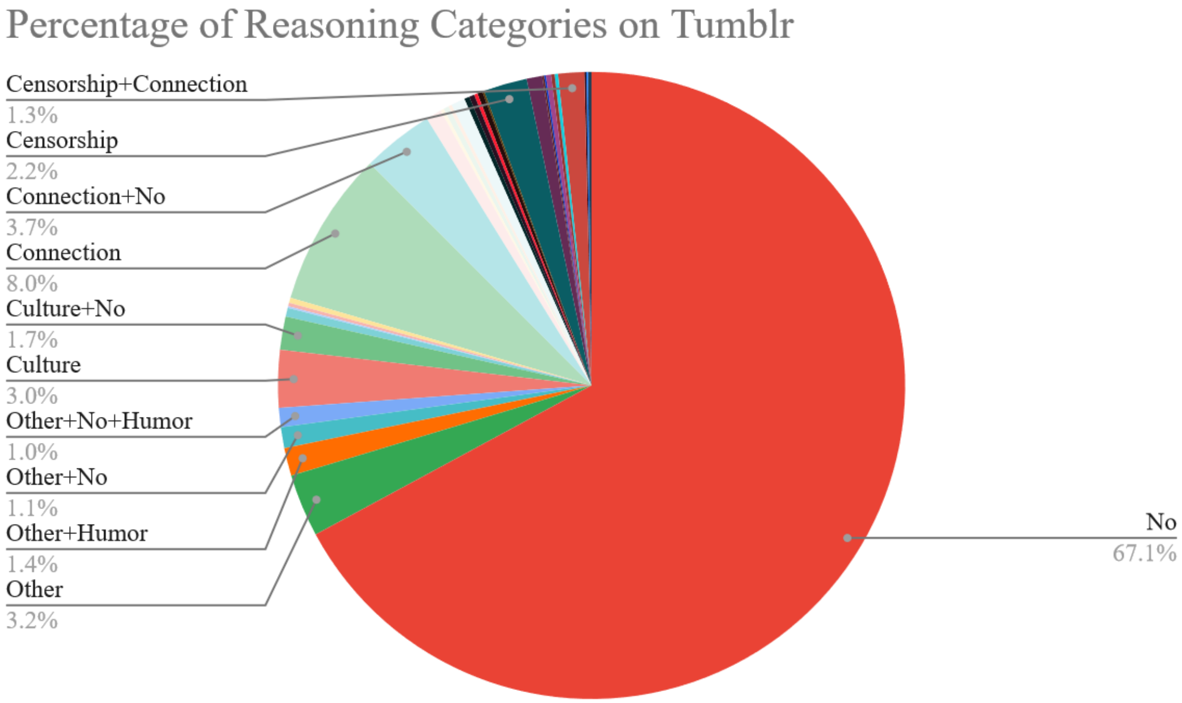

Fig. 7: A pie chart representing the results received for each reasoning category on Tumblr

For Tumblr, 67.12% of respondents selected only ‘No,’ continuing the trend from the previous section. After that, the next highest is only ‘Connection,’ with 8%, ‘No+Connection’ with 3.68%, ‘Other’ with 3.23%, only ‘Culture’ with 2.96%, only ‘Censorship’ with 2.25%, ‘Culture+No’ with 1.7%, ‘Other+Humor’ with 1.44%, ‘Censorship+Connection’ with 1.35%, and ‘Other+No’ with 1.08%.

Only ‘No’ dominates the responses once more, indicating only 32.9% of respondents reported self-censoring at all. Of those who self-censor, however, many selected ‘Connection,’ since the next two highest categories were ‘Connection’ and ‘No+Connection.’ ‘Censorship,’ ‘Culture’, ‘Other’, and ‘Humor’ were present but low, and of the reasons mentioned, ‘Connection’ is the most prominent.

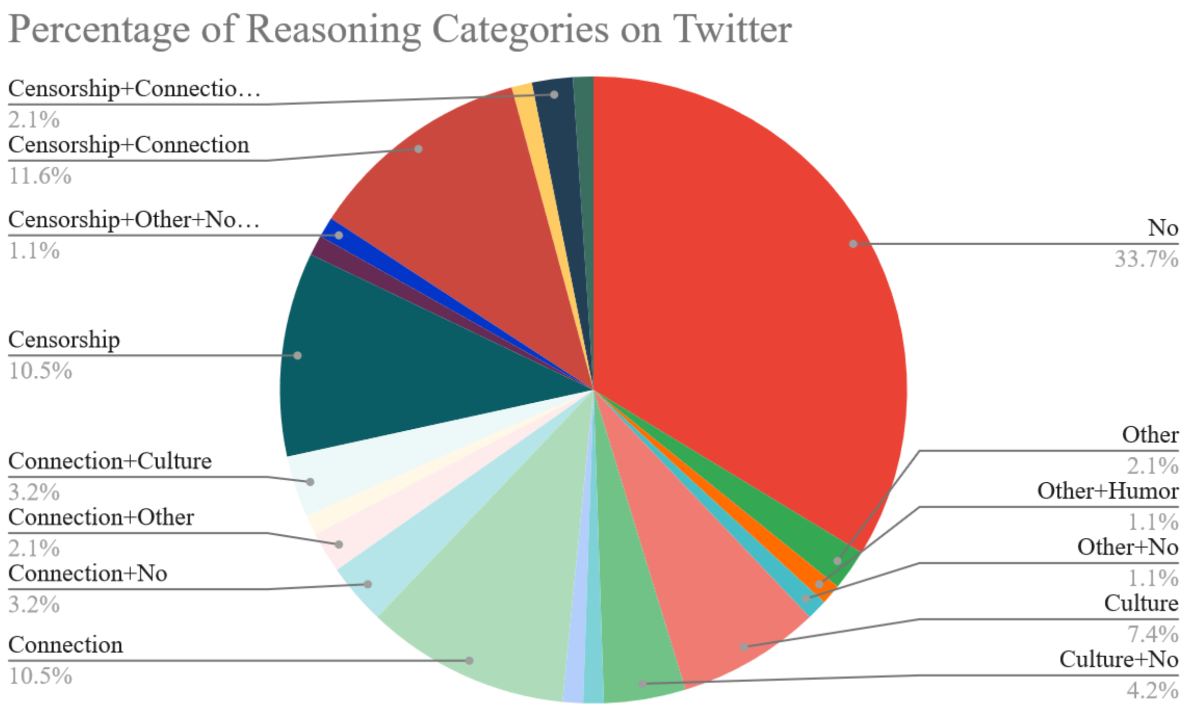

Fig. 8: A pie chart representing the results received for each reasoning category on Twitter

For Twitter, 33.68% of respondents selected only ‘No,’ also continuing the trend from the previous section. After that, the responses above 2% are ‘Censorship+Connection’ with 11.58%, only ‘Censorship’ and only ‘Connection’ tied at 10.53%, only ‘Culture’ at 7.37%, ‘Culture+No’ at 4.21%, ‘Connection+No’ at 3.16%, ‘Connection+Other’ and only ‘Other’ tied at 2.11%, and ‘Censorship+Connection+Culture’ at 2.11% as well. Below 2% and above 1% are ‘Other’ and ‘Humor’, ‘Other+No’, ‘Culture+Other’, ‘Culture+Other+Humor’, ‘Connection+Other+Humor’, ‘Censorship+No’, ‘Censorship+Other+No+Humor’, ‘Censorship+Connection+Other+Humor’, and ‘Censorship+Connection+Culture+Other’.

‘No’ remains the highest here, with but the majority of respondents - 66.3% - reported some reasoning. Of the categories, ‘Censorship+Connection,’ ‘Censorship’ and ‘Connection’ are most prominent, with ‘Culture’ not far behind and other combinations filling out the rest.

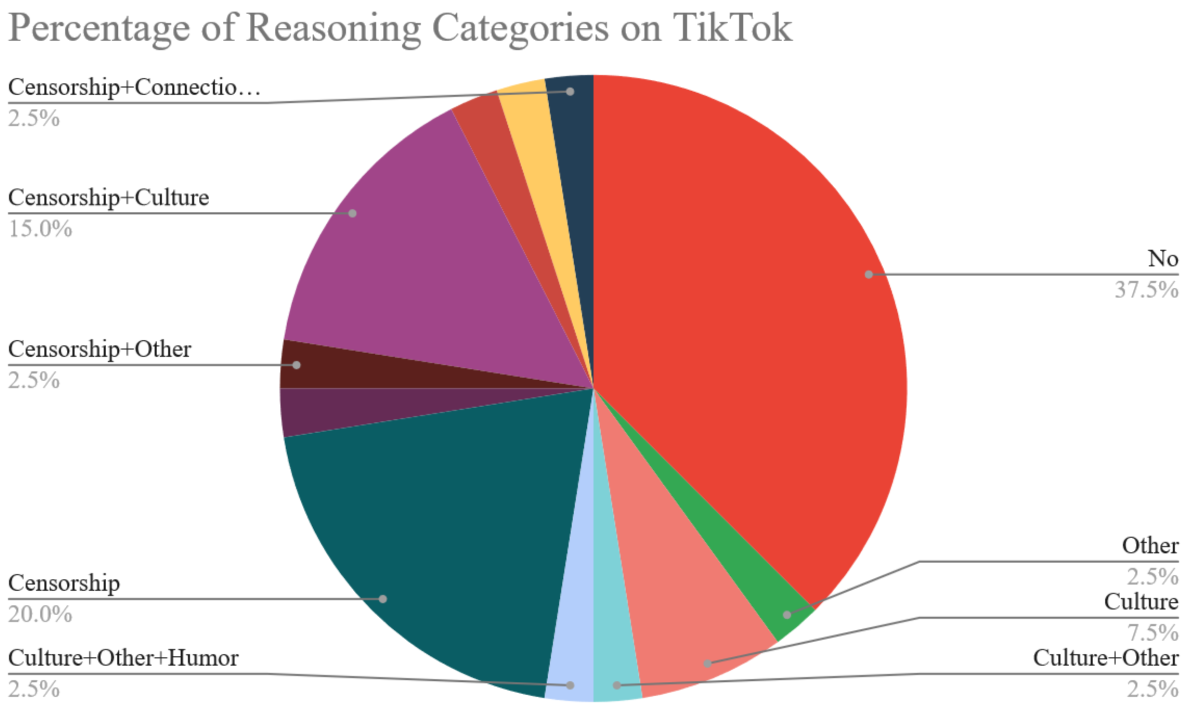

Fig. 9: A pie chart representing the results received for each reasoning category on TikTok

For TikTok, 37.5% of respondents said 'No.' 20% of respondents said only 'Censorship' while 15% said 'Censorship+Culture' and 7.5% said only 'Culture.' The rest of the categories were tied at 2.5%, these being only 'Other,' 'Culture+Other', 'Culture+Other+Humor', 'Censorship+No', 'Censorship+Other', 'Censorship+Connection', 'Censorship+Connection +Other+Humor', and 'Censorship+Connection+Culture.'

Interestingly, 'No' was not the most selected response for the total percentages - 'Censorship' was. However, 'No' by itself has the highest percentage of the separated categories because respondents who selected 'No' mostly did not select other responses. On the other hand, 'Censorship' was found both alone and alongside other responses - 'Censorship' and 'Censorship+Culture' had the highest percentages behind 'No'. The prominence of these two categories and the presence of 'Censorship' in some of the other categories explains why 'Censorship' was the highest for the total count. After 'Censorship', 'Culture' is the highest solo category, and appears in other combination categories, explaining its rank in the total.

Our raw and separated data clearly displays what reasonings respondents reported for each site. As such, we now have enough evidence of where these reasonings fall for each site to answer the question, 'What are the most commonly stated reasons for self-censorship on each site?' The majority of Tumblr users, once again, report no self-censoring and thus no reason to do so, but those who do are more likely to self-censor to avoid algorithmic connections than the other options. A good third of Twitter users also report no self-censoring reason at all, but of the two-thirds who do, many seek to avoid algorithmic connections or censorship, although adapting to site culture is not an uncommon reason. For TikTok, nearly 40% reported no self-censoring reason at all. Many of those that do self-censor are seeking to evade censorship, with some looking to adapt to site culture. Before we discuss the implications of these results, however, we need to address the last question.

5.3 Tactics and Reasonings

The third and final question is 'Is there an association between self-censoring reasonings and self-censoring tactics?' To calculate correlation, I employed a chi-squared test on a contingency table where the columns were reasonings and the rows were strategies. Contingency tables are matrices displaying the frequency distribution of the variables. Here, these variables are reasonings and strategies. However, for the first round, I excluded 'No censoring' as an option for each variable to only work with respondents who indicated regular self-censoring. For this round, the possible tactics were 'Asterisk,' 'Symbol', 'SS Euph', 'Diff Euph', and 'Other', which the possible reasonings were 'Censorship', 'Connection', 'Culture', 'Humor', and 'Other'.

For the second round, I allowed responses that included selections 'No censoring' for question 20 and for question 11, but did not examine them as variables, as their intersection would skew the data. For this round, I only investigated 'Asterisk', 'Symbol', 'SS Euph' and 'Diff Euph' as possible tactics and 'Censorship', 'Connection', and 'Culture' as possible reasons. Since chi-square tests work best when every cell has a value above 5, eliminating the low-response categories of 'Other' and 'Humor' narrows the scope of the test and increases the accuracy.

For each round, I investigated only Tumblr, only Twitter, only TikTok, and then all the data combined, including site overlaps and non-standard sites.

5.3.1 Round 1

Tumblr

For Tumblr only, this was the contingency table for the ninety-one responses that did not select 'No censoring':

	Censorship	Connection	Culture	Humor	Other
Asterisk	7	26	9	9	13
Symbol	7	10	6	1	2
SS Euph	4	4	4	2	4
Diff Euph	12	11	4	1	6
Other	5	8	2	4	7

The p-value was 0.275994358291285, which is more than the standard alpha value of 0.05, indicating no correlation.

TikTok

For TikTok only, this was the contingency table for the twenty-three responses that did not select ‘No censoring’:

	Censorship	Connection	Culture	Humor	Other
Asterisk	6	2	4	1	0
Symbol	8	2	6	1	2
SS Euph	11	3	8	1	2
Diff Euph	12	3	9	2	3
Other	1	0	0	0	0

The p-value was 0.9997032923168346, which is more than the standard 0.05, indicating no correlation.

Twitter

For Twitter only, this was the contingency table for the 39 responses that did not select ‘No censoring’:

	Censorship	Connection	Culture	Humor	Other
Asterisk	12	14	6	2	4
Symbol	10	12	5	2	5
SS Euph	11	11	6	0	3
Diff Euph	8	11	7	0	1
Other	2	3	1	0	0

The p-value was 0.948952048608702, which is more than the standard 0.05, indicating no correlation.

All

For all the data combined, this was the contingency table for the 188 responses that did not select ‘No censoring’:

	Censorship	Connection	Culture	Humor	Other
Asterisk	32	45	27	13	20
Symbol	31	27	20	5	11
SS Euph	31	20	20	4	10
Diff Euph	40	30	26	5	13
Other	11	13	4	4	10

The p-value was 0.345849465097053, which is more than the standard 0.05, indicating no correlation.

5.3.2 Round 2

Tumblr

For Tumblr only, this was the contingency table for the full 1116 responses, investigating a limited number of variables:

	Censorship	Connection	Culture
Asterisk	14	46	21
Symbol	9	18	10
SS Euph	7	9	13
Diff Euph	13	19	11

The p-value was 0.21448928095857195, which is more than the standard 0.05, indicating no correlation.

TikTok

For TikTok only, this was the contingency table for the full forty responses, investigating a limited number of variables:

	Censorship	Connection	Culture
Asterisk	7	2	4
Symbol	8	2	6
SS Euph	11	3	8
Diff Euph	14	3	9

The p-value was 0.9996213883994364, which is more than the standard 0.05, indicating no correlation.

Twitter

For Twitter only, this was the contingency table for the full ninety-six responses, investigating a limited number of variables:

	Censorship	Connection	Culture

Asterisk	13	15	8
Symbol	12	13	8
SS Euph	12	11	8
Diff Euph	12	12	12

The p-value was 0.1428390965577175, which is more than the standard 0.05, indicating no correlation.

All

For all the data combined, this was the contingency table for the full 1380 responses, investigating a limited number of variables:

	Censorship	Connection	Culture
Asterisk	42	68	42
Symbol	36	37	25
SS Euph	36	27	31
Diff Euph	48	41	39

The p-value was 0.16391047115833407, which is more than the standard 0.05, indicating no correlation.

5.3.3 Summary

Overall, for every test in every round, the p-value was more than 0.05, indicating no statistical correlation in any situation. As such, the answer to my main question ‘Is there an association between self-censoring reasonings and self-censoring tactics?’ is a definitive no.

With all of my questions answered, we can move on to analyzing what exactly the answers mean, and what we can learn from them.

6 Analysis, Limitations and Future Work

I set out to explore my main question by first investigating two auxiliary questions through percentage observation and then using a chi-squared test to answer the main question. Each question investigated some intersection of three variables: self-censoring tactics, reasons for self-censoring, and social media sites. The first question examined what tactics users employ on which sites, the second question examined what reasons were listed for which sites, and the main question examined whether tactics and reasons were correlated. The auxiliary questions rely on observation rather than statistical analysis, and as such I cannot use them as definitive evidence for or against my main question. However, they do provide insight into the variables I investigated, and point towards other questions that deserve further exploration.

6.1 Tactics and Sites

As noted above, the answer to the question 'What self censoring tactics are used most on which sites?' is clear. The majority of Tumblr respondents reported no censoring, while a small percentage use asterisk replacements and other tactics are low. Some Twitter users reported no censoring, but there appeared to be an even spread of asterisk replacement, symbol or number replacement, phonetic X-phemisms, and nonphonemic X-phemisms among those who reported censoring. TikTok users mostly do self-censor, and many use nonphonemic X-phemisms, but the other tactics are not rare. As such, there are definitive patterns of tactic usage - or lack of tactic usage, for Tumblr - that differ among the three sites.

It appears that users who selected 'No' were less likely to select another response than users who selected another tactic - 'No' had less overlap than other options. This makes sense, as many who select 'No censorship' would be less likely to select a self-censorship tactic, given they have indicated no censorship. However, 'No' did appear alongside other responses sometimes. This may be because the survey question used for this question asked respondents about patterns used by "you/other users of your primary site," meaning that they could be responding based on behavior patterns from both themselves and other users. Alternatively, they may be indicating that they rarely self-censor, but when they do they use a specific tactic. However, this is pure speculation, as reasoning for responses was not recorded.

While illuminating, these results rely on observation rather than statistical analysis and require further investigation before any definitive claims are possible. This preliminary investigation points to a number of questions. Firstly, is there a statistical correlation between social media sites and self-censoring tactics? Answering this question definitely would guide the rest of the questions and could incite further research itself.

Secondly, do users use the same tactics for all words, or do different categories of words garner different tactics? I did collect data on different words and the self-censoring tactics used in questions 12 through 15, these being ‘sex’, ‘lesbian’, ‘commit suicide’, and a controversial name. However, I did not divide these words into distinct categories, and their investigation was outside the scope of this particular paper. Nonetheless, further study is possible.

Finally, if users on different sites do indeed use different tactics, why? Do the algorithms on one site catch one tactic and not another? Is it a result of site culture, where one tactic is perpetrated over others because it has been in use the longest? Did some users bring their self-censoring tactics from older internet communities, and the practice spread? There are innumerable answers and a wide array of avenues to explore for this question, but my next auxiliary question incites additional avenues as well.

6.2 Reasonings and Sites

Once again, the answer to this question, ‘What are the most commonly stated reasons for self-censorship on each site?’, is clear. Tumblr users mostly report no self-censoring, but those who do lean towards avoiding algorithmic connections, with all other reasons lower. Around a third of Twitter users also do not self-censor, but of the rest, many are trying to avoid algorithmic connections or censorship, and some are trying to adapt to site culture. For TikTok, many also reported no self-censoring, but many those who do self-censor are seeking to evade censorship, while others are trying to adapt to site culture.

Like before, users who selected ‘No’ were less likely to select another response than users who selected another reasoning, likely because respondents who selected ‘No’ would not have a reason to self-censor because they do not do so. However, there were some overlaps, and for this question the survey question used only asked about personal behavior. As before, they may be indicating that while they rarely self-censor, when they do, they do so for a specific reason, but this is merely a possible reason with no support.

As noted in the results section, there was no way to separate ‘Humor’ from ‘Other’ for the separated results. While unfortunate, neither ‘Humor’ nor ‘Other’ were common responses, and even when the ‘Other’ section included both ‘Humor’ responses and ‘Other’ responses, it did not rival other reasons. I should also note that ‘Humor’ and ‘Other’ required write-in responses.

If ‘Humor’ had been a selectable response, it is possible it would have garnered more responses, but this is unknown.

Once again, while there appear to be patterns of behavior, these results are based on percentage observation. The first step to determining if there is a correlation between social media sites and reasons for self-censoring would be a true statistical analysis. From there, however, several other questions emerge.

For one, are these reasons true for all kinds of words? For instance, someone may self-censor words related to self-harm because they see others doing so, but may self-censor words related to sexuality because they fear the site would ban their content. As mentioned above, I did ask questions about distinct categories of words, but I did not ask why people would censor them, only how. As such, this question remains unexplored.

Another important question is from where precisely these reasons stem. As discussed in the background section, many users believe that the algorithms or administration of sites will censor them or force unwanted connections. How does the site functionality influence these beliefs? What about site culture, which may perpetuate either the beliefs or the desire to self-censor? This question would require further research into not only people’s self-described habits, but the history of self-censoring on the internet and the inner functionality of the social media sites.

With the auxiliary questions examined, we have enough information to move on to the main question.

6.3 Tactics and Reasonings

As mentioned before, the answer to the question ‘Is there an association between self-censoring reasonings and self-censoring tactics?’ is a definitive no. For each of the tests in both rounds, we found no statistical correlation. Each test covered a number of angles.

For the first round, we only examined respondents who indicated that they self-censor. This reduced the scope to only look at people who do self censor, rather than those who observe self-censorship on other sites. One of the survey questions that this question draws on, question 11, asked survey takers about patterns used by “you/other users of your primary site”, meaning that this question incorporates both personal practice and observations about the site as a whole. This may have encouraged a wider view of the site as a whole, as respondents could report both, but also distances the responses from the reasons self-censoring was done. As such, respondents who did not self-censor may have reported their observations here, but since they themselves do not self-censor, could not report the reasoning behind the self-censoring. Excluding these respondents narrowed the scope of the first round, but also significantly reduced the sample size, especially for Tumblr.

For the second round, we allowed respondents who selected ‘No censoring’ for either of the variables. This widened the scope of the tests by allowing a higher sample size. However, those options were not included in the variable list for testing, since there is an obvious correlation between those who selected ‘No censoring’ for both questions. A preliminary test run including those options produced a very small p-value, indicating high correlation, but when run without those options the resulting p-value was not above 0.05. Since including them would have massively skewed the results, they were excluded.

I also eliminated ‘Other’ and ‘Humor’ as options for the reasoning variable and ‘Other’ as an option for the tactic variable. Chi-square tests perform optimally when every cell has a value above 5, and responses for the above options skewed low. As such, eliminating these options both narrowed the scope of the test from five variables to three and four variables and increased the accuracy of the test for round two.

However, both rounds are flawed in one significant aspect - overlapping data. Chi-squared tests are not the optimal tool for analyzing such data, and typically require mutually exclusive variables. However, separating the variables into distinct categories, as was done for the auxiliary questions, would result in 63 categories for each variable, for a total of 3969 cells in each contingency table. Most of these cells would be empty or contain a value below 5, making a chi-squared test entirely ineffective. Additionally, most statistical tests are not built to analyze overlapping data, as variables must be independent and mutually exclusive. As such, I could find no tool better than a chi-squared test for my data.

As it stands, the results of both rounds are flawed but illuminating. Firstly, the p-values of the second-round were generally smaller than those of the equivalent tests in the first round, indicating that the changes made for the second round by increasing the sample size but narrowing the scope of variables likely helped highlight any slight trends hidden in the first round. This is especially true for Twitter, whose p-value went from 0.949 in the first round to 0.143 in the second. While still not statistically significant, this is a major drop.

Second, I had expected the p-value for the test including all sites to be lower than the separated sites, since in the first and second auxiliary questions there appeared to be different trends towards specific tactics and reasons on different sites. As such, it stood to reason that if, for instance, both ‘Diff Euph’ and ‘Censoring’ were common choices for TikTok and not the other sites, that could indicate a correlation between those two. However, while the p-values for these tests were generally lower than some of the site specific tests, they never were low enough to indicate statistical significance. This indicates that even without separating the data into sites to account for possible skewing, there is no correlation between self-censoring tactics and reasons.

Future work on this question would necessarily need to tread the same ground covered here, as the statistical tests performed were flawed. The easiest fix for this would be to simply have the relevant survey questions take only one choice, rather than many. However, as shown by the survey responses, many respondents use some combination of tactics or self-censor for multiple reasons. As such, the questions themselves would need to be reframed as well. Since the results of this question indicate no correlation between the two variables, the next step is investigating what does influence these variables, as mentioned in the previous sections.

I have covered limitations of each individual question and the answers they produced, but there are other factors that may have influenced the research that require discussion.

6.4 Other

While not strictly relevant to the research questions, the demographics of survey respondents may have also influenced the results. The full demographic data can be found in the appendix, but there are some specifics that should be noted here, as they may limit the scope of the results.

For all social media sites, the highest answer for gender was ‘Female’, then ‘Other’, then ‘Male’, and finally ‘Prefer not to say’. This shows that the data skews towards female respondents and nonbinary respondents over male respondents. For all sites, the majority of respondents identified as white. Twitter had the highest percentage of Asian respondents, at 11.46%, while 15% of TikTok identified as multiracial. All other categories were below 10%. The majority of respondents were only native English speakers, with a small percentage being both native English speakers and native speakers of another language, and an even smaller amount being native speakers of another language only.

For all sites, over half of the respondents were in the 21 to 29 age range, but TikTok respondents skewed especially young, with none being above 40 and nearly 30% being 18 to 20. Twitter had the highest amount of respondents in the 40 to 49, 50 to 59, and 60 or older ranges, but none of these ranges had more than 6%. Over half of respondents for all the sites started using the internet between 2000 and 2010. TikTok respondents skewed later than other sites, with only 10% between 1991 and 2000 and nearly 30% between 2011 and 2020. In contrast, the other sites had around a third between 1991 and 2000 and less than 20% between 2011 and 2020. The distribution of what social media sites respondents used first followed similarly - Twitter users skewed towards earlier sites, TikTok users skewed towards later ones, and Tumblr stayed in the middle.

Overall, while my main question appears to have been answered in the negative, the lack of appropriate statistical tools for the data renders the results flawed. Besides the statistical tools, the investigation is limited by the demographic scope of the results, as mentioned above. However, despite these flaws, both the main question and the auxiliary questions deserve further examination and inspire more questions. Do people use different tactics on different categories of words? For that matter, do the reasons people self-censor these categories differ as well? Why do users on different sites appear to use different tactics? And what inspired the various reasonings for self-censoring? While fascinating, these questions are outside the scope of this paper, which has come to its conclusion.

7 Conclusion

In this paper, I had one main question, ‘Is there an association between self-censoring reasonings and self-censoring tactics?,’ and two auxiliary questions, ‘What self censoring tactics are used most on which sites?’ and ‘What are the most common stated reasons for self-censorship on each site?’ In order to learn more about those sites, self-censoring reasons, and self-censoring tactics, I delved into the rules, functionality and culture of the social media sites Tumblr, TikTok, and Twitter, researched the function and effect of taboos, investigated terms and taxonomies surrounding self-censoring practices to create my own categorization scheme, and examined how self-censoring can be used in online communication and memes. Then to answer my questions, I conducted a survey. Using that survey data, I performed percentage observations on the data for my auxiliary questions and a chi-square test on the data for my main question.

Finally, I had results. I found no correlation between self-censoring tactics and reasons for self-censoring, proving the answer to my main question to be “No.” There were also significant defects with my statistical analysis, indicating that a more definitive answer would require more research or an alternate analysis. However, through investigation of my auxiliary questions I did observe that different sites do display different patterns of tactics used and reasons for self-censoring. As such, while my main question produced flawed results, it does deserve further examination, and the auxiliary questions point towards new questions.

Unfortunately, the nature of my results means that I have no clear answers for why the trends I observed exist. Fortunately, this means there is still much to explore on this particular topic, and many more papers to write. As conversations and communication happens more and more in online spaces, the nature of these spaces and sites and the possible linguistic pressures they exert become more and more relevant. An examination of why and how our language changes in these spaces is vital for the future of communication, and the subtle shifts in what words we use must be brought to light. Self-censoring deserves our attention.

8 References

Allan, Keith & Kate Burridge. 2006. Forbidden Words: Taboo and the Censoring of Language. *Cambridge University Press*. Print.

astraltricker. 2022. Okay I know this was mostly a joke but I just feel…*Tumblr*. https://astraltrickster.tumblr.com/post/682820467119259648/okay-i-know-this-was-mostly-a-joke-but-i-just-feel (4 November, 2022).

Burridge, Kate. 2012 . Euphemism and language change: The sixth and seventh ages. *Lexis. Journal in English Lexicology*, (7). https://journals.openedition.org/lexis/355#citedby (3 October, 2022).

bundibird. 2022. Hi all -- newbies in particular, and newbies from tiktok extra in particular...*Tumblr*. https://www.tumblr.com/bundibird/680384295193477120/hi-all-newbies-in-particular-and-newbies-from?source=share (4 November, 2022).

Chayka, Kyle. 2022. How Tumblr Became Popular for Being Obsolete. *The New Yorker*. https://www.newyorker.com/culture/infinite-scroll/how-tumblr-became-popular-for-being-obsolete (6 December, 2022).

Cho, Won Ik & Soomin Kim. 2021. Google-trickers, Yaminjeongeum, and Leetspeak: An Empirical Taxonomy for Intentionally Noisy User-Generated Text. *Proceedings of the Seventh Workshop on Noisy User-generated Text* (W-NUT 2021), 56–61. Association for Computational Linguistics. https://aclanthology.org/2021.wnut-1.7.pdf (15 September, 2022.)

Day, Faithe J. 2021. Are Censorship Algorithms Changing TikTok's Culture? Medium. https://onezero.medium.com/are-censorship-algorithms-changing-tiktoks-culture-17f7912e0064 (4 November, 2022)

Delkic, Melina. 2022. Leg Booty? Panoramic? Seggs? How TikTok Is Changing Language. New York Times. https://www.nytimes.com/2022/11/19/style/tiktok-avoid-moderators-words.html (25 November, 2022).

Dias Oliva, Thiago, Dennys Marcelo Antonialli & Alessandra Gomes. 2021. Fighting Hate Speech, Silencing Drag Queens? Artificial Intelligence in Content Moderation and Risks to LGBTQ Voices Online. *Sexuality & Culture 25, 700–732* https://doi.org/10.1007/s12119-020-09790-w (4 November, 2022)

etakeh. 2022. I'm sure this has gone around before, but…*Tumblr*. https://etakeh.tumblr.com/post/680836214877798400/im-sure-this-has-gone-around-before-but (14 January 2023).

Eugene. 2013. Intro to Voldemorting. *A Life in Juxtaposition*. http://skepticmystic.blogspot.com.au/2013/02/into-to-voldemorting.html (3 October, 2022)

Fishbein, Rebecca. 2022. Welcome to Tumblr. Now Go Away. *New York Times*. https://www.nytimes.com/2022/11/23/style/tumblr-twitter-elon-musk.html (6 December, 2022).

Gorwa, Robert, Reuben Binns & Christian Katzenbach. 2020. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society, 7(1)*. https://doi.org/10.1177/2053951719897945 (4 November 2022)

Mashable SEA. 2022. Despite censorship controversy, Tumblr creators remain loyal to their favorite hellsite. *Mashable SE Asia*. https://sea.mashable.com/tech/19341/despite-censorship-controversy-tumblr-creators-remain-loyal-to-their-favorite-hellsite (12 January 2023).

McCulloch, Gretchen. 2020. Because internet: Understanding the new rules of language. *Penguin*. Print.

Nagel, Emily van der. 2018. 'Networks that work too well': intervening in algorithmic connections. *Media International Australia*, 168(1), 81-92. https://journals.sagepub.com/doi/full/10.1177/1329878X18783002 (15 September, 2022.)

pockopeas. 2022. its so funny to me that people on twitter n tiktok…*Tumblr*. https://pockopeas.tumblr.com/post/682602207280054272 (4 November, 2022).

Ryan, Fergus, et al. 2020. TikTok Censorship. TikTok and WeChat: Curating and Controlling Global Information Flows, Australian Strategic Policy Institute pp. 04–24. http://www.jstor.org/stable/resrep26120.5. (4 November, 2022)

Smith, Katie Louise. 2018. Twitter Is Deleting Accounts And These Are The Words That Might Get You Suspended. *PopBuzz*. Global Media Group Services Limited. https://www.popbuzz.com/internet/social-media/twitter-account-suspension-trigger-words/ (4 November, 2022).

TikTok. 2022. Community Guidelines. *TikTok*. https://www.tiktok.com/community-guidelines?lang=en (25 November, 2022).

Twitter Help. 2022. Violent threats policy. *Twitter*. https://help.twitter.com/en/rules-and-policies/violent-threats-glorification (4 November, 2022).

Twitter Help. 2022. Suicide and Self-harm policy. *Twitter*. https://help.twitter.com/en/rules-and-policies/glorifying-self-harm (4 November, 2022).

Twitter Help. 2023. About your Home timeline on Twitter. *Twitter*. https://help.twitter.com/en/using-twitter/twitter-timeline (12 January 2023).

Tumblr. 2022. Community Guidelines. *Tumblr*. https://www.tumblr.com/policy/en/community (6 December, 2022).

Yus, Francisco. 2005. Attitudes and emotions through written text: the case of textual deformation in internet chat rooms. *Pragmalingüística 13: 147-173.* https://revistas.uca.es/index.php/pragma/article/view/120/131 (27 October, 2022).

Yus, Francisco. 2018. Identity-related issues in meme communication. *Internet Pragmatics 1.1: 113-133.*https://rua.ua.es/dspace/bitstream/10045/89247/1/2018_Francisco-Yus_InternetPragmatics_preprint.pdf (27 October, 2022)

Wu, Xiaoping & Richard Fitzgerald. 2020. 'Hidden in plain sight': Expressing political criticism on Chinese social media. *Discourse Studies*, 23(3), 365-385. https://journals.sagepub.com/doi/10.1177/1461445620916365 (15 September, 2022.)

A Appendix

A.1 Data Collection

Note: bullet points indicate only one option could be selected, while boxes indicate multiple options could be selected.

A.1.1 Consent Form

This is a survey gathering linguistic data for a senior thesis conducted by Talia Feshbach. The purpose of the study is to determine how, where, and why users self-censor on social media sites, specifically Tumblr, Twitter, and TikTok. Participants are found or selected by their use of social media websites, including but not limited to those listed. As the survey is public, participation is generally self-selected. I anticipate 100 participants, but may receive more or less. For legal and logistical reasons, this survey is only open to people above 18 years of age and residents or citizens of the United States of America. This survey will remain open until November 15, 2022.

Participation in this study only involves filling out the 20 question survey. This will take approximately 5 minutes. A question about the consent form will be included in the survey to ensure comprehension. Names, emails, and other identifying information will not be collected, as the survey is entirely anonymous. Some demographic questions will be asked, but this is for the purpose of data analysis, and there will be no way to use answers to link specific responses to specific individuals. The data is being collected using a secure connection to the host survey service provider. Results are stored in a password protected account accessible by only the researchers and system administrators. While no absolute guarantees can be made regarding security, these measures provide safeguards against outside agents accessing the electronic data.

There are some questions mentioning words and euphemisms for sex, suicide, death, and drug use. These topics are not described in depth or explored beyond the words used to refer to them. If you anticipate this would cause psychological discomfort, I suggest you do not complete the survey. If you experience unanticipated discomfort or long-lasting distress as a result of the survey and are comfortable breaking anonymity, then please be in touch with me and I will provide some suggestions about who to talk to. If you would prefer to remain anonymous, please call your local or national mental health hotline.

There are no direct benefits to you for participating in this research. However, you may find it interesting to talk about the issues addressed in the research, and it may be beneficial to the field of linguistics. There will be no compensation for participating in this research. There is no deception used in this study.

Your participation is completely voluntary. You can withdraw from the study at any time. You do not have to answer any questions that you don't want to answer. If you choose not to participate, there will be no penalty for not participating.

If you have any questions about the research, please feel free to call or email the Principal Investigator, Talia Feshbach, at tfeshbach@brynmawr.edu, or the student's supervisor, Jane Chandlee, and jchandlee@haverford.edu. If you have questions about your rights as a research participant, please contact the Chair of the Bryn Mawr College IRB (irb@brynmawr.edu). You can find a copy of this consent form here: https://docs.google.com/document/d/1oLXCf87ZGLOor3BOObudGqzn6dIGgoqc4wSIAtvnU-4/edit?usp=sharing

By clicking yes, you certify that you are 18 or older, a US Citizen and/or resident, have read this consent form or it has been read to you, have had all your questions answered to your satisfaction, have been provided a copy of the consent form, and have agreed to participate in research.

- Yes
- No

A.1.2 Initial Survey

1. How would you describe your gender?

- Male
- Female
- Other___
- Prefer not to respond

2. Are you White, Black or African-American, American Indian or Alaskan Native, Asian, Native Hawaiian or other Pacific Islander, or some other race?

- White
- Black or African-American
- American Indian or Alaskan Native
- Asian
- Native Hawaiian or other Pacific islander
- From multiple races
- Some other race (please specify)___

3. Are you a native English speaker? If not, please state your native language.

- Yes
- Yes, but I am also a native speaker of __
- No, I am a native speaker of ___

4. What category below includes your age?

- 18-20
- 21-29
- 30-39
- 40-49
- 50-59
- 60 or older

5. When did you first start using the Internet?

- before 1980
- 1980-1990
- 1991-2000
- 2000-2010
- 2011-2020
- after 2020

6. When you first started using the Internet, what social media sites did you use?

- Usenet, forums, IRC, BBS, listservs, or similar
- AIM, MSM Messenger, blogs, LiveJournal, MySpace, or similar
- Facebook, Twitter, Gchat, YouTube, or similar
- Instagram, Snapchat, iMessage, WhatsApp, TikTok, or similar

7. How often do you use Tumblr?

- 1 - Never used it
- 2 - Used it previously/use it infrequently
- 3 - Use it occasionally
- 4 - Use it often
- 5 - Use it constantly/daily

8. How often do you use Twitter?

- 1 - Never used it
- 2 - Used it previously/use it infrequently
- 3 - Use it occasionally
- 4 - Use it often
- 5 - Use it constantly/daily

9. How often do you use TikTok?

- 1 - Never used it
- 2 - Used it previously/use it infrequently
- 3 - Use it occasionally

- 4 - Use it often
- 5 - Use it constantly/daily

10. Choose your primary social media/s:

- Tumblr
- TikTok
- Twitter
- Other ___

11. Do you/other users of your primary site censor words/phrases, e.g. kill, by asterisk replacement (like k*ll), symbol or number replacement (like k!ll or k1ll), euphemisms/misspellings that sound similar (like krill), euphemisms that sound different (like unalive), or other? As a side note, euphemisms in this case are any words or phrases used to refer to a concept without stating it outright.

- Asterisk replacement
- Symbol/number replacement
- Similar-sounding euphemisms/misspellings
- Euphemism that sounds different
- Other ___

12. How would you/other users of your primary site censor the word ‘sex’?:

- Asterisk replacement (like s*x)
- Symbol or number replacement (like s3x)
- Euphemisms/misspellings that sound similar (like seggs)
- Euphemisms that sound different (like intercourse)
- Other ____

13. How would you/other users of your primary site censor the word ‘lesbian’?:

- Asterisk replacement (like l*sbian)
- Symbol or number replacement (like le$bian)
- Euphemisms/misspellings that sound similar (like lessbien)

- Euphemisms that sound different (like sapphic)
- Other ____

14. How would you/other users of your primary site censor the phrase ‘commit suicide’?:

- Asterisk replacement (like commit s*icide)
- Symbol or number replacement (like commit suic!de or commit su1cide)
- Euphemisms/misspellings that sound similar (like kermit super slide)
- Euphemisms that sound different (like unalive themself)
- Word omission (like commit)
- Other ____

15. How would you/other users of your primary site censor the name Voldemort?:

- Asterisk replacement (like V*oldemort)
- Number/symbol replacement (like V0ld3mort)
- A euphemism/misspelling that sounds or looks similar to the name (Moldyvort)
- A euphemism that sounds different from the name (like He Who Shall Not Be Named)
- Other

16. This is the question that checks for comprehension of the consent form. How many participants am I expecting for this survey? If you don’t know, feel free to go back and review the consent form.

- 25
- 50
- 100
- 200

17. You’re texting a friend about a person you both don’t like named John Smith. Would you censor their name? If so, how?

- No, I wouldn’t.
- An asterisk replacement (like J*hn Sm*th)
- Number/ symbol replacement (like J0hn Sm!th)

- A euphemism/misspelling that sounds similar to the name (like Jihn Smoth)
- A euphemism that sounds different from the name (like Example Man)
- Other

18. You’re chatting in a Discord/Slack/other group chat where you know some people well and some people little. You’re talking about a piece of media where someone attempts suicide. Which of these phrases would you say (assuming other people in the chat are aware that the discussion involves suicide)?

- I think she was trying to kill herself
- I think she was trying to commit suicide
- I think she was trying to krill herself
- I think she was trying to k!ll herself
- I think she was trying to unalive herself

19. You’re posting on a new website with unknown moderation/censorship rules about drug use. How would you say ‘weed’?:

- weed
- w33d
- w**d
- wheed (or other misspelling/phonetically similar euphemism)
- mary jane (or some other unrelated euphemism)

20. If you self-censor on your website, why do you do so?

- I am avoiding shadowbanning/censorship from site administration
- I am trying to avoid attention from other users/detection from searches
- I am adopting the terms I see around me to participate in site culture
- Other

A.1.3 Edited Survey

Note: edits are in bold.

1. How would you describe your gender?

- Male
- Female
- Other___
- Prefer not to respond

2. Are you White, Black or African-American, American Indian or Alaskan Native, Asian, Native Hawaiian or other Pacific Islander, or some other race?

- White
- Black or African-American
- American Indian or Alaskan Native
- Asian
- Native Hawaiian or other Pacific islander
- From multiple races
- Some other race (please specify)___

3. Are you a native English speaker? If not, please state your native language.

- Yes
- Yes, but I am also a native speaker of __
- No, I am a native speaker of ___

4. What category below includes your age?

- 18-20
- 21-29
- 30-39
- 40-49
- 50-59
- 60 or older

5. When did you first start using the Internet?

- before 1980
- 1980-1990

- 1991-2000
- 2000-2010
- 2011-2020
- after 2020

6. When you first started using the Internet, what social media sites did you use?

- Usenet, forums, IRC, BBS, listservs, or similar
- AIM, MSM Messenger, blogs, LiveJournal, MySpace, or similar
- Facebook, Twitter, Gchat, YouTube, or similar
- Instagram, Snapchat, iMessage, WhatsApp, TikTok, or similar

7. How often do you use Tumblr?

- 1 - Never used it
- 2 - Used it previously/use it infrequently
- 3 - Use it occasionally
- 4 - Use it often
- 5 - Use it constantly/daily

8. How often do you use Twitter?

- 1 - Never used it
- 2 - Used it previously/use it infrequently
- 3 - Use it occasionally
- 4 - Use it often
- 5 - Use it constantly/daily

9. How often do you use TikTok?

- 1 - Never used it
- 2 - Used it previously/use it infrequently
- 3 - Use it occasionally
- 4 - Use it often
- 5 - Use it constantly/daily

10. Choose your primary social media/s:

- Tumblr
- TikTok
- Twitter
- Other ___

11. Do you/other users of your primary site censor words/phrases, e.g. kill, by asterisk replacement (like k*ll), symbol or number replacement (like k!ll or k1ll), euphemisms/misspellings that sound similar (like krill), euphemisms that sound different (like unalive), or other? As a side note, euphemisms in this case are any words or phrases used to refer to a concept without stating it outright.

- Asterisk replacement
- Symbol/number replacement
- Similar-sounding euphemisms/misspellings
- Euphemism that sounds different
- Other ___
- **No censoring**

12. How would you/other users of your primary site censor the word ‘sex’?:

- Asterisk replacement (like s*x)
- Symbol or number replacement (like s3x)
- Euphemisms/misspellings that sound similar (like seggs)
- Euphemisms that sound different (like intercourse)
- Other ____
- **No censoring**

13. How would you/other users of your primary site censor the word ‘lesbian’?:

- Asterisk replacement (like l*sbian)
- Symbol or number replacement (like le$bian)
- Euphemisms/misspellings that sound similar (like lessbien)

- Euphemisms that sound different (like sapphic)
- Other ____
- **No censoring**

14. How would you/other users of your primary site censor the phrase ‘commit suicide’?:

- Asterisk replacement (like commit s*icide)
- Symbol or number replacement (like commit suic!de or commit su1cide)
- Euphemisms/misspellings that sound similar (like kermit super slide)
- Euphemisms that sound different (like unalive themself)
- Word omission (like commit)
- Other ____
- **No censoring**

15. How would you/other users of your primary site censor **a controversial name - for example, Voldemort**

- Asterisk replacement (like V*ldemort)
- Number/symbol replacement (like V0ld3mort)
- A euphemism/misspelling that sounds or looks similar to the name (Moldyvort)
- A euphemism that sounds different from the name (like He Who Shall Not Be Named)
- Other
- **No censoring**

16. This is the question that checks for comprehension of the consent form. How many participants am I expecting for this survey? If you don’t know, feel free to go back and review the consent form **here: https://docs.google.com/document/d/1oLXCf87ZGLOor3BOObudGqzn6dIGgoqc4wSIAtvnU-4/edit?usp=sharing**

- 25
- 50
- 100

- 200

17. You're texting a friend about a person you both don't like named John Smith. Would you censor their name? If so, how?

- No, I wouldn't.
- An asterisk replacement (like J*hn Sm*th)
- Number/ symbol replacement (like J0hn Sm!th)
- A euphemism/misspelling that sounds similar to the name (like Jihn Smoth)
- A euphemism that sounds different from the name (like Example Man)
- Other

18. You're chatting in a Discord/Slack/other group chat where you know some people well and some people little. You're talking about a piece of media where someone attempts suicide. Which of these phrases would you say (assuming other people in the chat are aware that the discussion involves suicide)?

- I think she was trying to kill herself
- I think she was trying to commit suicide
- I think she was trying to krill herself
- I think she was trying to k!ll herself
- I think she was trying to unalive herself

19. You're posting on a new website with unknown moderation/censorship rules about drug use. How would you say 'weed'?:

- weed
- w33d
- w**d
- wheed (or other misspelling/phonetically similar euphemism)
- mary jane (or some other unrelated euphemism)

20. If you self-censor on your website, why do you do so?

- I am avoiding shadowbanning/censorship from site administration

- I am trying to avoid attention from other users/detection from searches
- I am adopting the terms I see around me to participate in site culture
- Other
- **No censoring**

A.2 Demographic Data

Gender	**Tumblr**	**TikTok**	**Twitter**	**All**
Male	10.75%	17.50%	17.71%	11.52%
Female	50.72%	45.00%	52.08%	51.45%
Other	32.89%	37.50%	25.00%	31.81%
Prefer not to say	5.65%	0.00%	5.21%	5.22%

Race	**Tumblr**	**TikTok**	**Twitter**	**All**
White	83.32%	85.00%	78.13%	82.73%
Black or African-American	0.99%	0.00%	1.04%	0.87%
American Indian or Alaska Native	0.63%	0.00%	0.00%	0.58%
Asian	4.57%	0.00%	11.46%	5.30%
Native Hawaiian or Pacific Islander	0.09%	0.00%	0.00%	0.07%
From multiple races	7.98%	15.00%	8.33%	8.27%
Some other race (please specify)	2.42%	0.00%	1.04%	2.18%

Language	**Tumblr**	**TikTok**	**Twitter**	**All**

Native English Speaker	93.63%	97.50%	92.71%	93.84%
Native English and Other Language Speaker	4.66%	0.00%	4.17%	4.50%
Not Native English Speaker	1.70%	2.50%	3.13%	1.67%

Age	**Tumblr**	**TikTok**	**Twitter**	**All**
18-20	15.14%	27.50%	8.33%	15.22%
21-29	55.11%	62.50%	54.17%	53.33%
30-39	23.84%	10.00%	25.00%	24.42%
40-49	4.30%	0.00%	5.21%	4.49%
50-59	1.34%	0.00%	4.17%	1.81%
60 or older	0.27%	0.00%	3.13%	0.72%

Year Started Using Internet	**Tumblr**	**TikTok**	**Twitter**	**All**
before 1980	0.00%	0.00%	0.00%	0.00%
1980-1990	1.79%	0.00%	3.13%	2.03%
1991-2000	28.43%	10.00%	35.42%	28.64%
2000-2010	55.52%	62.50%	54.17%	55.40%
2011-2020	14.26%	27.50%	7.29%	13.92%
after 2020	0.00%	0.00%	0.00%	0.00%

First Social Media	Tumblr	TikTok	Twitter	All
Usenet, forums, IRC, BBS, listservs, or similar	10.58%	0.00%	15.79%	11.26%
AIM, MSM Messenger, blogs, LiveJournal, MySpace, or similar	44.94%	27.50%	50.53%	45.10%
Facebook, Twitter, Gchat, YouTube, or similar	40.96%	55.00%	31.58%	39.99%
Instagram, Snapchat, iMessage, WhatsApp, TikTok, or similar	3.53%	17.50%	2.11%	3.65%