# The Moderation of Contentious Content on Twitter

by

Wei Hu

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Mathematics
in
Computer Science

Waterloo, Ontario, Canada, 2023

## Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

## Statement of Contributions

A portion of this thesis incorporates material from a workshop paper [38] for which I was first author. Some content, particularly in the introduction and in the section discussing self-moderated topics, is taken verbatim.

## Abstract

Retweeting posts is Twitter's most important feature, playing a vital role in enabling the platform to be a virtual town hall that fosters timely discussions. This attribute has been instrumental in drawing a younger, wealthier, and more educated user-base, distinguishing Twitter from its competitors. We were motivated by the observation that the retweet count on popular tweets diminishes over time. In particular, this reduction is greater for contentious tweets. Since, retweets represent endorsements, it is pertinent to understand how self-moderation and platform moderation play a role in their retractions.

We collected our own datasets and tracked various reasons for retweet loss over time. Leveraging Kaggle datasets, we trained models to predict which tweets would see a significant decrease in retweets; the model's performance extended to previously unseen datasets. Additionally, we proposed an algorithm to estimate the timeline of retweet loss and explored factors that contribute to individual unretweeting behaviour. Finally, our data collection period coincided with the volatile phase on Twitter following Elon Musk's acquisition. As a result, we were able to observe the impact of various changes in platform moderation through our analysis.

## Acknowledgements

First, I would like to thank Professor Kate Larson for the freedom, patience and support she extended me throughout the process. Her flexibility and encouragement afforded me the opportunity to explore my own interests, which made this experience most rewarding. I am particularly grateful for her thoughtful feedback, insights into proper research methodology, and gentle guidance, as they were instrumental in bringing the thesis together while ensuring that it remained an authentic reflection of my own voice.

I thank Professor Diogo Barradas for teaching an excellent course on censorship, and for first introducing me to the subject. I am grateful for his dedication both during and after the course, and thankful for his writing and presentation tips. I would also like to thank him along with Professor Robin Cohen for agreeing to review this thesis.

Finally, I would like to thank my partner Allyson for her encouragement, companionship and continual support throughout the journey.

# Table of Contents

# List of Figures

# List of Tables

## Quote

“If liberty means anything at all, it means the right to tell people what they do not want to hear.”

– George Orwell

# Chapter 1

# Introduction

Twitter is an important source of information across a variety of domains, ranging from understanding political polarization [18], capturing stock behaviours [19], detecting sport fan-bases [73], finding market influencers [61], predicting music popularity [21], to identifying significant public events [54, 37].

As a point of differentiation to previous multi-purpose social networking platforms, Twitter presents itself mainly as a town square where views are shared freely, and the debate of ideas is encouraged [89]. There is indication that the platform includes feedback mechanisms that urges users to use a confrontational tone [58]. This is in part due to the constraints Twitter sets on Tweet length, originally capped at 140 characters (later doubled to be 280), which encourages users to be terse. However, the constraint also nudges users to be more timely in their responses. Thus, the platform is often used to express real-time commentary on personal and news events [45]. While other social media platforms, like Facebook, often carry reposts of news, Twitter is seen as the platform where news breaks first [63, 84]. For example, financial news which has traditionally been spread through press releases and regulatory disclosures, is now being initially discovered on Twitter [23]. Furthermore, Twitter has changed how journalists view their jobs. Through the platform, journalists now report on a broader range of topics, and moreover post Tweets to give story updates and participate in online discussions with users [81].

Furthermore, Twitter has become an important component of many organisations' communication strategies whether it be coffee-houses [82] or political candidates [12, 64]. It also plays an important role in global political events, rising to prominence in the Arab Spring protests [52]. For corporations, Twitter users are a desirable demographics to target, because they are younger, more educated, and have higher income [89] than Americans in general and users of other social media platforms, such as Facebook [91].

A key component of Twitter's differentiation and its appeal to users seeking open discussion is its emphasis on *Retweets*. Retweeting allows a user to conveniently share another author's content and have it be associated with his/her profile. It can be seen as a form of endorsement, and it allows users to align themselves with a group's opinion instead of articulating their own beliefs [54]. As the platform evolved, the interface was refined to make Retweeting easier to do [74, 14]. In short, Retweets are a scarce commodity and serve as the most important mean of information propagation [85, 53].

Given Twitter's role as a source of truth for many of its users, there has been significant public debate over the balance between moderation and free speech on Twitter. My thesis will explore the moderation of contentious content on Twitter through the lens of its most important mechanism: Retweets. We will examine how Retweet counts and Retweeting behaviours are affected by platform moderation, as well as self-moderation.

We are motivated by the observation that popular Tweets experience a decrease in Retweets over time. In this context, we are referring to a decline in the total count of Retweets received, as opposed to the rate at which new Retweets are acquired, since the latter is expected as Tweets gradually become less relevant. While this Retweet loss is also to be expected because Retweeting users occasionally leave the platform and deactivate their account, no doubt the platform's moderation and the users' self-moderation also play an important role. The significance of moderation becomes more apparent when we observe that contentious posts experience a greater reduction in Retweets compared to uncontentious posts. Differences in the extent of self-moderation is possibly a factor, since the act of manually unretweeting represents unendorsing an opinion and disassociating it from the user profile.

## 1.1 Related Works

Broadly speaking, there are two forms of content moderation: platform moderation, and self-moderation. The former refers to how social media employees and the platform's automated tooling interferes with content deemed unsuitable. The latter refers to users manually retracting content or holding back from posting content in the first place.

We first introduce some works that study the impacts and trade-offs of stricter moderation by social media platforms. From there, we explore self-moderation by examining the underlying causes, and describing past approaches to measure its prevalence. Lastly, we look at existing studies on deleted social media content, such as posts users retracted.

### 1.1.1 Social Media Platform Moderation

Social media platforms are increasingly taking a stance on what to censor (e.g., banning a sitting President's accounts [3]), and applying soft moderation (e.g., attaching warning labels to users' posts that question elections' integrity [92] or vaccines' side-effects [75]). While there are obvious differences between high-profile public figures using the platform to create chaos and the average person using it to exercise freedom of speech, it remains unclear to what extent users are still comfortable posting contentious opinions on Twitter.

Following the January 6th Capitol Attack, the moderation of content on Twitter has become a heavily politicized issue [3], and the moderation of misinformation [30, 70] and fake accounts [86] on Twitter is under heavy scrutiny. There are claims of platforms demonstrating a liberal bias when it comes to moderation, though Novak et al. [57] notes some political asymmetries in enforcement are to be expected. Meanwhile, some emerging social media platforms have taken different stance; these competitors to Twitter such as Gab, Parler, and Truth Social orient themselves as free-speech focused alternatives.

Parler claims to "champions free speech, individual liberty, and the free flow of information online," [39] while Gab puts "people and free speech first" and appeals to users banned elsewhere [93]. Truth Social, a platform by former President Trump, positions itself against"Big Tech" censorship [27]. Israeli and Tsur [39] notes these platforms have recently seen significant growth in users, which is ascribed to increased moderation on mainstream platforms. There are previous works noting concerns about the relationship between decreased moderation and increase in hate speech, which presents arguments that these platforms may be alt-right echo-chambers, rather than free-speech town-halls [39, 93, 87]. Furthermore, it is known that platforms need some degree of moderation to be functional, and even Truth Social sees the need to moderate [27].

While it is unclear to what extent platforms should interfere with hate speech, it is evident that a platform's finances are impacted by those decisions. For example, they must balance the threat of competition against being perceived as not being advertiser-friendly [55]. Elon Musk's acquisition of Twitter in late 2020 was centered around moderation policy. Musk is a self-prescribed free-speech absolutist by principle; his motivations seem to stem from wanting to shape public opinion with fewer restrictions and maximize freedom on the platform [43] more so than his optimism about the platform's business potential.

Musk marketed his acquisition as an attempt to increase transparency in the moderation process and to make those algorithms open-source. Following his acquisition, there were multiple changes to moderation, such as increased automation in the moderation process [66]. While an ensuing increase in hate speech [10] and a rise in contentious actors [9] have both been observed, the platform continues to welcome back banned users.

### 1.1.2 Social Media Self-Moderation

A variety of people ranging from politicians to academics have faced negative repercussion for their online content, which can result in individuals being more self-aware and hesitant to share their beliefs and abstain from participating in online discourse [60]. This abstinence, often referred to as *self-moderation* or self-censorship [7], can be perceived as a form of repression that imposes challenges on the proper functioning of a democratic society. Indeed, self-censorship undermines freedom of speech, which is of paramount importance to the flow of information and fair democratic elections. For instance, Ong et al. [62] modelled how fear of state surveillance, harassment, and legal prosecution in Southeast Asian countries can reduce the expected utility of online expression, quieting dissenters and discouraging collective action. Likewise, following the 2016 Turkish coup attempt, Turkish citizens engaged in self-censorship, expressing less of their opinions on social media and voluntarily removing old posts, unfavourable to the government, due to fear of persecution [83]. Self-censorship is a major risk for authoritarianism and autocratization.

A major reason for engaging in self-moderation in social media is to avoid professional repercussions. Aktas et al. [2] describe Turkish academics' self-restraint in posting on social media. Larsen et al. [50] describe how journalists in Central American countries abstain from using social media to express their views due to job security concerns. Rudnik [71] shows that these same concerns lead Russian and Belarusian bloggers to self-censor.

Individuals also engage in self-censorship behaviour on social media due to safety and privacy concerns. In effect, there is a vast body of literature on the analysis of self-censorship in countries ruled by repressive regimes [50, 62, 15, 13, 25] where, for safety reasons, journalists avoid publishing or otherwise exchanging information about certain topics. In addition, Warner and Wang [88] revealed that the online self-censoring behaviour of individuals living in the United Kingdom has increased as new online surveillance methods were introduced by intelligence agencies. Privacy-conscious users would refrain from discussing topics suspected to be under active monitoring.

Social media self-moderation is also prominent in North America. Reddit users have been resorting to throwaway accounts when discussing divisive political events in the United States [59]. Powers et al. [67] examined American college students' view of social media discourse and showed that students preferred to discuss their political views offline, mostly due to a rather politically homogeneous nature of social networks and the desire to avoid frictions. This result comes as no surprise, given that a recent study by Gibson and Sutherland [33] revealed that 40% of Americans engage in self-censorship behaviour because they worry that expressing unpopular views will alienate people from their close circles. In [38], we explicitly examined North American motivations for self-moderating online.

### 1.1.3 Deleted Content on Social Media

It is challenging to keep track of retracted content on social media as such an effort involves analysing social media posts that either user or the platform tried to suppress. There are ethical concerns to collecting this type of content, and it is important to be abide by platform policies, and to protect users' personally identifiable information when possible. On the other hand, users are aware when they decide to post publicly on social media that their content has no reasonable expectation of privacy.

Studying deleted content is an effective way to scrutinize platform moderation policy and to understand self-moderation motivations. For example, Bhattacharya and Ganguly [11] characterized the Big-Five personality traits of Twitter users who deleted Tweets, as well as the vocabulary found in those Tweets. In a large-scale study, Almuhimedi et al. [4] explored the reasons behind Tweet deletion. While they found many Tweets were deleted for superficial reasons (e.g., spelling mistakes), they also found evidence suggesting user regret to be a likely factor. Building on that, Zhou et al. [94] designed a classifier for determining which Tweets were deleted as a result of regret, while Bagdouri and Oard [6] were able to predict which Tweets will be deleted in the future.

There are also ways to indirectly examine content that would be retracted. In one attempt, Sleeper et al. [78] conducted a user study on MTurk asking Twitter users about content they regretted posting. This analysis enabled an understanding of the context around the regret, but is susceptible to social desirability bias, where participants may not want to reveal their more shameful regrets. Sleeper et al. [77] also designed a study where users kept a log of statements they wanted to post but ultimately did not post. There is also Das and Kramer [20] who used Facebook internal data to capture content that users started writing but ultimately refrained from posting.

There are also a collection of volunteer services that keeps track of retracted content. Xia et al. [90] used `polititweet.org` – a service that tracks messages that were posted on Twitter (i.e., Tweets), but later deleted – to understand how deleted influencers' Tweets helped spread disinformation. A similar tool Reveddit [1] allows user to track content deleted

[1] https://www.reveddit.com/

on Reddit (which has users volunteering as moderators that can delete content). However, these services are often at the mercy of the platform's policies. Reveddit, along with many user-built applications which requiring streaming Reddit content, is built on top of the PushShift API [2]. Yet, Reddit has recently changed its data API policy, and as of May 2, 2023, PushShift no longer has access, and Reveddit is no longer available.

## 1.2 Proposed Research

The goal of the proposed research is to examine self-moderation and platform moderation behaviour by looking at the retraction of Retweets on popular Tweets. We noticed that popular Tweets generally fail to retain their Retweets, and that a decline in Retweet count is observed overtime for almost all Tweets. Here, we are looking at a decline in the absolute number of Retweets a post has and not the rate of gaining new Retweets. We consider a Tweet to be *popular* if, at any point in time, it has over 50 Retweets. This definition is motivated by the fact that fewer than 1% of Tweets receive more than 50 Retweets (most Tweets received no Retweets at all). If we were to choose a number higher than 50, we would have too few Tweets to work with. If we chose a value less than 50, there would be high variance when calculating Retweet loss.

**Popular Tweets:** Let $T$ be Tweet, and let $r(T,t)$ denote the number of Retweets $T$ has at timestamp $t$. $T$ is considered *popular* if:

$$\exists t,\ r(T,t) \geq 50$$

The life-cycle of popular Tweets falls into three stages. First, there is a period of rapid Retweet growth. Then, we see relative stability. Lastly, we have a period of gradual Retweet loss. The first two stages, which looks like $f(x) = log(x)$, are illustrated in Figure 2.12 in greater detail. This third stage where Tweets, particularly contentious ones, lose Retweets is our focus. As far as we know, there is no previous work studying this phenomenon.

To measure Retweet loss, we initially rely on finding datasets that capture a snapshot of Tweets some time in the past (at least a year). We then use the Twitter API to collect up-to-date information about the Tweet, and compare the current state to the historical state to measure Retweet loss. In datasets we will be collecting ourselves, we have the flexibility of more frequent snapshots and can thus define Retweet loss to be a ratio of the maximum Retweet count across all snapshots to the final Retweet count.

[2] https://github.com/pushshift/api

**Retweet Loss:** Let $t_*$ be the latest timestamp we collected information about $T$ in our study. Let $t_{max} = \arg\max_t r(T, t)$ We define Retweet loss ($rt_{loss}$) as follows[a]:

$$rt_{loss}(T) := \frac{r(T, t_{max}) - r(T, t_*)}{r(T, t_{max})}$$

[a] When dealing with historical datasets that which often consist of a single snapshot, we would not know the Retweet count at $t_{max}$. Instead, we set $t_{max}$ to be the earliest timestamp we have information collected for a Tweet $T$.

One would expect to see some Retweet loss because Retweeting users occasionally leave the platform and deactivate their accounts. However, we observed that Tweets of contentious content are more likely to lose a larger portion of Retweets, and since Retweeting is often seen as a form of endorsement [47], we hypothesize some element of self-moderation is at play. Similarly, platform moderation is also expected to be a factor.

### 1.2.1 Research Question 1

We being by re-emphasizing that the Retweet loss we are measuring here is not the rate at which a Tweet picks up Retweets, which is expected to decline as the Tweet becomes less relevant over time. Instead, the Retweet loss refers to a decline in a Tweet's cumulative Retweet count, i.e., the list of active users whose profiles show them Retweeting the Tweet.

While it is unclear why Retweets are lost, we suspect it is one of these three scenarios: (S1) the Retweeting user manually deleted his/her account; (S2) the Retweeting account was suspended by Twitter; (S3) the Retweeting user manually unretweeted. To be able to measure to proportion between these scenarios, it is necessary for us to collect our own dataset, because we could not find any existing datasets that tracked Retweeting users over time. In our data collection, described more in Chapter 2, we will create and track two datasets of Tweets, one of which is contentious while the other is uncontentious.

Using our custom datasets, we can track how many re-tweets were lost across various periods of times, by continually fetching the current status of Tweets using the Twitter API. We will also have snapshots of the list of Retweeting users, and would therefore know which users stopped Retweeting. We will be able to know definitively that unretweeting happened due to which of the three aforementioned scenarios. We will be able to track the rate of growth of these scenarios over time. The relative proportion of the scenarios and their rate of growth will yield many insights about moderation. For example, by tracking when account bans occur, we can get a sense of Twitter's platform moderation timeline.

Furthermore, by looking at the delay in manual unretweets, we can infer motivations and examine the nature of regret.

We previously hypothesized that there are three scenarios that contributes to Retweet loss: (S1) users deleting their accounts; (S2) Twitter suspending user accounts; (S3) users manually going on their profiles and undoing the Retweet. We want to investigate the proportion of each of those three scenarios.

**Research Question 1: What is the breakdown and timeline of various unretweeting scenarios?**

### 1.2.2 Research Question 2

We will also be designing features and trying to predict which Tweets are likely to lose a large portion of Retweets. These features based on Tweet content, author characteristics and Tweet metadata will give indications of what is often targeted by self-moderation and platform moderation. The goal is to find characteristics that are independent of specific Tweet topics and that generalize across various datasets of Tweets. To verify generalization, we will be training our classifer using existing historical datasets, and then validating them on the custom datasets we will be building and that the model has never seen.

We want to know the characteristics common to Tweets that face a lot of moderation. First, we will design and examine a set of features that are hypothesized to be predictive of Retweet loss. From there, the aim is to design a binary classifier trained to determine whether a Tweet will fall into the high-loss category.

This model could be a useful tool for authors wishing to know whether their posts will expect to face Retweet retractions. It would also allow us to examine what type of content leads to retractions, and how Retweet deletion behaviour varies by topic.

**Research Question 2: Which features are useful for identifying Tweets that will experience significant Retweet loss?**

### 1.2.3 Research Question 3

Lastly, we will be trying to model Retweet loss behaviour. We plan to approach this task from two levels: macro and micro. At the macro-level, we will try to construct the curve that represents the third stage of a Tweet's life cycle and illustrates how Retweet loss occurs over time. Meanwhile, at the micro-level, we aim to model individual unretweeting behaviour. To achieve this, we will analyze the network structure of users who have unretweeted and try to identify commonalities among them. We will consider several factors that may cluster unretweeting users, such as their proximity to the author, their connections to other Retweeters, and the number of authors they follow. We want to see if users who Retweet controversial content might be more likely to be suspended by Twitter. We also want to know if there is any difference in self-moderation and account self-deletion behaviour between users who are Retweeting an uncontentious post versus a contentious one.

At the macro-level, we want to model the timeline of Retweet-loss and to determine characteristics common to accounts that unretweeted. The goal is to validate the hypothesized three-staged Tweet life-cycle of Retweet growth, stability and then Retweet decline. We want to know at what point Tweets begin to lose Retweets and to measure the rate of Retweet loss over time. We will try to determine whether contentious Tweets have a different timeline compared to uncontentious ones. At the micro-level, we will examine characteristics of accounts that unretweeted, by analyzing the network structure, account details and behaviours of unretweeting users.

**Research Question 3: How can we model the timeline of Retweet loss as well as the properties of unretweeting users?**

# Chapter 2

# Data Collection and Exploratory Analysis

In this chapter, we go over the various data sources used in our analysis. We discuss their collection process, highlight their strengths and limitations, and examine potential ethical and privacy concerns regarding the data collection process. Lastly, we present some high-level exploratory analysis of the datasets to better motivate our research questions.

For our analysis, we required datasets of Tweets that capture a snapshot of Retweet counts at some timestamp in the past. Furthermore, they would ideally also have some Tweet metadata that can be made into useful prediction features, such as information about the author and the platform used to post the Tweet. However, it is rare to find such datasets. The papers mentioned in our related work do not publish full datasets, because Twitter has an updated policy that requires Tweet data to be shared only in aggregate, and it restricts sharing individual Tweet data beyond just the ID.

To comply with the new content redistribution policy, datasets that were previously available were retracted. For example, Lamsal [49] re-designed her Covid-19 dataset on March 20, 2020 to include nothing but the IDs. These datasets would need to be reconstructed by refetching the IDs using the Twitter API (known as Tweet *hydration*). However, the hydration process would not yield datasets that are useful for us, because the hydrated Tweets would only show the current Retweet count. It would therefore be impossible to know how many Retweets the Tweet had when it was posted, and we would not be able to calculate what portion of Retweets were lost.

While there are definitely legitimate ethical reasons for Twitter to be restrictive when it comes to data collected through its APIs, such as to comply with European GDPR [31]

(General Data Protection Regulation), there are also more pragmatic factors. Twitter has previously reduced third-party distribution of Tweet data due to monetization concerns [46]. Similar motivations are suspected to be behind other platforms, such as Reddit, reducing third-party redistribution [16].

In our study, we were open to responsibly exploring available online datasets that were made public by other users, even if those users might not be in harmony with Twitter's redistribution policy. However, we were cautious to ensure our data collection respects the privacy of users, and we only publish aggregated metrics. Furthermore, we also benefit from the Twitter academic APIs, and will ensure we comply by its policies.

On Kaggle[1], we found three suitable datasets that took snapshots of the Retweet count some time in the past. These datasets were hobbyist datasets curated by individual users. However these datasets do not track individual Retweeters, and thus, to study the properties of unretweeting users, we needed to create our own datasets. We went through two iterations collecting Tweets and tracking lists of Retweeters using the Twitter API. Unexpectedly, our data collection periods overlapped with the volatile period associated with Elon Musk's acquisition, and we observed interesting phenomena in our dataset that coincides with changes in moderation policy, which we will elaborate on in Chapter 4.

To be consistent with Twitter policy, we will not be making our datasets public. Meanwhile, although the Kaggle datasets themselves are a violation of Twitter policy, using them in our analysis and only posting the results in aggregate would not be. Meeks [56] discusses the ethical considerations of studying deleted content in light of Twitter's policy. In particular, the concerns focus on individuals being analyzed and put into the spotlight without their knowing consent. Due to the large scale nature of studies and the anonymity of online communications, this consent can be difficult to obtain. Fiesler and Proferes [28] notice that most users are unaware their Tweets might be used to conduct research and feel that they should be asked consent. However, attitudes towards this issue vary depending on the nature of the research. The main take-away is to ensure results are presented in aggregate and that individual users are not spotlighted.

Our study of retracted Retweets has some similarities to studying deleted Tweets. To ensure we are compliant, we will study retractions only in aggregated and not post identifying information of unretweeting users. Furthermore, our emphasis is not on the deleted content (the Tweet still exists), but rather on the unretweeting, which further reduces risk.

[1] https://www.kaggle.com/

## 2.1 Hobbyist Kaggle Datasets

We were looking for datasets that satisfies (1) contains the Retweet count; (2) has a variety of distinct authors; (3) the snapshot was taken close to when the Tweets were posted. These were rare, since, to begin with, few datasets tracked Retweet count. And of those, many either tracked a single celebrity, or took snapshots far after the Tweets were posted (it is mostly datasets built using the Twitter stream API that captured Tweets close to creation).

Two of the three datasets (*Election 2020*, *Covid-19*) satisfy all three conditions. Meanwhile, although there were fewer distinct authors (500 authors for 2417 Tweets) in the *NASDAQ* dataset, it is still large enough to capture a broad set of content. And although there is also a larger gap between Tweet creation and the snapshot taken, this difference can likely be compensated by the fact that the dataset had more time to lose Retweets (two years have gone by between the Kaggle snapshot and our snapshot).

### 2.1.1 Election 2020 Dataset

This election dataset[2] by Manch Hui was the starting point of our analysis. It consists of Tweets related to the 2020 USA presidential election by containing either the hashtag #trump or #biden from 2020-10-15 to 2020-11-08. We extracted a dataset consisting of the $\approx 0.5\%$ of Tweets that had 50 Retweets or more. The dataset was generally of high quality, with most Tweets being collected within three days of being posted (and all within seven days), as shown in Figure 2.1. We took our own snapshot of this dataset between 2022-09-17 to 2022-10-04. However, the Tweet IDs were stored as floats (with 53-bit precision), which was not sufficient to preserve the entire ID (which is 64 bits). As a result, it was not possible to hydrate some of the Tweets. Fortunately, the last 12-bits of the id are often unused, and by assuming they are all zeroes, we were able to hydrate around 30% of the Tweets. We confirmed that the creation time of the hydrated Tweet matched what was indicated in the Kaggle dataset and verified that the text of the hydrated Tweets mostly matched the original (though some Tweets have been since edited). As for the Tweets that could not be hydrated, we cannot tell whether they have been deleted or we do not have the full ID.

This Kaggle dataset also came with some useful metadata, such as the source platform, and some characteristic of the author (e.g., username, profile description). However, it

[2] https://www.kaggle.com/datasets/manchunhui/us-election-2020-tweets

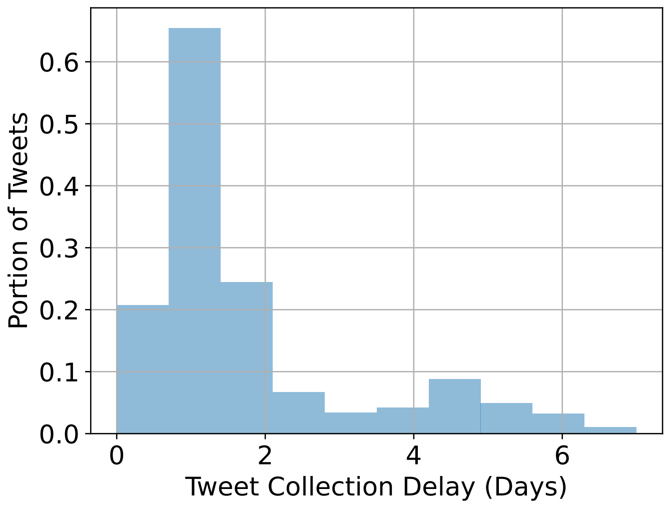


Figure 2.1: There was little delay in the data collection process, and we verified all Tweets in the *Election 2020* dataset were collected within 7-days.

lacks some other useful metadata, such as the author's follower count at the time of the Tweet, and whether the author was verified.

The posting of this dataset publically is in violation of Twitter API policies, since it shares detailed information about the Tweets. We will be using this data cautiously, and abiding by the ethical principles discussed earlier, such as only posting our analysis in aggregate, and not redistributing the dataset.

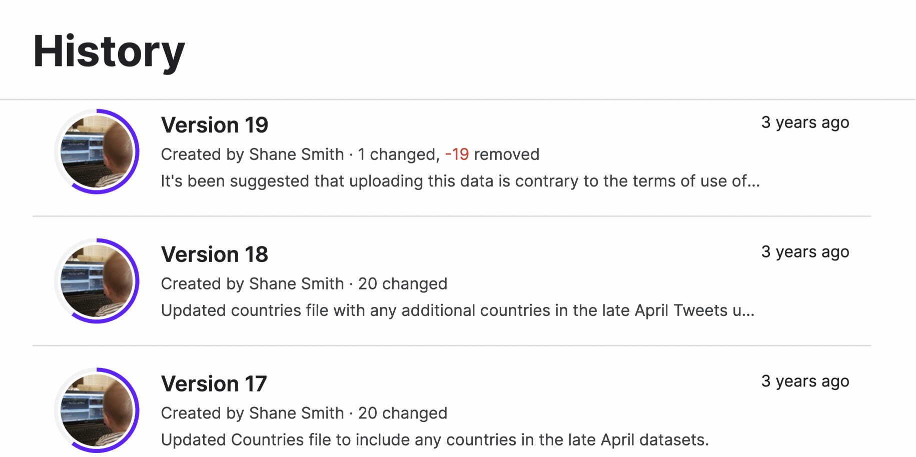
### 2.1.2 Covid-19 Dataset

This Covid-19 dataset[3] by Shane Smith was created by pulling Tweets that contain one of these hashtags: #coronavirus, #coronavirusoutbreak, #coronavirusPandemic, #covid19, #covid_19, #epitwitter, #ihavecorona, #StayHomeStaySafe, #TestTraceIsolate. The author indicated he was pulling new Tweets on a daily basis, and the last update time verifies that the snapshot was fully taken no later than 2020-04-03.

This dataset had the most comprehensive metadata, containing a snapshot of the author's profile along with the information about how the Tweet was posted (platform, language and location). The author went through 19 iterations when constructing the dataset,

[3] https://www.kaggle.com/datasets/smid80/coronavirus-covid19-tweets/versions/18

incrementally adding more Tweets. In the final (19th) version, the author appears to have self-censored the dataset; as shown in Figure 2.2, the 19 Tweet data files were removed, and only an unrelated file mapping country codes to country names were kept. The author attached the following commit message to the change "It's been suggested that uploading this data is contrary to the terms of use of the Twitter API. I'm removing the Tweets until either an exemption is granted or the terms change." However, it is possible to view previous commits, by specifying the version in the URL (the link in the footnote specifying the version still give access to the Tweets), and Google continues to index the previous versions.



History

Version 19
3 years ago
Created by Shane Smith · 1 changed, -19 removed
It's been suggested that uploading this data is contrary to the terms of use of...

Version 18
3 years ago
Created by Shane Smith · 20 changed
Updated countries file with any additional countries in the late April Tweets u...

Version 17
3 years ago
Created by Shane Smith · 20 changed
Updated Countries file to include any countries in the late April datasets.

Figure 2.2: In the final version of the *Covid-19* dataset, the .csv files containing Tweets were removed to comply with Twitter policy. However, we could still access previous versions.

Unlike the *Election 2020* dataset, the Tweet IDs were properly preserved. We could therefore know which Tweets were actually deleted and study properties of deleted Tweets. We narrowed the dataset down to the 1.4% of Tweets that had over 50 Retweets, and took our own snapshot between 2022-10-19 to 2022-10-20; we noticed that 10.2% of those popular Tweets were deleted. The breakdown of those deleted Tweets is: 5.1% author account suspended, 1.6% author account deactivated, 3.5% author account still exists, but the Tweet had been deleted. Since this dataset is our most comprehensive one, we will be training our models to predict Retweet loss based on this dataset in Section 3.2.

### 2.1.3 NASDAQ Dataset

This NASDAQ dataset[4] was created by Doğan et al. [22] and contains Tweets from 2015 to 2020, but the snapshot was taken in November, 2020. The goal of collecting this data was to attempt to predict stock prices using the public sentiment, as reflected in Tweets, about the respective companies.

The authors did not make use of the Twitter API; instead they built their own tooling based on the Selenium[5] web-scrapper. Their tool looked for Tweets that mentioned the NASDAQ ticker for Amazon, Apple, Google, Microsoft, and Tesla. Since they did not use Twitter's API, they are not bound by its policies and can freely share the scrapped Tweets. The ethical concerns are partly mitigated here, because without access to the Twitter APIs, the data gathered here did not include information collected about the author. Conversely, the drawback would be the lack of access to any of the Tweet metadata. The only information collected was: Tweet_id, author_id, post_date, Tweet_text, num_Retweets, num_comments, num_likes.

We only kept Tweets after 2018, to make the dataset more comparable with the other Kaggle datasets. Of those Tweets, 0.12% are popular (has 50 Retweets or more). While there are around 80000 authors across more than 1.5 million Tweets, once we have filtered down to only popular Tweets, we are left with 500 distinct authors. It would seem that the spread of financial information on Twitter is dominated by a few high impact authors. We took our own snapshot of the 2529 popular Tweets between 2022-10-25 and 2022-10-26, and noted that 4.42% of those Tweets have been deleted.

## 2.2 Exploratory Data Analysis

This section discusses some initial analysis we performed on the Kaggle datasets that motivated our research questions and shaped the design of our custom datasets. First, we present some overall characteristics relating to Retweets. We will look at the overall distribution of Retweet counts, and how that is related to characteristics about the author. From there, we will examine more in depth the characteristics of Retweet loss. We will compare the relative level of Retweet loss across the three datasets, and better understand the distribution of Retweet loss.

As shown in Table 2.1, the three Kaggle datasets have different columns available, and the sort of exploratory analysis that can be done on them thus varies. For example, having

[4] https://www.kaggle.com/datasets/omermetinn/tweets-about-the-top-companies-from-2015-to-2020

[5] https://www.selenium.dev/

| Dataset Name | Full Tweet ID | Tweet Metadata | Author Metadata |
|---|---|---|---|
| Election 2020 | ✗ | ✓ | ✗ |
| Covid-19 | ✓ | ✓ | ✓ |
| NASDAQ | ✓ | ✗ | ✗ |

Table 2.1: For each dataset, we show what types of data were available for our analysis.

the full Tweet ID is necessary to determine what portion of the dataset had been deleted, by checking to see if the ID still exists. The *Covid-19* dataset is the most complete of the three, and thus a lot of our analysis will focus on it.

### 2.2.1 Properties of Retweets

To start, we want to get a sense of the Retweeting behaviour, and how Retweet count is related to the size of the author's following, which is a strong measure of his/her influence [74]. The *Election-2020* dataset was the first one we explored, and some of the analysis on the properties of Retweets are shown for that dataset. However, we confirm that the observations are general and apply to our other datasets.

We show the raw distribution of the Retweet count in Figure 2.3, 2.4, and 2.5. The data is shown in $log_{10}$ scale, we see that few Tweets have ≥50 Retweets (≈ 1.7 on that scale).

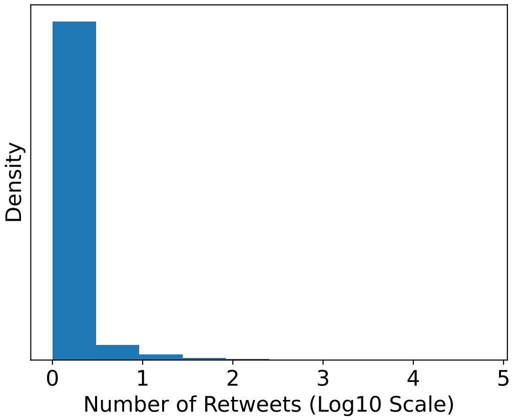


Figure 2.3: The distribution of Retweets in the Election Kaggle dataset, our *Election-2020* consists of just the Tweets with more than 50 Retweets.

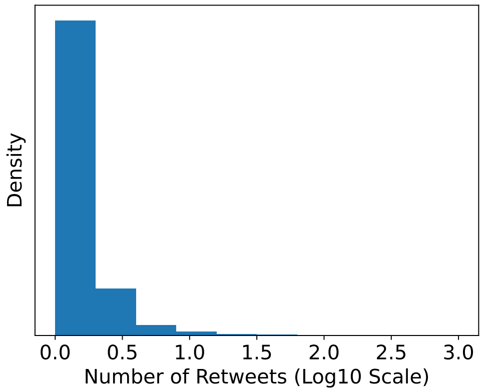


Figure 2.4: Dist. of Retweets in the Covid-19 Kaggle dataset.

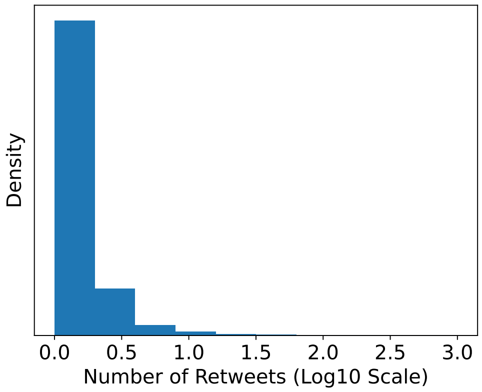


Figure 2.5: Dist. of Retweets in the NASDAQ Kaggle dataset.

From there, we look at the relationship between Retweet count and the size of the author's following. In Figure 2.6 and 2.7, we see that the authors who produced Tweets with more than 50 Retweets typically had more followers than the average Twitter user. The difference is statistically significant (t-statistic$=-201$, p-value$\approx 0$). As an observation, 99.9% of authors of popular Tweets had more than 50 followers.

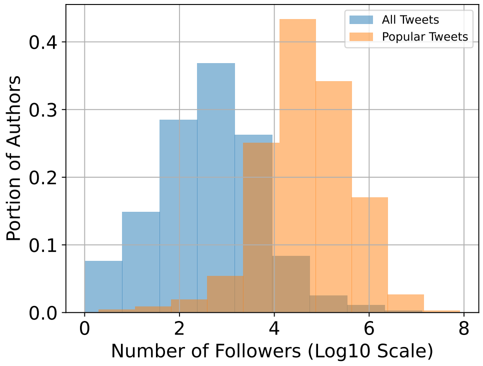


Figure 2.6: Authors of popular Tweets tend to have more followers than the average, as shown here for the *Election-2020* dataset.

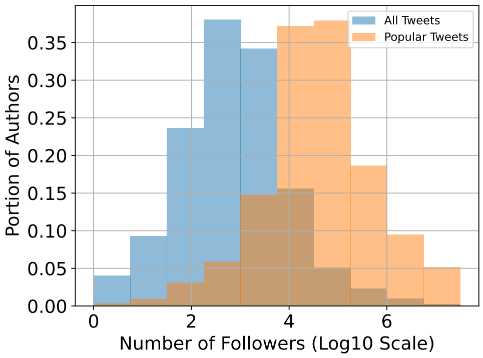


Figure 2.7: The same phenomenon for authors of popular Tweets is shown here for the *Covid-19* dataset.

Conversely, when we compared the Retweet count of post by popular authors with larger followings (≥1000 followers), we noticed their Retweets had higher than average Retweets. This effect is shown in Figure 2.8 and 2.9 for the *Election-2020* and *Covid-19* datasets respectively.

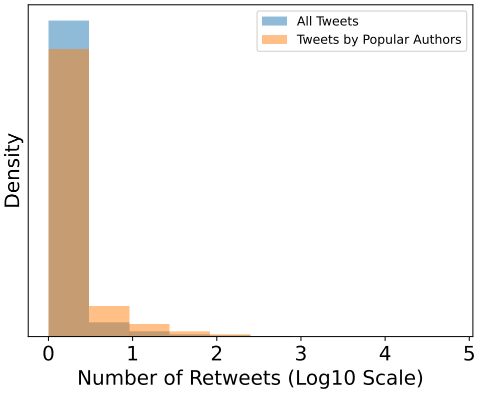


Figure 2.8: Authors with more than 1000 followers had more Retweets than average, as shown here for the *Election-2020* dataset.

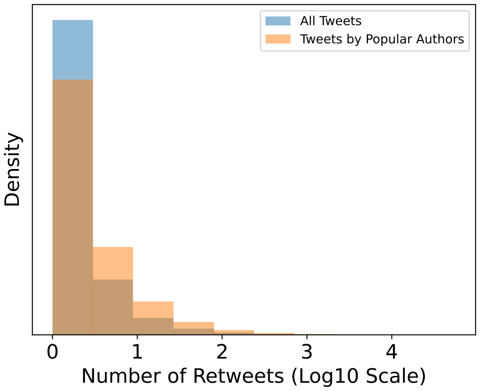


Figure 2.9: The same phenomenon for authors with large followings is shown here for the *Covid-19* dataset.

More generally, beyond our arbitrary cut-off of 1000 in the previous figures, there continues to be a positive correlation between the size of the author's following and and their post's Retweet count. We show this relationship for *Election-2020* dataset in Figure 2.10.

Meanwhile, if we were to condition on Tweets that have more than 50 Retweets, the impact of the author having more followers was small when it comes to gaining additional Retweets. We show this effect in Figure 2.11, where having more followers was insignificant, unless the author has more than one million followers.

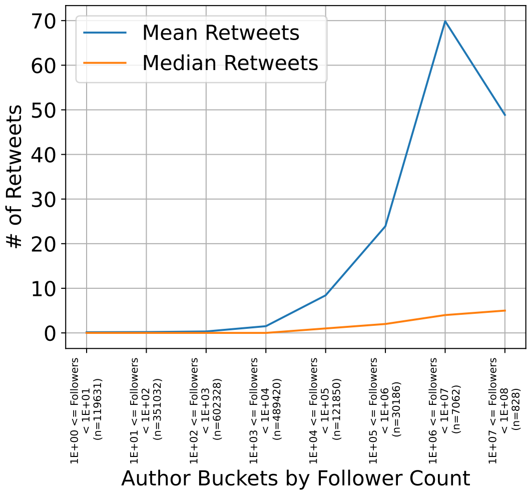


Figure 2.10: Rel. between having more followers and Retweet count for all Tweets.

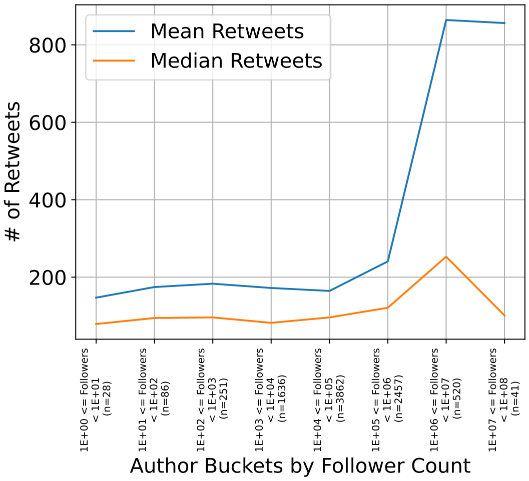


Figure 2.11: Rel. between having more followers and Retweet count for popular Tweets.

Lastly, we look at the timeline over which Retweets are accumulated. In Figure 2.12, we plot, for all 3 Kaggle datasets, the CDF of the portion of total Retweets accrued averaged across all the Tweets in the dataset.

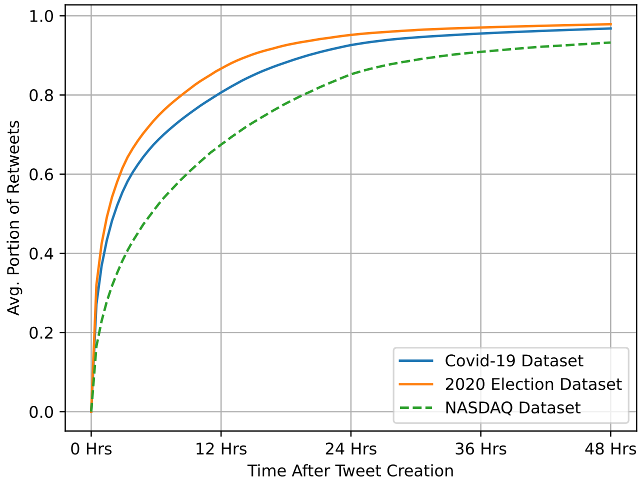


Figure 2.12: The CDF of the portion of total Retweets accrued averaged across the three Kaggle datasets.

We observe, as expected, that most Retweets are accrued in the early periods following when the Tweet was first posted. Note: only the Retweets that still existed at the time we took our snapshot are accounted for here. Thus, this timeline will look different than the timelines we are hoping to produce for RQ1 and RQ3, which include lost Retweets.

### 2.2.2 Properties of Retweets Loss

We now turn to our phenomemon of interest: Retweet loss. In Figure 2.13, we plot the CDF of $rt_{loss}$ for the three datasets. We notice a sharp rise centered around 20% for every dataset. However, the CDF of the Covid-19 and election datasets differ meaningfully from the NASDAQ dataset in that they are bimodal. They have a higher portion of Tweets with $rt_{loss} \geq 50\%$, and also have generally higher Retweet loss.

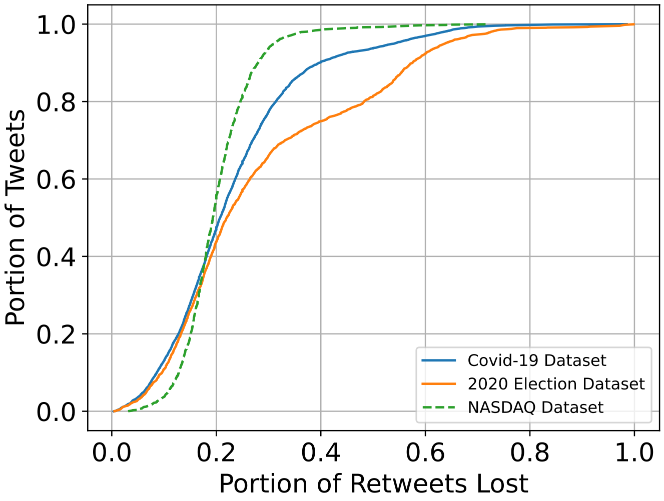


Figure 2.13: The CDF of the portion of Retweets lost ($rt_{loss}$) for the Election, Covid-19, and NASDAQ datasets.

The NASDAQ dataset consists of Tweets that mention large companies and has the least contentious content of the three. Its Tweets are also older than those in the other datasets. These two factors could explain why it has a less steep slope. In the NASDAQ dataset, we observed that 99.8% of Tweets have fewer Retweets now than they did at $t_0$. For the Covid-19 and the election datasets, it was 94.8% and 96.7% respectively. The higher percentage for NASDAQ is likely due to its Tweets being 1-2 years older than those in the other datasets. Our hypothesis is that as time goes by, old Tweets are unlikely to gain new Retweets; however, some Retweeting accounts are likely to be deactivated, leading to $rt_{loss} > 0$.

We hypothesize the peak centered around $rt_{loss} = 20\%$ to be some universal loss based mostly on users closing their accounts. If so, our focus will be on examining on Tweets with $rt_{loss} \geq 50\%$, which appear to be a property of contentious content. Of the Tweets that lost Retweets, the mean $rt_{loss}$ was 28.3%, 23.1%, and 20.1% for the Election, Covid-19, NASDAQ dataset respectively.

We now take a closer look at the *Covid-19* dataset, whose Retweet loss is sandwiched between the other two datasets. In Figure 2.14, we show a histogram of its Retweet loss. We note around 49.4% of Tweets lost more than 20% of their Retweets (the dashed line on the histogram).

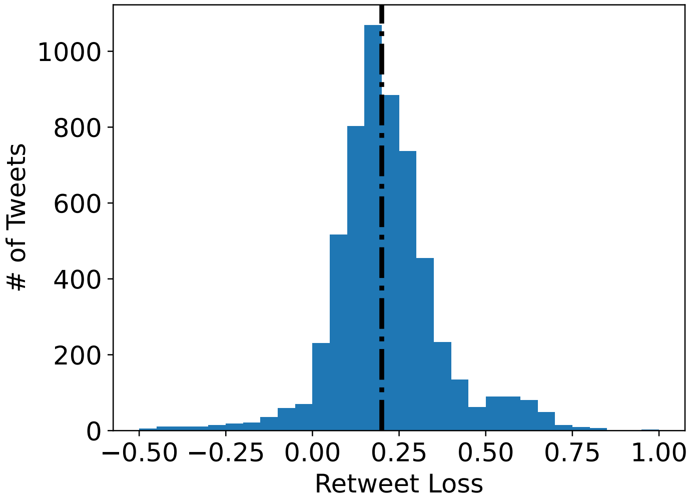


Figure 2.14: Histogram of the Retweet loss from the *Covid-19* dataset. The dashed line at 0.2 seperates Tweets into roughly equal halves.

The *Covid-19* dataset will be of special focus in our analysis, due to the completeness of its metadata, as shown in Table 2.1. We will be presenting a classifier that predicts which side of the dashed line in Figure 2.14 a Tweet will likely fall into.

## 2.3 Custom Datasets

One of the limitations of the existing Kaggle datasets is that they do not provide a list of users that Retweeted at $t_0$. Thus, while we are able to calculate $rt_{loss}$, we cannot understand the reasons behind the loss (RQ1). We also cannot examine common factors between unretweeting accounts (RQ3). Furthermore, of the three Kaggle datasets, only one of them had the necessary metadata needed for the training features we had in mind. Thus, it was necessary to build our own dataset.

In our custom datasets, we will be maintaining a list of Retweeting users over time, so we clearly see which users unretweeeted. We will also be collecting information about the followers of the Retweeters and examining their networks structure.

The availability of the Twitter Academic API[6] was what allowed us to search through large number of Tweets and keep only the popular ones (with over 50 Retweets), and to

[6] At the time of writing in June 2023, the Academic API has been deprecated and is no longer available.

continually take periodic snapshots. Without access to the Academic API, we would not have been able to build our own custom dataset.

### 2.3.1 The *Custom-Autumn* Dataset

To create the *Custom-Autumn* dataset, we gathered a corpus of Tweets posted between 2022-10-26 and 2022-11-03 and tracked them over 75 days. We created a list of search phrases we thought would likely lead to contentious Tweets and a list for yielding uncontentious ones. This list was inspired by investigating currently controversial political topics. Furthermore, we analyzed currently trending topics on Twitter and attempted to identify innocuous hashtags as well as topics that would likely lead to contentious discussions. We created dataset $D_u$ using 1000 uncontentious Tweets, and dataset $D_c$ using 1000 contentious Tweets. We capped the number of Tweets contributed by each search query at 50.

In Table 2.2, the search queries used and the number of Tweets each query contributed to the dataset are shown. We acknowledge that the types of content deemed contentious varies meaningfully by demographics; as an example, "lest we forget" could be contentious to an audience who have live through the Vietnam war period. Nonetheless, for a quick sanity check, we selected 100 Tweets, with 50 from each of $D_u$ and $D_c$, and had them labelled manually by three peers as being either contentious or uncontentious. The labelling process involved active discussion, and ultimately it was possible to achieve consensus on all Tweets. We called the combined dataset, consisting of $D_u$ and $D_c$, the *Custom-Autumn* dataset.

Inspired by Jimenez-Sotomayor et al. [44], we used the labelled results to calculate 95% confidence intervals for each of set of queries. Assuming normal approximation of random error, we calculate our interval using:

$$\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Here, $\hat{p}$ is our sample proportion, and the multiplier $z^*$ is set to 1.96 for a 95% level of confidence. With that, the labelled data suggests 20% ± 11% of the $D_u$ Tweets are expected to be contentious , while it is 66% ± 13% for those in $D_c$.

Throughout the 75-day observation period, we took ten snapshots tracking all the users currently Retweeting each Tweet. Furthermore, for the 50 first Retweeters on each of the Tweets, we collected the list of all of their followings, capped at 75000 followers. The reason we impose this cap is due to Twitter API rate limits.

While the limits on using the API to search for Tweets is quite generous for the academic endpoints (10 million Tweets per month), the constraints on fetching user follower information are quite restrictive. Only 15 requests can be made in any 15 minute period, and each request returns at most 5000 user IDs. In practice, this translate to 15-minutes of wait time between every 15 requests. The limit of 75000 was set to avoid spending more than 15 minutes on any one user (as some users have more than 1 million followers).

This information would allow us to understand the network structure of the users Retweeting a given Tweet, which would be useful in modelling unretweeters for RQ3.

| Search Query | # of Tweets | Search Query | # of Tweets |
|---|---|---|---|
| abortion | 50 | baking | 50 |
| affirmative action | 50 | graduation | 50 |
| transgender | 50 | wedding | 50 |
| queer | 50 | trick-or-treat | 50 |
| putin | 50 | snowman | 50 |
| protest | 50 | reunion | 50 |
| obama | 50 | promotion | 50 |
| inflation | 50 | harvest | 50 |
| vaccine | 50 | donation | 50 |
| diversity | 50 | dogs | 50 |
| covid19 | 50 | celebration | 50 |
| communism | 50 | cats | 50 |
| censor | 50 | wholesome | 50 |
| dr. oz | 48 | hiking | 47 |
| fracking | 47 | new home | 43 |
| feminism | 37 | relaxation | 43 |
| elizabeth II | 37 | food | 39 |
| midterm election | 35 | bread | 35 |
| prostitution | 32 | lest we forget | 33 |
| critical race theory | 32 | pumpkin | 31 |
| student debt | 30 | birthday | 27 |
| homelessness | 26 | gardening | 26 |
| immigrant | 26 | fitness | 26 |

Table 2.2: Contentious (Left) and uncontentious (Right) search queries and the corresponding number of Tweets contributed to *Custom-Autumn* Dataset.

### 2.3.2 The *Custom-Winter* Dataset

After we finished collecting data for the *Custom-Autumn*, we scaled up and collected a larger-scaled dataset: the *Custom-Winter* dataset. We made several improvements when gathering data in this dataset. First to be more representative of Twitter, we note some search queries should be weighted more heavily than others (i.e., the topic is more popular). We thus no longer adjust the ratios by capping the number Tweets contributed by a search query to 50. We reduced the number of search queries, so it would be easier to vet them. For example, "cats" which was used in *Custom-Autumn* turns out to be different than expected, as they turn out to often be advertising about cats available for adoption. Second, we are more careful with the quality of the data by ensuring duplicate Tweets are avoided and that no single author is over-represented. Lastly, we automated large portions of the data collection process which allows us to take much more frequent snapshots. Our dataset *Custom-Winter* is a combination of 2000 uncontentious Tweets (which we will call $D_u$), and 2000 contentious Tweets (called $D_c$). Table 2.3 shows the search queries used and the number of Tweets each query contributed to the dataset. We have higher confidence that content found through the queries actually match what is orignally expected. The ratio also reflects the topics' relative popularity on Twitter.

In Appendix A, we discuss how we conducted a user study to confirm that $D_c$ does indeed contain more controversial Tweets than $D_u$.

| Search Query | # of Tweets |
|---|---|
| BlackHistoryMonth | 449 |
| putin | 442 |
| abortion | 258 |
| diversity | 248 |
| feminism | 55 |
| homelessness | 48 |

| Search Query | # of Tweets |
|---|---|
| food | 959 |
| celebration | 315 |
| wedding | 155 |
| wholesome | 50 |
| baking | 21 |

Table 2.3: Contentious (Left) and uncontentious (Right) search queries and the corresponding number of Tweets contributed to *Custom-Winter* Dataset.

Due to the heavy limitation Twitter set on its APIs for collecting user network structure data, we did not attempt to do the collection for the *Custom-Winter*. Our user network observations would all come from the *Custom-Autumn* dataset.

## 2.4 Dataset Overview

In Table 2.4 and 2.5, we summarize the high-level features of the Kaggle datasets and the custom datasets that we will be using to answer the research questions discussed in Chapter 1. We show information about the dates when data collection took place, as well as some high-level summary metrics. For the Kaggle datasets, we complemented the existing data by collecting our own snapshot, i.e., we fetched the current status of the Tweets using the Tweet IDs from the datasets. These snapshots enable us to get up-to-date information about those Tweets and to calculate the Retweet loss. If the Tweet IDs are complete (which is the case for *Covid-19* and *NASDAQ*, but not for *Election-2020*) we can also determine the number of Tweets that were deleted. The dates when our own snapshots are taken are also shown in the tables.

Overall, the data we were able to gather for our analysis was of high-quality. However, we were not able to gather network structure data for the *Custom-Winter* dataset due to highly constraining nature of the Twitter API rate limits for gathering user follower data. The Twitter Academic API endpoints were crucial to our research and allowed us to take continual snapshots of the Tweets we were tracking. However, on April 27, 2023, we received an email from Twitter indicating that our application has been "suspended from accessing the Twitter API" (see Appendix B). This suspension is a result from Twitter's February announcement [8] that it intends to eliminate free-tier access. At this point, we have already collected most of the data needed, and the impact of the suspension is therefore quite limited. The only data we had failed to collect was the final breakdown of the unretweeting users for the *Custom-Winter* dataset. However, we do have the full breakdown data for the *Custom-Autumn* dataset and the preliminary breakdown (one third of the way into the data collection period) for the *Custom-Winter* dataset.

Lastly, we benefitted greatly from access the Twitter Academic APIs. This acess binds us to its terms of service. We are very cautious in ensuring that the APIs were used appropriately and that we have not shared any information in this thesis that would be contradictory to the terms agreed upon.

| Dataset Name | Source | Tweet Posted | Collection Dates |
|---|---|---|---|
| Covid-19 | Kaggle | 2020-03-09 to 2020-03-11 | The Kaggle dataset consists of a snapshot taken some time before 2020-04-03. Additionally, we took our own snapshot between 2022-10-19 to 2022-10-20. |
| Election 2020 | Kaggle | 2020-10-15 to 2020-11-08 | The Kaggle dataset was collected between 2020-10-21 and 2020-11-09. Tweets were always captured within seven days of being posted. Additionally, we took our own snapshot between 2022-09-17 to 2022-10-04. |
| NASDAQ | Kaggle | 2018-01-01 to 2020-12-31 | The Kaggle dataset consists of a snapshot taken some time around 2020-11-26. Additionally, we took our own snapshot between 2022-10-25 to 2022-10-26. |
| Custom-Autumn | Twitter API | 2022-10-26 to 2022-11-03 | We collected ten snapshots between 2022-11-04, to 2023-01-18 (75 days). |
| Custom-Winter | Twitter API | 2023-01-30 to 2023-02-06 | We collected 850 snapshots between 2022-02-06 and 2023-04-22 (75 days). |

Table 2.4: For each dataset, we show the data source, the dates Tweets were posted, and the dates the snapshots were taken.

| Dataset Name | Tweet Count | Mean Retweet Count | Median Retweet Count | Number of Distinct Authors |
|---|---|---|---|---|
| Covid-19 | 6488 | 410.3 | 108 | 3057 |
| Election 2020 | 7654 | 237.4 | 101 | 2813 |
| NASDAQ | 2417 | 95.3 | 71 | 500 |
| Custom-Autumn | 2000 | 154.4 | 71 | 1721 |
| Custom-Winter | 3000 | 257.4 | 97 | 2933 |

Table 2.5: For each dataset, we show some high level metrics about Retweets and the number of distinct authors.

# Chapter 3

# Data Analysis

In this chapter, we set out to answer our research questions. Most of our results are contained in this chapter, with each subsection corresponding to a research question. At a high level, we will examine the relative proportion of the different unretweeting scenarios for RQ1 using our two custom datasets: *Custom-Autumn* and *Custom-Winter*. From there, we will address RQ2 by training a model using the *Covid-19* dataset to predict which Tweets will lose a large portion of their Retweets. The trained model will be applied to our custom datasets, which it would never have before seen. Lastly, for RQ3, we will present a curve-fitting algorithm to model the timeline of Retweet loss. We will also model individual unretweeting behaviour using network structure data from our *Custom-Autumn* dataset.

## 3.1 Breakdown of Unretweeters

This section addresses research question 1, and analyzes the relative proportion of unretweeting scenarios across the two datasets we built: *Custom-Autumn* and *Custom-Winter*. As a reminder, we were hoping to investigate the following.

**Research Question 1: What is the breakdown and timeline of various unretweeting scenarios?**

As part of this analysis, we discovered a new unretweeting scenario that was not part of the original hypothesis. This scenario consists of users changing the status of their accounts to *protected*, making their profile only visible to an approved list of followers. As a result, their past Retweets no longer count in the Retweet total. With that, unretweeters fell into one of four categories. The first category (S1) consists of users that have *deactivated* their accounts; we also get an error from the API in that case, but there is no message indicating they have been suspended. The second category (S2) consists of *suspended* users. When looking up these user by their IDs with the Twitter API, we get an error saying they have been suspended. In the third category (S3), the Retweeter is still active on the platform but have manually unretweeted. Lastly, we have (S4) the aforementioned case where the user protected his or her account.

The analysis process consists of building a list of users that have at some point unretweeted, and then using the *Users* API endpoint `https://api.twitter.com/2/users` to gather information from their profile and infer their current status. There are however steep rate limits constraining us to 300 requests per 15-minute window. Our intention was to collect this user information at two different timestamps for each dataset: once at the very end of the data collection period, and once a third of the way into the process (75 days and 25 days respectively). This timeline is better visualized in Figure 3.1. In Chapter 4, we will be putting some major events following the acquisition into the context of this timeline (see Figure 4.1).

The goal of taking the two snapshots is to better understand the delay in platform moderation, particularly in light of the volatility following Musk's acquisition. Our academic access to the Twitter API was revoked before getting the final snapshot for the *Custom-Winter* dataset; however, we were able to get the preliminary snapshot. Ideally, we would have liked to have taken much more frequent snapshots.

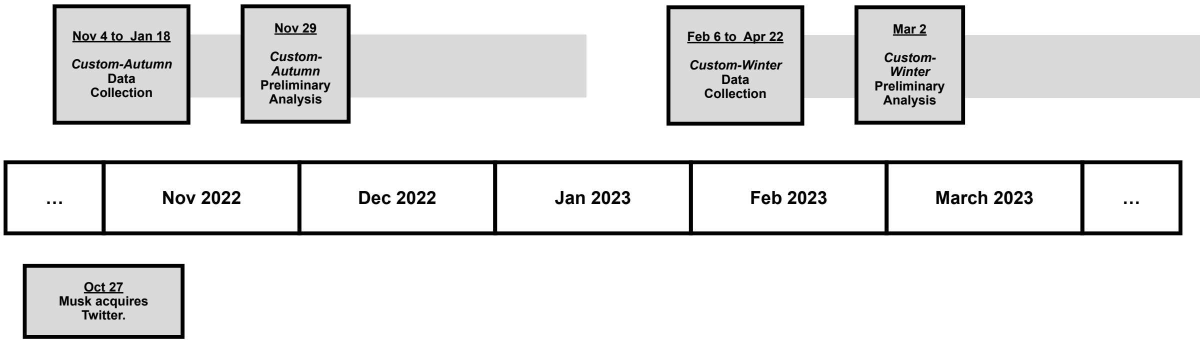

Nov 4 to Jan 18 Custom-Autumn Data Collection
Nov 29 Custom-Autumn Preliminary Analysis
Feb 6 to Apr 22 Custom-Winter Data Collection
Mar 2 Custom-Winter Preliminary Analysis
...
Nov 2022
Dec 2022
Jan 2023
Feb 2023
March 2023
...
Oct 27 Musk acquires Twitter.

Figure 3.1: A timeline showing the preliminary as well as the full data collection periods for our *Custom-Autumn* and *Custom-Winter* datasets.

### 3.1.1 Users Per Unretweeing Category

We now examine the breakdown of unretweeters across S1, S2, S3, S4. For our *Custom-Autumn* datasets, we have both the preliminary results and the final results, referring to snapshots taken in the timeline of Figure 3.1. These results are shown in Figure 3.2. For the *Custom-Winter* dataset, we only have the preliminary results and it is shown in Figure 3.3.

First, we note the Retweet loss was much lower in our custom datatsets than in the Kaggle datasets. It was on average 4.2% in the *Custom-Autumn* dataset and 1.3% in the *Custom-Winter* dataset. While the higher Retweet loss percentage in the Kaggle datasets are certainly partly due to the Tweets being older, we believe it is also due to platform moderation differences following Musk's acqusition (e.g., reinstating suspended accounts). We discuss this difference more in Section 3.3. For example, most of the Retweet loss from suspended accounts came after more than 25 days. This lack in account suspension in the early periods overlaps with the period following the acquisition where moderation staff were laid off [72], and there is a rise in contentious content on the platform [10, 9]. The portion of suspensions rose significantly in the final breakdown. Furthermore, this lack in suspension is no longer observed in *Custom-Winter*, and the breakdown of the preliminary results is quite close to the final breakdown for *Custom-Autumn*.

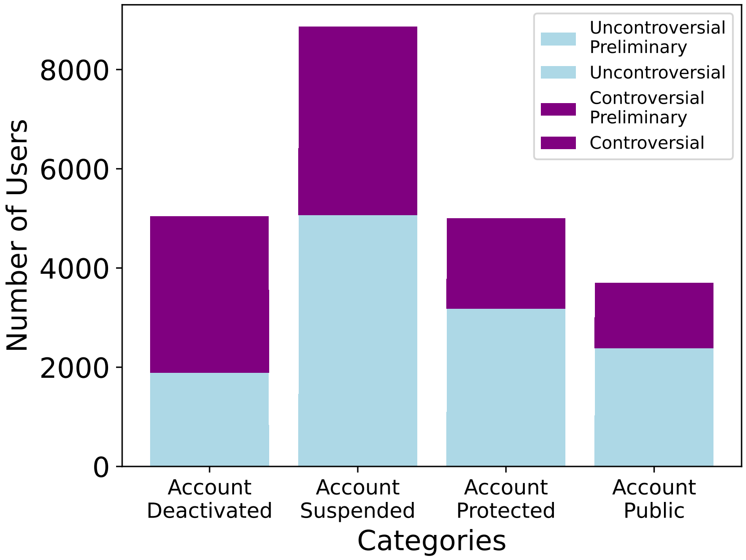


Figure 3.2: The preliminary and final breakdown of unretweeters across the 4 categories for the *Custom-Autumn* dataset.

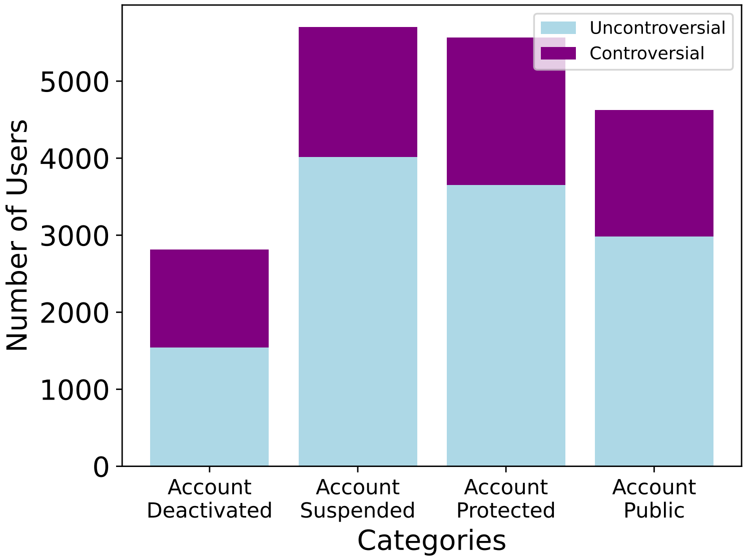


Figure 3.3: A preliminary snapshot breakdown of unretweeters across the 4 categories for the *Custom-Winter* dataset.

### 3.1.2 Contentious Versus Uncontentious Tweets

To better compare the uncontentious Tweets to the contentious ones, we show their percentages of each unretweeting scenarios on bar graphs found in Figure 3.4 and Figure 3.5. The first figure is for the *Custom-Autumn* dataset, and consists of the breakdown at the 75-day mark. The second figure is for the *Custom-Winter* dataset, at the 25-day mark. We note the relative percentages are quite similar between the two figures.

Somewhat surprisingly, account suspensions by the platform were more prevalent for Retweeters of uncontentious content. We have two hypotheses for this observation. First, self-moderation could be at play, since contentious unretweeters were more often voluntarily deactivating their accounts. Second, a larger portion of the uncontentious content is of an advertising or self-promotion nature. There might be more incentives to use bots or fake accounts to boosts Retweets on those posts, and thus these Retweeting accounts are more likely to be suspended. Comparing Figure 3.4 and Figure 3.5, we noticed the account deactivation rate to be lower. It is possible this might be related to the Musk acquisitions. Immediately following the acquisition, many users expressed discontent and discussed leaving the platform. The relative proportion of the four scenarios stayed mostly similar for contentious Tweets across the two datasets. Meanwhile, for uncontentious Tweets, we see a rise in self-moderation, but a decline in platform moderation when contrasting Figure 3.4 and Figure 3.5. This phenomenon might be due to policy changes on the platform.

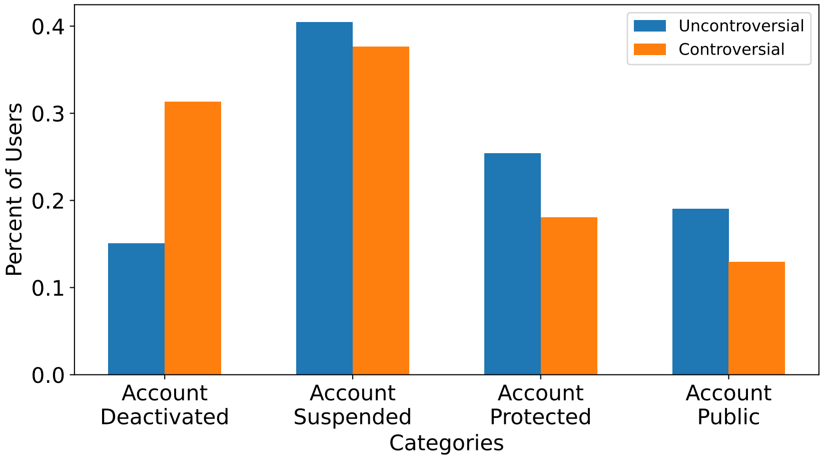


Figure 3.4: The relative percentage of Retweet loss contributed by each of the 4 breakdown categories for the *Custom-Autumn* dataset.

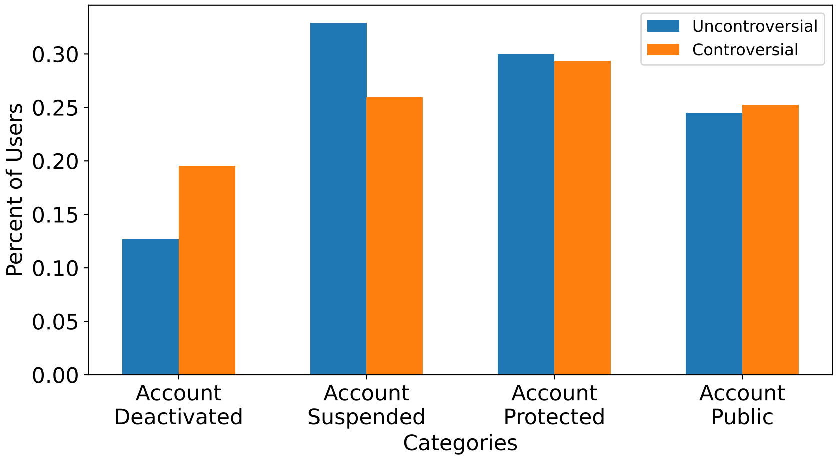


Figure 3.5: The relative percentage of Retweet loss contributed by each of the 4 breakdown categories for the *Custom-Winter* dataset. This breakdown is from the preliminary results.

### 3.1.3 Timeline of unretweeting Cases

The overall Retweet loss often does not tell the full story, since it is the sum of new Retweeters and disappearing unretweeters. Thus, the Retweet loss percentage values aforementioned are actually underestimates. Section 3.3 will discuss this idea in more detail. For now, we will look at the absolute number of new Retweets gained and lost over time. In Figure 3.6, we show the number of new Retweeters and new unretweeters that were accumulated between snapshots for the *Custom-Autumn* dataset. The rate of losing Retweeters appears somewhat constant over time; however, it is initially masked by the inflow of new Retweeters. As the rate of new Retweeters drops off over time, the overall Retweet count begins to diminish.

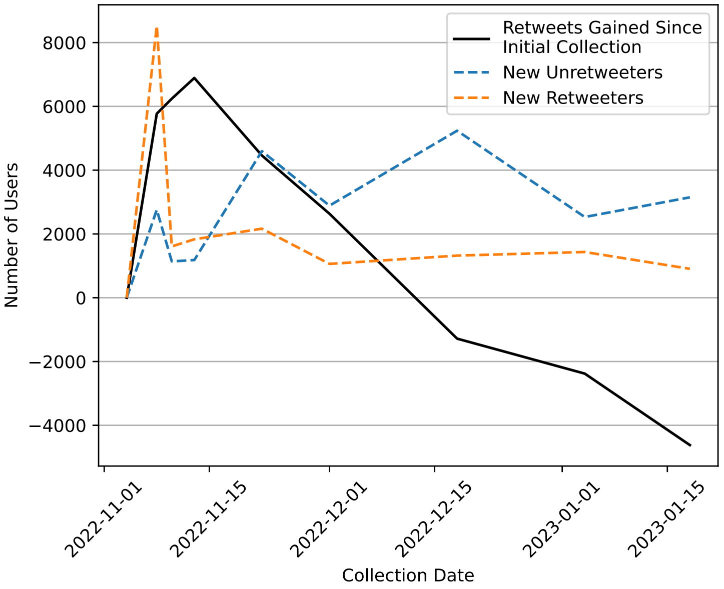


Figure 3.6: Tracking the flow of new Retweeters and unretweeters over time for the *Custom-Autumn* dataset.

In Figure 3.7, we show the same breakdown for the *Custom-Winter* dataset. We were able to capture more frequent snapshots in the winter data collection process; as a result, the timeline is much more granular. Two characteristics of the *Custom-Winter* dataset comes across in Figure 3.7. First, the overall rate of Retweet loss is lower compared to *Custom-Autumn*. Second, there was shorter delay between when Tweets were posted and when our snapshotting began for the *Custom-Winter* dataset. As a result, Tweets were still in the process of rapidly gaining Retweets. Thus, even after an extended period of Retweet loss, the Retweet count is still higher than the initial Retweet count at the time of collection, when the count was still far from the peak. Furthermore, the data collection for *Custom-Winter* overlaps with a period of time when Twitter was reinstating suspended accounts (discussed in Appendix C). This period is shaded in Figure 3.7; we see sudden jumps in Retweet count, even though the Tweets were already supposed to be in the Retweet loss phase of their lifecycle.

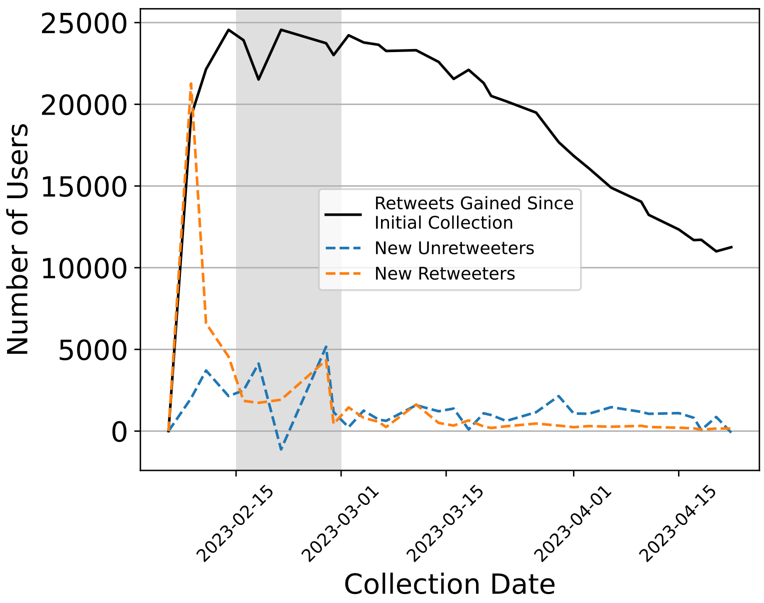


Figure 3.7: Tracking the flow of new Retweeters and unretweeters over time for the *Custom-Winter* dataset. The shaded region reflects a period of time when Twitter was reinstating suspended accounts.

**Answer to RQ1:** We discovered a new unretweeting reason through the process - users changing their account status to protected. Overall, (S2) Twitter suspending accounts is the greatest cause of Retweet loss. Deactivated (S1) accounts and protected accounts (S4) make up a similar portion of Retweets lost. The higher Retweet loss on contentious content is due to a higher likelihood of account deactivation. As time passes, new unretweeting cases accumulate at roughly the same rate, while the Tweet stops gaining Retweets, leading to overall Retweet loss.

## 3.2 Predicting Retweet Loss

This section addresses research question 2 by looking at how predictive features can be determined to predict which Tweets will lose a significant portion of their Retweets.

**Research Question 2: Which features are useful for identifying Tweets that will experience significant Retweet loss?**

Using the *Covid-19* dataset, we designed and trained a classifier to predict which Tweets would have high Retweet loss. We did not consider the *NASDAQ* nor the *Election 2020* dataset when designing this model, as they did not contain all of the metadata needed for many of our features (e.g., account age, author follower count). The *Covid-19* dataset contains 6488 Tweets. After removing the Tweets that no longer existed (e.g., deleted or made private), we are left with 5764 Tweets. From there, we trained models with the usual 80-20 train-test split, leaving 20% for validation. This ratio is used in many other Twitter classification tasks [69, 40, 76].

The intent of our model is to separate Tweets into two groups based on their Retweet loss. As shown in Figure 2.14, around half of the Tweets in the dataset lost more than 20% of their Retweets. Our model will take in information about a Tweet and try to classify it to the correct side of the dashed line in the figure. In other words, it will assign a $label = 1$ to Tweets it believes have lost fewer than 20% of their Retweets, and a $label = 0$ to those believed to have lost more.

By categorizing instead of predicting an explicit Retweet loss percentage, we make it easier to validate the model in other datasets. Different datasets have varying base Retweet-loss rates (as shown in Figure 2.13). Furthermore, the age of the dataset also determines the raw Retweet-loss percentage (since Tweets lose Retweets over time). If we assume the relative ranking of Retweet loss percentage between Tweets of the same dataset is somewhat stable, we can validate our classification even on a relatively-new, unseen dataset.

### 3.2.1 Feature Design and Justification

In this section, we discuss how we designed and settled on our prediction features. We present some analysis of the features individually and propose explanations as to why these features might be related to Retweet loss. The full list of features we settled on are presented in Table 3.2.

**Feature Group 1 - This first group of features are designed based on the actual textual content of the Tweet.**

**Sentiment Score and Sentiment Magnitude:** We had the sentiments of the Tweets labelled with Google Cloud API[1]. The API is able to parse text in multiple languages, and it returns a discrete score ranged between -1 and 1, at 0.1 intervals. There were a small number of Tweets that could not be parsed, and those were discarded in our analysis. The score is an overall evaluation of the emotional content in the Tweet. For example, a score of -0.2 corresponds to a text with slightly negative emotion; meanwhile, a score of 0.8 suggests the text expresses very strong positive emotions. We suspect sentiment score is related to Retweet loss, as Bhattacharya and Ganguly [11] observed that negative sentiment is more likely to lead to Tweet deletion.

To get a sense of the sentiment score distribution, Figure 3.8 shows the histogram of the sentiment scores for both the contentious and uncontentious Tweets from the *Custom-Autumn* dataset. Unsurprisingly, contentious Tweets had more negative sentiments, which has a median of -0.1 compared to a median of 0.2 for uncontentious Tweets. This median score was 0.0 for all three Kaggle datasets. From there, we compared the sentiment of Tweets with high $rt_{loss}$ to those with low $rt_{loss}$. We chose two thresholds of 0.2 and 0.5 based on critical points in the Figure 2.13 CDFs. For each threshold $t$, we form a control group (CG) consisting of Tweets whose $rt_{loss} < t$ and a target group (TG) whose $rt_{loss} \geq t$. (We did not include the NASDAQ dataset, because there were few Tweets with $rt_{loss} \geq 0.5$.)

To compare the sentiment score distribution between TG and CG, we use the Mann-Whitney U (MWU) test. The test assigns a probability that the median sentiment score of TG and CG are the same. The p-value of the test states how likely we are to see the observed difference between the groups or an even greater difference, if the groups indeed came from the same distribution. A smaller p-value means it is less likely the groups came from the same distribution.

As shown in Table 3.1, it is highly probable that the sentiment scores of high $rt_{loss}$ Tweets fall into a different distribution than those with low $rt_{loss}$. We see the portion of positive Tweets (sentiment score > 0) is higher for the CG with lower $rt_{loss}$, and that the mean sentiment score of those Tweets is more positive. Both means are close to 0, because most Tweets have a neutral sentiment (as seen in Figure 3.8).

**Flesch Reading Score:** This readability metric proposed by Flesch [29] measures how the effort required to read a body a text with the formula:

$$206.835 - (1.015 * ASL) - (84.6 * ASW)$$

where ASL measures the Average Sentence Length and ASW measures the average word length. Flesch claims a text with a score of 0 is "practically unreadable," while score

[1] https://cloud.google.com/natural-language

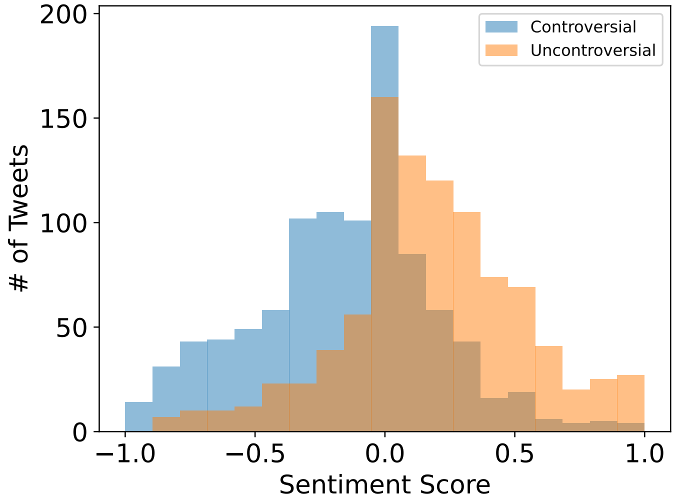


Figure 3.8: The contentious dataset has a more negative sentiment score distribution than the uncontentious one.

of 100 means it is easy to read by any literate person. Leonhardt and Makienko [51] showed that the Flesch readability metric is related to Tweet engagement and more readable posts have significantly more Retweets.

**Number of Characters in Tweet:** In terms of Tweet length, we observed the same phenomenon as Gligorićet al. [34], where authors try to go up to the character limit. This behaviour is illustrated in Figure 3.9. Currently the character limit is 280 characters (the dashed line). However, when there is an attachment associated with the Tweet, the API attaches a URL of the form `https://t.co/xxxxxxxxxx` to the end of the Tweet text. This link is not shown when viewing the Tweet on Twitter, and it therefore does not count towards the character limit. This 23 character link is what leads to a second peak around 303 characters (the solid line). There are also other invisible characters that show up when fetching from the API (such as new line characters). These characters explain why there are Tweets with more than 303 characters when fetched with the API.

In short, it is clear that authors typically try to go up to the character limit. We suspect there is a fundamental difference between authors that are interested in going up to the limit and those who do not. There is probably also a difference between content that needs to use up all the characters in order to be properly expressed. For example, Ghenai [32] found character count to be a significant predictor of whether a Tweet is health misinformation, and Jenders et al. [41] found it predictive of Tweet virality. The character

|  | **Covid-19** |  | **Election** |  |
|---|---|---|---|---|
|  | $t = 0.2$ | $t = 0.5$ | $t = 0.2$ | $t = 0.5$ |
| CG Positive Tweets | 47.08% | 44.78% | 39.96% | 37.89% |
| CG Avg. Sentiment | 0.053 | 0.043 | 0.0070 | 4.6e-4 |
| TG Positive Tweets | 40.55% | 28.36% | 32.10% | 25.68% |
| TG Avg. Sentiment | 0.029 | 0.010 | -0.025 | -0.059 |
| MWU p-value | 1.6e-5 | 0.0060 | 0.0012 | 1.5e-7 |

Table 3.1: Comparing the proportion of positive Tweets and the average sentiment score of a control group (CG) with $rt_{loss} < t$ to a target group (TG) with $rt_{loss} \geq t$.

count of posts are likely related to the post quality, which likely affect the Retweet loss.

**Feature Group 2 - This second group of features are considered Tweet metadata.**

**Tweet Source:** Groshek and Cutino [36] showed that Tweets and Retweets that originate from a mobile platform were more uncivil and impolite. This observation is consistent with Suler's model [80] on online disinhibition, which proposed that less delay in responses between users reduces disinhibition, which is generally the case for mobile communication. Furthermore, we studied properties of deleted Tweets using the *Covid-19* dataset, and noticed that Tweets posted from a mobile platform were more likely to be deleted (though we are unsure if it is by the platform or by the user). Of the Tweets that were from deleted accounts in the dataset, 76.3% were posted from one of these mobile platforms: "Twitter for iPhone", "Twitter for Android", "Twitter for iPad"; meanwhile, of the Tweets that still exist, only 46.3% were posted from a mobile platform. Examining those 434 deletions, we note 330 (76%) of them were due to account suspension.

As observed in 3.1, a meaningful portion of the Retweet loss is due to account suspension. Although here we are looking at author account deletion rather than the deletion of Retweeting accounts, it is likely the two are related. We suspect a relationship between the quality of a post and the medium used to post the Tweets. Tweets posted from a desktop platform could be more deliberate and thought out.

**Retweet Momentum:** This custom feature of ours is a measurement of how quickly a Tweet gained Retweets. It does so by measuring what portion of the total Retweets were gained within time period $\delta$ of the Tweet being posted. It can be defined as follows:

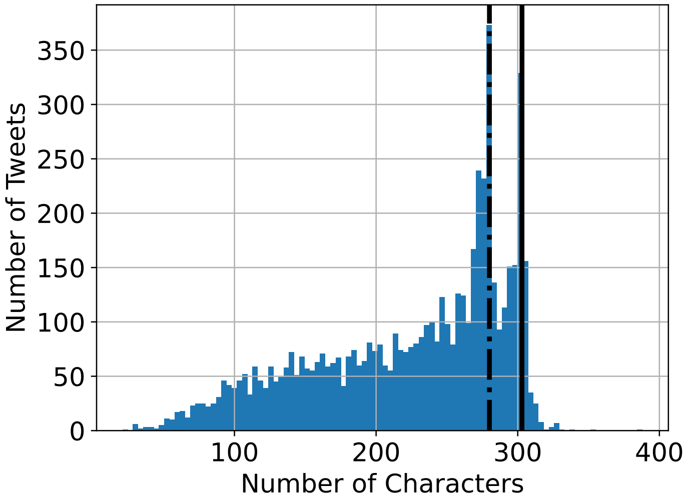


Figure 3.9: Tweets have a 280 character limit. There is peak around 280 (dashed line), which suggests authors try to go up to the character limit. A second peak at 303 characters (solid line) is due to Tweets with attachments (the API returns a 23-character URL to attachment). The Tweets that are even longer contain invisible characters.

$$rt_{momentum}(T, \delta) := \frac{r(T, t_{creation} + \delta)}{r(T, t_0)}$$

Here, $t_{creation}$ is the timestamp when the Tweet was posted. It is impossible to determine $rt_{momentum}(\delta)$ precisely in the Covid-19 dataset, even though we have the timestamp of the Retweets, due to deleted Tweets. With our custom datasets, we get a closer approximation, as we are have continual snapshots. In any case, the approximations should be close. The $\delta$ in our analysis was 2 hours.

This Retweet momentum is likely to not only reflect the content of the Tweet, but also the make-up of its Retweeters, which would impact Retweet loss. Fan et al. [26] confirms the speed of gaining Retweets is affected by the composition of users that are Retweeting the post. The presence of Retweeting hub users (users with large followings) is related to the momentum with which Tweets gain Retweets. It is possible that the types of users Retweeting is related to Retweet loss, as they might exhibit different levels of self-moderation.

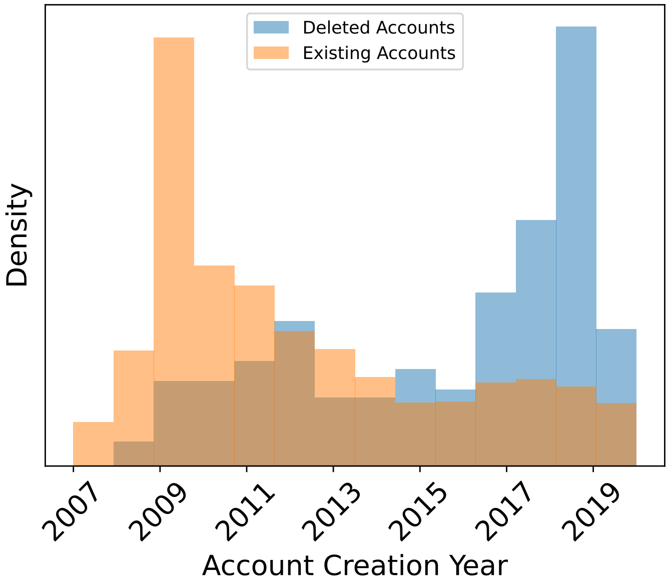


Figure 3.10: Deleted accounts (containing both suspended and deactivated accounts) in the Covid-19 dataset were much younger than those still existing.

**Feature Group 3 - This last group of features are all characteristics of the posting account.**

**Age of Account:** The age of the author's account also differed meaningfully between Tweets that were deleted and Tweets that still exist. We show this difference in Figure 3.10, where the the distribution of the account creation year for both set of accounts are plotted. Based on this difference, and the aforementioned motivations, we propose a feature based on the age of the author's account. Suh et al. [79] showed age of account to be a predictor of a Tweet's Retweet count.

**Account Verification:** In Covid-19 dataset, around 49% of the Tweets were from verified accounts. This value drops to 25% in our dataset (28% in the prototype dataset). Prior to Musk's acquisition and the launch of the Twitter Blue subscription program, verified accounts were vetted by Twitter based on principles surrounding "Active, Notable, and Authentic." It is reasonable to expect that these accounts would, in general, post higher quality content. Paul et al. [65] showed verified users have meaningfully larger influence and reach than non-verified users; Ghenai [32] found verified accounts are a lot less likely to post health misinformation.

**Follower Ratios:** Lastly, we have two features based on the author's follower count: Retweet-to-Follower Ratio, Following-to-Follower Ratio. First, we have the number of Retweets garnered by the Tweet divided by the author's follower count. Second, we look at the number of account followed by the author divided by the author's follower count. This ratio is commonly optimized by influencers, which Agam [1] discusses in the context of Instagram. In both cases, we add 1 to the follower count to avoid division by zero.

| Feature | Description |
|---|---|
| Sentiment Score | A value that reflects the overall sentiment direction of the Tweet as determined by Google Cloud NLP. |
| Sentiment Magnitude | A value that reflects the total sentiment content in the Tweet, also determined by Google Cloud NLP. This metric is only weakly correlated with the sentiment value (because positive and negative content in the same Tweet can cancel each other out). |
| Retweet-to-Follower Ratio | The number of Retweets received on the Tweet divided by the author's follower count. |
| Age of Account | The age of the author's account (in days). |
| Following-to-Follower Ratio | The number of accounts the author follows (which Twitter refers to as friends) divided by the number of followers the author has. |
| IsVerified | Is posted by an verified author (i.e., has a blue checkmark). |
| IsMobile | Is posted from a mobile platform. |
| Characters in Tweet | The number of characters in the Tweet. |
| Flesch Reading Score | A score given by $206.835 - (1.015 * ASL) - (84.6 * ASW)$ where ASL = average sentence length and ASW = average word length in syllables. A higher score represents a text that is easier to read. |
| Retweet Momentum | The percentage of Retweets gained within the two hours of being posted. It measures whether the Tweet was quick to gain traction. |

Table 3.2: The initial set of features used to train our classifier. First, an OLS regressor is trained; the p-values of those coefficients are used to perform feature selection, before training a final logistic classifier.

### 3.2.2 Feature Selection & Model Training

Using these features, we initially applied an Ordinary Least Squares (OLS) regression to predict on the raw Retweet-loss rate. From there, to do feature selection, we dropped all features which had p-value higher than 0.1. The features that were dropped are: Sentiment Magnitude Score, IsVerified, IsMobile. The p-value of the various coefficients are shown in Table 3.3.

| Feature | OLS P-Values |
|---|---|
| Sentiment Score | $< 0.001$ |
| Sentiment Magnitude | 0.371 |
| Retweet-to-Follower Ratio | 0.934 |
| Age of Account | $< 0.001$ |
| Following-to-Follower Ratio | 0.053 |
| Is Verified | 0.214 |
| Is Mobile | 0.158 |
| Characters in Tweet | $< 0.001$ |
| Flesch Reading Score | $< 0.001$ |
| Retweet Momentum | $< 0.001$ |

Table 3.3: The p-values of the features listed in the Table 3.2. Features with p-value $> 0.1$ are dropped from the final model.

From there we applied a logistic regression model to classify Tweets into two categories. Tweets that lost fewer than 20% of their Retweets (in the top half when it comes to preserving Retweets) are intended to receive a positive label.

### 3.2.3 Model Validation

Taking inspiration from Ghenai [32], we present our model features in Table 3.4. All coefficients had p-value $p < 0.05$. Validating the model on Covid-19 test set, we achieved a F1-Score of 0.65 on the test set. The TP and TN rate were both roughly the same, at 0.652 and 0.649 respectively. In Figure 3.11, we visualize the real Retweet-loss percentage of the two categories seperated by our classifier. Those with label=0 are the high Retweet-loss group. If our classifier was perfect, all of the label=0 would be on the right side of the dashed line, and all of the label=1 would be on the left. The average Retweet loss of the

| Variables | Coefficients | Std. Errors | P-Values |
|---|---|---|---|
| (Intercept) | 0.7129 | 0.1558 | *** |
| Sentiment Score | 0.5624 | 0.1193 | *** |
| Retweet-to-Follower Ratio | 0.0345 | 0.0148 | |
| Age of Account | 0.0002 | 0.0000 | *** |
| Following-to-Follower Ratio | -0.1792 | 0.0490 | ** |
| Characters in Tweet | 0.0023 | 0.0004 | *** |
| Flesch Reading Score | -0.0095 | 0.0006 | *** |
| Retweet Momentum | -1.8249 | 0.1377 | *** |

Table 3.4: Logistic regression predicting whether a post is in the top half when it comes to preserving Retweets. For each feature, we show the coefficient (unstandardized), standard error, and p-value. Significance levels: $p < 0.0001$ ***, $p < 0.001$ **, $p < 0.01$ *, $p < 0.05$ .

high loss group was 25% compared to 15% for the low group. From there, we applied the trained model to our two custom datasets. Even though Tweets those datasets were from a different time period and environment (post-acquisitions) and have not had as much time to accumulate Retweet loss, the classifier was nonetheless somewhat effective. In the *Custom-Winter* dataset, the high loss group had a rate of loss of 1.53% compared to 1.26% for the low group. In the *Custom-Autumn* dataset, it was 4.80% compared to 4.47%. Due to the lower Retweet loss rate, the mean is much more susceptible to anomalies. In any case, the *Covid-19* dataset was different enough from the two custom datasets that it was not obvious learning could be transferred over. Yet, it appears some learning has transferred over, as shown in the histograms in Figure 3.12 and 3.13, where we see Tweets classified with label=1 indeed have lower Retweet loss.

**Answer to RQ2:** There are trained features which generalize across unseen datasets that can predict which Tweets will lose a meaningful portion of their Retweets. These features fall into three broad categories: Tweet metadata, Tweet content, and characteristics of the author.

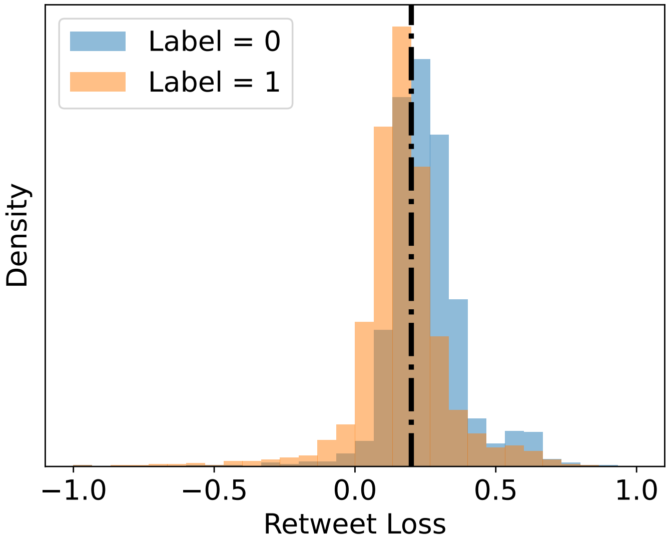
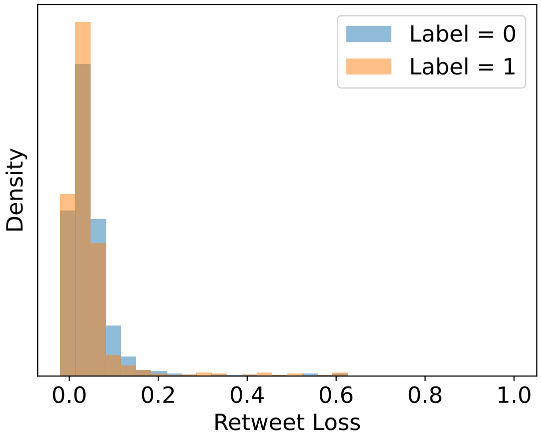


Figure 3.11: Histograms of the true Retweet loss for the *Covid-19* test set. If the classifier was perfect, the dashed line would perfectly separate the two histograms.



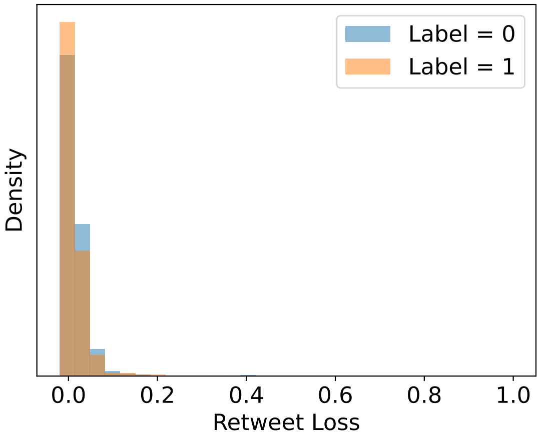


Figure 3.12: The results from applying the classifier the whole *Custom-Autumn* dataset.

Figure 3.13: The results from applying the classifier on the whole *Custom-Winter* dataset.

## 3.3 Modelling Retweet Loss

In this section, we first lay out the timeline of Retweet loss. From there, we will examine how we can model the types of users that are most likely going to unretweet.

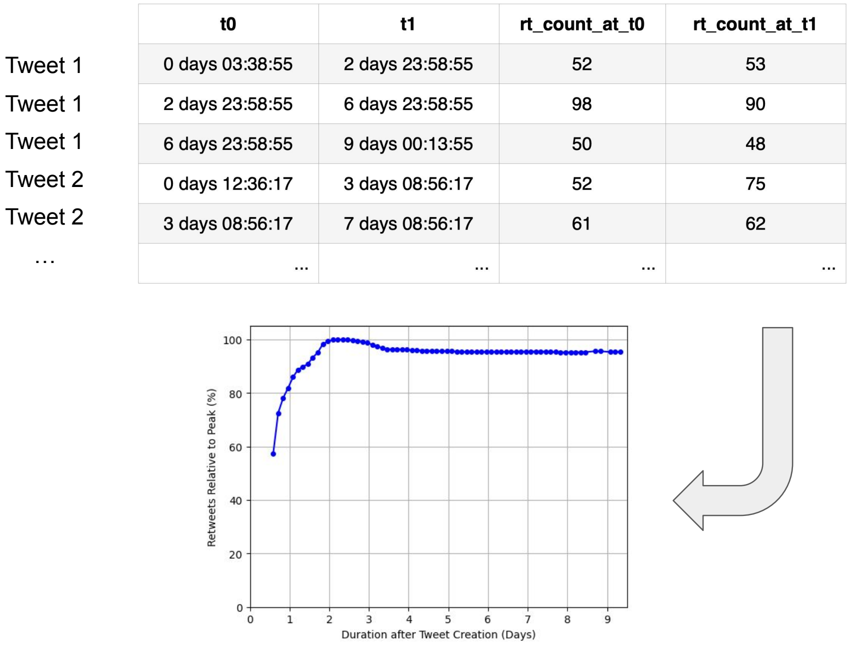
While in Section 3.1, we looked at the relative proportion of the unretweeting scenarios across a collection of Tweets, this section will shift perspective and model the expected life-cycle of an individual Tweet. In particular, we want to know at what point of the life-cycle do we expect cumulative Retweets to begin declining, and at what rate does the decline occur, as more time passes. Having an expectation of this timeline not only allows the unretweeting behaviour to be modelled better, it also allows authors to monitor their Tweets' relative to that of the average Tweet. Furthermore, we will also examine how unretweeters can be modelled by exploring their relationship to the author and properties of the Retweeters' network structure.

**Research Question 3: How can we model the timeline of Retweet loss as well as the properties of unretweeting users?**

### 3.3.1 A Tweet's Expected Life-cycle

When building a dataset, all of the Tweets collected are at somewhat different stages of their life-cycle, due to the fact that they were posted at different times. This difference in life cycle persists whenever we continue to take snapshots over time. As a result, the data we collected are in the format of the table shown in Figure 3.14. The goal, as the figure illustrates, is to take data in snapshot format and convert it to a curve which represents the trajectory of the average Tweet. Note each Tweet ID appears in the table multiple times, and for each Tweet, each collection timestamp appears once under $t_1$ and once again under $t_0$. The *rt_count* columns refers to the Retweet count of the Tweet at the two different timestamps. The challenge in this curve reconstruction is that the age of various Tweets are quite different at the times snapshots are taken. We need to make certain assumptions about the life-cycle and develop a technique to align those timestamps. Furthermore, the Retweet count of various Tweets are often different by orders of magnitude.

To address these problems, we present the following novel algorithm for the reconstruction of the curve.



| | t0 | t1 | rt_count_at_t0 | rt_count_at_t1 |
|---|---|---|---|---|
| Tweet 1 | 0 days 03:38:55 | 2 days 23:58:55 | 52 | 53 |
| Tweet 1 | 2 days 23:58:55 | 6 days 23:58:55 | 98 | 90 |
| Tweet 1 | 6 days 23:58:55 | 9 days 00:13:55 | 50 | 48 |
| Tweet 2 | 0 days 12:36:17 | 3 days 08:56:17 | 52 | 75 |
| Tweet 2 | 3 days 08:56:17 | 7 days 08:56:17 | 61 | 62 |
| … | … | … | … | … |

Figure 3.14: Fabricated data intended to demonstrate the purpose of the curve reconstruction algorithm. We want to combine the data for all the snapshots of the dataset into a single curve, which represents the average life-cycle.

### The Curve-Reconstruction Algorithm

This curve reconstruction problem is equivalent to trying to recover a continuous function f: $\mathbb{R}^+ \to [0,1]$ whose shape we do not know with a collection of *scaled* samples as follows:

$$\{s_i,\ t_i,\ p_i,\ q_i\}_n$$

with $s_i, t_i \in D(f)$ and $t_i > s_i$ and where $p_i =\ k_i f(s_i) + \epsilon$, $q_i =\ k_i f(t_i) + \epsilon$.

Here, $s_i$ and $t_i$ represent the timestamps, and $p_i$, $q_i$ represent the Retweet count at those respective timestamps.

Furthermore, let the value of $k_i \in \mathbb{R}^+$ vary for each sample, and assume $\epsilon \sim \mathcal{N}(0, \sigma^2)$ as in classical regression. We now set out to find $\hat{f}$, a discrete approximation of $f$. We begin by proposing a transformation of the original samples as follows:

$$\{x_i, y_i\}_n = \left\{ \frac{s_i + t_i}{2}, \frac{\frac{q_i - p_i}{p_i}}{t_i - s_i} \right\}_n$$

From there, cut the range $[min(\{x_i\}_n),\ max(\{x_i\}_n)]$ into $n_b$ equal sized ranges. For each range, bucket all the $y_i$'s whose corresponding $x_i$ fall into that range. Then, map each range to the mean of the $y_i$'s in its bucket. We now have:

$$\{[a_i,\ b_i],\ m_i\}_{n_b}$$

where $[a_i,\ b_i]$ are ranges such that $\bigcup\limits_{i=1}^{n_b} [a_i,\ b_i] = [min(\{x_i\}_n),\ max(\{x_i\}_n)]$ and that for every $1 \le j \le n_b$, we have:

$$\frac{\sum_{s \in S_j} s}{|S_j|} = m_j, \text{ when } S_j =\ \{y_i \mid x_i \in [a_j,\ b_j]\}$$

From there, let $g$ be a discrete function with domain $D = \bigcup\limits_{i=1}^{n_b} \{a_i\}$. We set $g(a_0) = 1$, and for define the remainder of the values with recurrence:

$$g(a_{i+1}) = g(a_i)[1 + m_i(a_{i+1} - a_i)]$$

A discrete approximation of $f$ can be found by normalizing $g$ with the max value of its image, i.e., $\hat{f} = \frac{1}{max(I(g))} * g$. Note: in practice, some buckets have too few elements. We can skip over those and set $g(a_{i+k}) = g(a_i)[1 + m_i(a_{i+k} - a_i)]$.

### Curves for the Custom Datasets

The hyper-parameters for the algorithm were determined based on a process we called the *snapshot approach*. Unlike our longitudinal approach where we take many snapshots overtime, this approach takes only two snapshots, a couple hours apart. However, data is collected for a much larger collection of Tweets spanning across a wide range of the life-cycle. If there were sufficiently many Tweets, we should be able reproduce the curve with high-fidelity. However, in reality, the number of Tweets for most topics are insufficient for the reconstruction. A few topics that worked well were the weekly recurring hashtags[2] and politically charged keywords (e.g., affirmative action, abortion, critical race theory). From there, we chose the best fitting hyper-parameters for the number of bins, and the minimum elements a bin had to contain for it not to be skipped over. Admittedly, these parameters were chosen partly based on priors we have for the shape of the curve: we expect the curve to be fairly smooth, and for the decline in Retweets to be continual. These priors are not that unreasonable and based on fairly sensible motivations to, on one end, not be too sensitive to anomalies, and on the other end, be sufficiently fine-grained to capture the true nature of the curve.

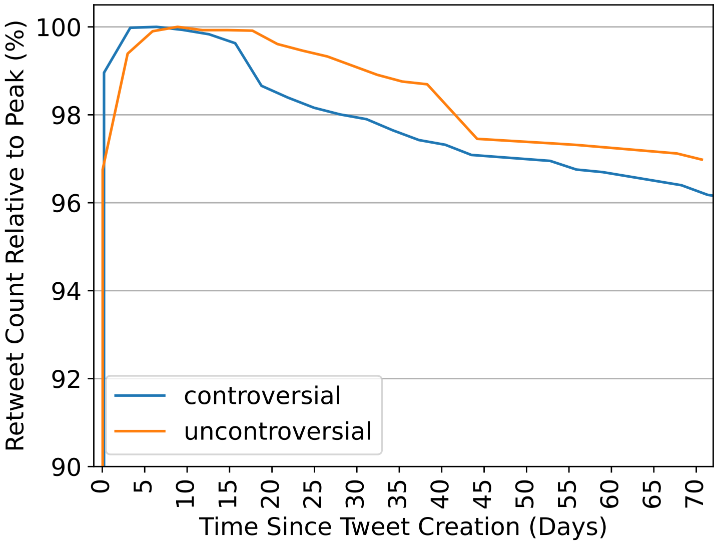


Figure 3.15: The constructed curve of the average Retweet loss over time for the *Custom-Autumn* dataset, separated by controversy.

[2] https://business.twitter.com/en/blog/recurring-twitter-hashtags-for-every-day-of-the-week.html

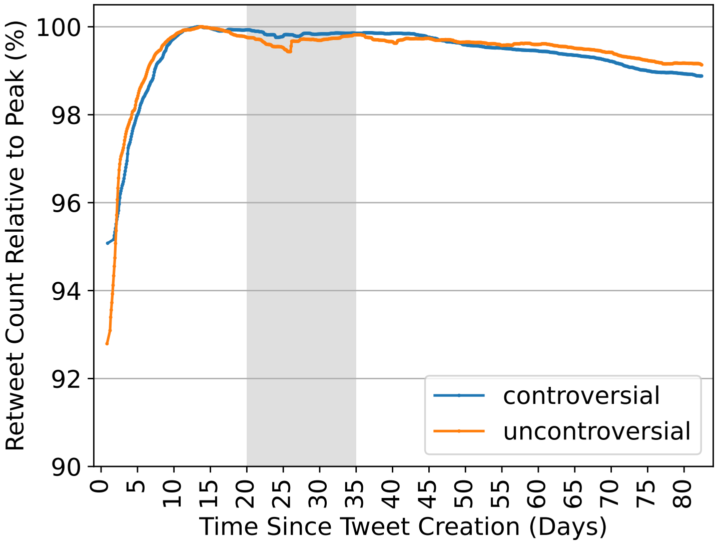


Figure 3.16: The constructed curve of the average Retweet loss over time for the *Custom-Winter* dataset, separated by controversy. The shaded region reflects a volatile period where Twitter was reinstating suspended accounts.

We settled on requiring there to be 150 data points for a bin to be considered. From there, the number of bins to use was chosen so that over 90% of the bins had sufficient data points. This translates to using using 25 bins for the *Custom-Autumn* dataset and 1000 bins for the *Custom-Winter* dataset.

The reason that we are able to use many more bins for the *Custom-Winter* datasets is improvements in our data collection process. Unlike the *Custom-Autumn* dataset, we were able to take snapshots of the Retweet count multiple times per day, because this process was modified to count Retweets without tracking the list user_ids that were currently Retweeting. In Figure 3.15, we show the plot generated from the *Custom-Autumn* dataset. Similarly, in Figure 3.16, we show the plot generated from the *Custom-Winter* dataset. The shaded area in Figure 3.16 corresponds to the volatile period where Twitter was reinstating suspended accounts (discussed in Appendix C). We see an increase in Retweet count, in particular for uncontentious Tweets. This observation is consistent with what we saw in Figure 3.4 and 3.5, where uncontentious Retweeters are more often suspended.

We note in *Custom-Winter*, the Retweet loss is less significant. This difference could be due to decreased moderation on the platform, and also possibly be related to the reinstatement of previously suspended accounts. In both custom datasets, we noticed that uncontentious Tweets had less Retweet loss, which is consistent with our original hypothesis and our observations in the Kaggle datasets.

### 3.3.2 Characteristics of Unretweeters

To begin, we examine the network structure of Retweeters, which refers to the follower-following relations of all the users that Retweeted a given Tweet. For the *Custom-Autumn* dataset, we were able to pull the follower list of 58437 Retweeters. These Retweeters are all one of the first 50 users to Retweet some Tweet in the dataset. Using this network structure data, we will present observations about how a Retweeter's connection to other Retweeters are linked to her propensity to unretweet.

In Figure 3.17, we compare the degree of connectivity of the unretweeting Retweeters to all Retweeters. Here, connectivity refers to the number of users each Retweeters is following. The maximum connectivity value is 50, because we are only considering the first 50 Retweeters. We see a meaningful difference between Retweeters that unretweeted and the Retweeters as a whole. Across all Retweeters, 56.6% of users were connected to at least another user, whereas this value drops to 26.3% if we look at only unretweeting users. With a Mann–Whitney U statistic of ≈2e8, there is a p-value of ≈0 of the two groups of users being from the same distribution. Furthermore, disappearing Retweeters are much less likely to be following the author. Across all Retweeters, 40% of users followed the author. This value drops to just 17% if we consider only unretweeters. Conversely, the author is also much less likely to be following unretweeters. Of all the Retweeting users, the author followed 16%. This is 4.8% for unretweeting users. From here, we shift our focus to the relationship between Retweeting early and unretweeting. We define the rank of a user to refer to the order in which the user Retweeted the post. The author has a rank of 0, and the first Retweeter has a rank of 1. In Figure 3.18, we show the relationship between this rank and probability of unretweeting. We notice, with p-value=0.013 that there is a negative relationship between rank and unretweeters, which is to say early Retweeters are more likely to unretweet.

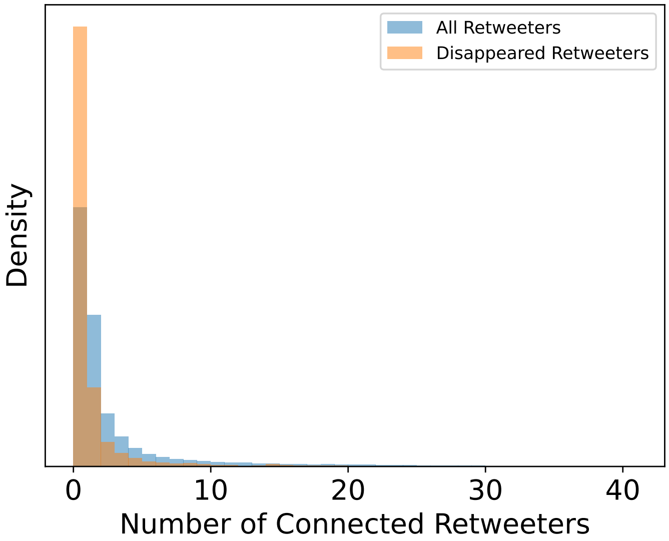


Figure 3.17: Two histograms comparing the number of follower-following connections for both all Retweeters and just the unretweeting Retweeters. We see unretweeting users are less connected.

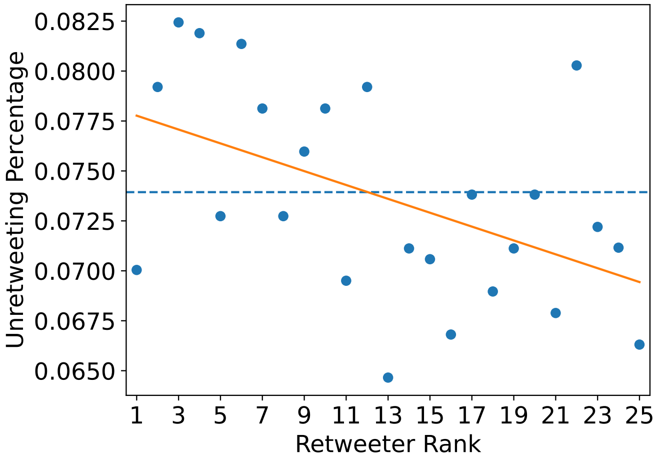


Figure 3.18: A plot of the relationship between the rank of a Retweeter (how early the Retweet took place) and how often that Retweet is retracted.

We propose two explanation for this phenomenon. First, it could be that the marginal impact of an additional Retweet diminishes as a post gathers more Retweets. For example, this impact could be the likelihood of the author noticing the Retweet or an effect on the probability of the Tweet becoming viral. Thus, when a controversial Tweet had few Retweets, a user might have been willing to Retweet it because he believes the impact he has exceeds the social cost of Retweeting. Yet, he might decide to not Retweet if the post had more Retweets. As the post gathered more Retweets, the calculus no longer makes sense for the Retweeter, and thus his Retweet is retracted. Second, early Retweeters might have a higher probability of being illegitimate users who are thus more likely to be suspended. When Tweets are first posted, authors have a stronger desire to have them gain Retweets, because it would allow the Tweet to have momentum and be recommended by the platform. Thus, it is more probable that these early Retweets are not from actual users. Conversely, once a Tweet has momentum, there is less incentive to add on fake Retweets.

Note, the average Retweet loss of  7% shown in the figure is higher than the base Retweet loss of around 1-4% shown in Section 3.1, because these look at the individual user's rate of unretweeting rather than the Retweet loss on the Tweet as a whole. The Retweet loss percentages from Section 3.1 are lower because they are offset by the new Retweeters which are gained over time. These characteristics could be useful for modelling this unretweeting phenomenon and simulating the behaviour as a game. We have begun modeling and exploring such a game, but will be looking to study it more rigorously in future work.

**Answer to RQ3:** We proposed a curve-reconstruction algorithm to produce the expected life-cycle of a popular Tweet. With that, we verified our prior that contentious Tweets were expected to lose a larger portion of their Retweets. We observed that network structure features like connectivity to other Retweeters and relationship to the author are related to the probability of unretweeting, with more connected users being less likely to unretweet. Furthermore, we observe a phenomenon that earlier Retweeters are more likely to unretweet, and proposed some explanations.

# Chapter 4

# Discussion

In this thesis, we presented a better understanding of how self-moderation and platform moderation work by examining the way the Retweet counts of popular Tweets diminish over time. To do so, we studied Kaggle datasets and also collected our own data during a highly volatile period on Twitter. We tracked Retweeters over time and examined the distribution across various reasons for unretweeting. We developed a model to predict which Tweets would lose the most Retweets. We studied the role that a user's network structure plays in determining whether she will be an unretweeter. In addition, we also conducted, in a separate work [38], a user survey to better understand what types of content various users are more inclined to self-moderate.

In addition to our analysis of moderation, we also present a snapshot of Twitter during a transitional period, following Musk's acquisition. We observed an initial reduction in account suspensions, possibly due to lay-offs and changes in moderation policy. Furthermore, later in the year, when Twitter was reinstating suspended accounts, our own datasets captured this resurgence of returning Retweeters. Retweeting is seen as a form of endorsement [47], as the Retweeted content is attached to the user's profile. We initially saw greater Retweet loss for contentious content over uncontentious content, and suspected the retraction of Retweets to be a form of self-moderation. However, this difference was much less significant in our custom dataset (following the Musk acquisition) than in the Kaggle datasets. While contentious content still had higher Retweet loss, we realized this difference was due to a higher rate of account deactivations, as shown in Section 3.1.

This reduction in Retweet loss could be due to changes in Twitter platform moderation, such as automating moderation [66] and reinstating suspended users (see Appendix C). It could also be due to changes in self-moderation behaviour, such as increased account deactivations associated with migrating users [95, 42] following the acquistions or fewer Retweeters of controversial content feeling the need to self-moderate due to the now looser platform moderation policies [9].

## 4.1 Timeline of Events

To get a stronger sense as to the role of the acquisition played in our custom data collection process, we overlay in Figure 4.1 the major events following the acquisition and our data collection periods.

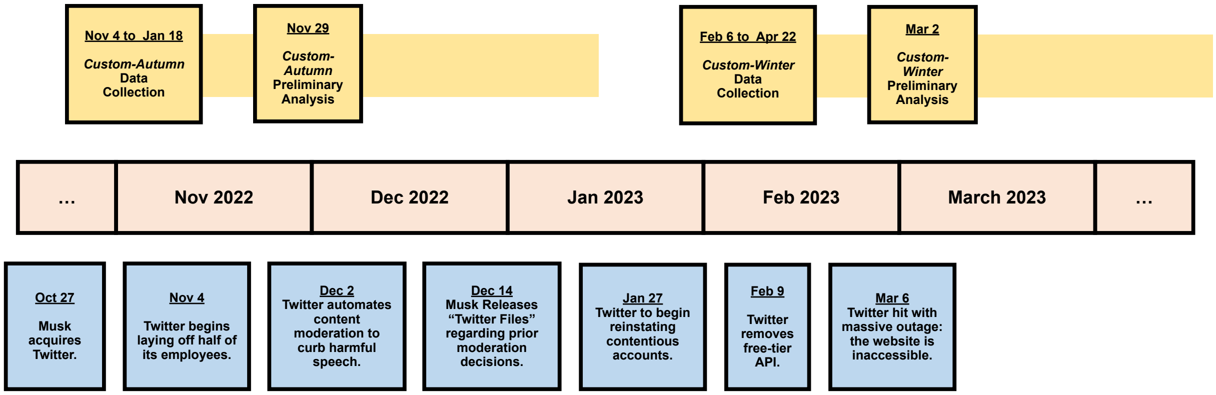

Nov 4 to Jan 18
Custom-Autumn Data Collection
Nov 29
Custom-Autumn Preliminary Analysis
Feb 6 to Apr 22
Custom-Winter Data Collection
Mar 2
Custom-Winter Preliminary Analysis
...
Nov 2022
Dec 2022
Jan 2023
Feb 2023
March 2023
...
Oct 27
Musk acquires Twitter.
Nov 4
Twitter begins laying off half of its employees.
Dec 2
Twitter automates content moderation to curb harmful speech.
Dec 14
Musk Releases "Twitter Files" regarding prior moderation decisions.
Jan 27
Twitter to begin reinstating contentious accounts.
Feb 9
Twitter removes free-tier API.
Mar 6
Twitter hit with massive outage: the website is inaccessible.

Figure 4.1: A timeline of some major events following Elun Musk's acquisition of Twitter overlapping with our data collection periods. News releases relating to the major events are cited in the discussion below.

### 4.1.1 Timeline for the *Custom-Autumn* Dataset

Our first dataset, *Custom-Autumn*, consists of Tweets that were posted around when news about the acquisition [17] was first released on October 27, 2022. Soon after, on November 4th, Twitter announced lay offs, reducing its workforce by half [72]. Following these events, Twitter saw a rise in hate speech [10]. The preliminary data collected for *Custom-Autumn* dataset (one third of the way into the full data collection timeline) captured the account deactivations and the lack of moderation that occurred during this time period.

These initial events were followed by responses from Twitter focusing on platform moderation. Twitter focused on finding ways to automate the moderation process [66]. Their approach also less frequently removed contentious speech outright, but instead set restrictions on how that content is distributed on the platform. Twitter also released what is commonly deemed the "Twitter Files" [68], a set of internal documents outlining the considerations that had be given to previous high-profile moderation events. Musk's Twitter appeared to want to juxtapose themselves those previous policies. The remainder of our data collection period for *Custom-Autumn* overlapped with these events. Compared to the preliminary data, we saw a meaningful rise in platform moderation, as measured by the number of suspended accounts, in the latter two-thirds of the collection period.

### 4.1.2 Timeline for the *Custom-Winter* Dataset

On January 27, 2023, Twitter announced through a Tweet (shown in Appendix C), an intent to reinstate previously suspended accounts. Our second round of data collection for *Custom-Winter* took place soon after that announcement. When modelling the Retweet loss of this second round of collection, we observed the effects of this reinstatement, particularly amongst Retweeters of uncontentious Tweets. This difference is somewhat expected, as Figure 3.6 and 3.7 show account suspensions making up a larger portion of Retweet loss for uncontentious Tweets.

Throughout data collection for *Custom-Winter*, Twitter continued to be a volatile environment. For example, on March 6, 2023, Twitter experienced one of its biggest outages [24]. Another challenge in collecting for *Custom-Winter* is Twitter's Feb 3, 2023 announcement to start eliminating free access to their APIs on Feb 9, 2023 [8]. Following the acquisition, Twitter has lost some of its status as a suitable environment for academic research and discussion [48]. Following the announcement to remove free-tier APIs, this effect was intensified, as many academic projects became priced out. Without access to the academic APIs, these projects would have required enterprise-level access, which is priced at $42,000 per month. Ultimately, our Academic API access was revoked on April 27, 2023 (see Appendix B). Fortunately, the effect of this suspension was minimal for our project. We had already completed all of the data collection for *Custom-Winter*, except for fetching the status of unretweeting users in the final snapshot.

In Table 4.1, we present dates, description and a source discussing the events mentioned in this discussion and highlighted in Figure 4.1.

| Date | Event | Source |
|---|---|---|
| Oct 27, 2022 | Elon Musk completes his acquisition of Twitter. | New York Times [17] |
| Nov 4, 2022 | Twitter lays off half of its staff. | The Guardian [72] |
| Dec 2, 2022 | Twitter leans on automated processes to perform platform moderation. | Reuters [66] |
| Dec 14, 2022 | Twitter releases "Twitter Files", a set of internal documents regarding platform moderation. | Vox [68] |
| Jan 27, 2023 | Twitter introduces a process to reinstate suspended account, and an intent to actively review past suspensions. | Twitter Safety Tweet (Appendix C) |
| Feb 3, 2023 | Free-tier access to Twitter's APIs is eliminated, which includes the academic access tier. | Forbes [8] |
| Mar 6, 2023 | Twitter experiences one of its largest outages since Elon Musk's takeover. | CNN [24] |

Table 4.1: A list of relevant events which took place during the volatile period on Twitter overlapping with our data collection process.

## 4.2 Future Work

An important area of future research is to better understand the Retweet loss due to self-moderation. In particular, we are interested in explaining the self-moderation case where the user account is still active and public, but the Retweet was manually retracted (the fourth category in Section 3.1). A preliminary analysis of this behaviour shows that this retraction is not uniformly distributed across all Tweets. Instead, it follows more of a power law distribution, which would not be expected if this retraction was random. It would be interesting to examine whether there are certain topics that are more likely to experience this form of self-moderation. In particular, we would like to examine whether there is a connection between the topics experiencing more moderation and the hierarchy of self-censored content we examined in our previous user studies [38]. The motivation behind Retweet loss could also be better understand by taking more frequent snapshots of the users' unretweeting reasons. From those snapshots, we would be able to better study the delay in self-moderation.

In terms of predicting Retweet loss, there are other potential features that would be worth examining, such as the author's profile description and the content shown in the author's profile picture. For example, it might be possible to infer the author's gender and age from the profile picture. In addition, the nature of being verified has changed following the acquisition. The nature of this change should be examined, as the data our current model is trained on is prior to the acquisition.

Another avenue of future research involves building on the observations in Section 3.3.2 and 3.2.1 to explicitly simulate the unretweeting behaviour. The model could capture various Retweeting and self-moderation incentives, such as relationship to the author, relationship to other Retweeters as measured by their network structure, and the diminished impact of an additional Retweet as the Retweet count gets higher. Since Retweeting can be seen as a collaborative exercise to spread certain information, this has similarities with existing work studying multi-agent collaboration on social networks [5].

Lastly, methodology similar to our current work can be extended to other platforms, such as Reddit, where there is meaningful moderation. For example, the same curve-reconstruction algorithm in Section 3.3 can be applied to track total number of comments on a Reddit thread, where both self- and platform moderation play an important role. In particular, Reddit discussions are overseen by volunteer moderators, which play an active role in removing comments. Furthermore, its upvote and downvote systems allow users a real-time sense of how their comments received, providing greater incentive for self-moderation. Reddit, like Twitter, has generally made access to its API much more difficult [35, 16]; however, it is committed to preserving free-access of its academic APIs.

## 4.3 Conclusion

In this thesis, we analyzed a surprising phenomenon regarding Retweets on Twitter: popular Tweets see significant Retweet loss over time. In particular, contentious content saw greater Retweet loss than uncontentious content. To investigate the role that self-moderation and platform moderation plays in that Retweet loss, we built our own dataset to study the phenomenon.

We saw that the most common reason for Retweet loss was account suspension, a form of platform moderation. From there, users making their accounts private and users deactivating their accounts made up a similar proportion, with Retweeters of contentious content being more likely to deactivate their accounts. We also illustrated how, as time passes, Retweet loss dominates over new Retweets gained. From there, we designed features based on sentiment labels from Google Cloud API, Tweet metadata, author characteristics, and the Tweet text itself, to predict which Tweets would lose the most Retweets. We trained a logistic regression model using a Kaggle dataset of *Covid-19* Tweets. The model performed well out-of-sample, and even generalized to newer datasets we collected ourselves.

This generalization to datasets of a different era and environment gives faith that although online content moderation is continually evolving, something fundamental was captured. Also, since the features that are used to make the prediction are known at the time the Tweet is posted, they can provide actionable steps for the author, especially those who are interested in attracting a long-term following (i.e., those who will not unretweet). An example of such an action could be to make the sentiment of the Tweet more positive. From there, we presented a curve-reconstruction algorithm to model the timeline of Retweet loss for contentious and uncontentious Tweets. Furthermore, we show that the expected curve of the Retweet loss for contentious Tweets is steeper than that of uncontentious ones. We also modelled individual Retweeters, highlighting in particular the relationship between user network-structure and Retweet loss. In particular, users who are more connected to the author and to the other Retweeters are less likely to unretweet.

Twitter runs on Retweets, which is arguably its fundamental differentiator. Retweeting allows users to quickly endorse an idea with a single click, enabling the platform to be a source of timely information as well as a town-hall for debate. This thesis presents a first attempt at understanding the circumstances around which this endorsement is retracted. The hope is that these observations can be used to better model and simulate the moderation of contentious content on social networks.

# References

[1] Darel Nicol Luna Anak Agam. Followers Ratio on Instagram Affects the Product's Brand Awareness. *Australian Journal of Accounting, Economics and Finance (AJAEF)*, 3(2):86, 2017.

[2] Vezir Aktas, Marco Nilsson, and Klas Borell. Social scientists under threat: Resistance and self-censorship in Turkish academia. *British Journal of Educational Studies*, 67(2):169–186, 2019.

[3] Meysam Alizadeh, Fabrizio Gilardi, Emma Hoes, K Jonathan Klüser, Mael Kubli, and Nahema Marchal. Content Moderation As a Political Issue: The Twitter Discourse Around Trump's Ban. *Journal of Quantitative Description: Digital Media*, 2, 2022.

[4] Hazim Almuhimedi, Shomir Wilson, Bin Liu, Norman Sadeh, and Alessandro Acquisti. Tweets are Forever: A Large-Scale Quantitative Analysis of Deleted Tweets. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, pages 897–908, 2013.

[5] Ben Armstrong. Coordination in a Peer Production Platform: A study of Reddit's /r/Place experiment. Master's thesis, University of Waterloo (Waterloo), 2018.

[6] Mossaab Bagdouri and Douglas W Oard. On Predicting Deletions of Microblog Posts. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 1707–1710, 2015.

[7] Daniel Bar-Tal. Self-censorship as a socio-political-psychological phenomenon: Conception and research. *Political Psychology*, 38:37–65, 2017.

[8] Jenae Barnes. Twitter Ends Its Free API: Here's Who Will Be Affected. *Forbes*, 2023.

[9] Christopher Barrie. Did the Musk Takeover Boost Contentious Actors on Twitter? *arXiv preprint arXiv:2212.10646*, 2022.

[10] Bond Benton, Jin-A Choi, Yi Luo, and Keith Green. Hate Speech Spikes on Twitter After Elon Musk Acquires the Platform. *School of Communication and Media, Montclair State University*, 2022.

[11] Parantapa Bhattacharya and Niloy Ganguly. Characterizing Deleted Tweets and Their Authors. In *Proceedings of the 10th International AAAI Conference on Web and Social Media (ICWSM)*, 2016.

[12] Devan Bissonette. "Modern Day Presidential:" Donald Trump and American Politics in the Age of Twitter. *The Journal of Social Media in Society*, 9(1):180–206, 2020.

[13] Svetlana S Bodrunova, Anna Litvinenko, and Kamilla Nigmatullina. Who is the censor? Self-censorship of Russian journalists in professional routines and social networking. *Journalism*, 22(12):2919–2937, 2021.

[14] Danah Boyd, Scott Golder, and Gilad Lotan. Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter. In *Proceedings of the 2010 43rd Hawaii International Conference on System Sciences*, pages 1–10. IEEE, 2010.

[15] Anne Shann Yue Cheung. *Self-censorship and the struggle for press freedom in Hong Kong*. Brill Nijhoff, 2021.

[16] Bary Collins. Reddit Strike One Week On: A Third Of Subreddits Still Down. *Forbes*, 2023.

[17] Kate Conger and Lauren Hirsch. Elon Musk Completes $44 Billion Deal to Own Twitter. *New York Times*, 2022.

[18] Michael Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Gonçalves, Filippo Menczer, and Alessandro Flammini. Political polarization on twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 5, pages 89–96, 2011.

[19] Francesco Corea. Can Twitter Proxy the Investors' Sentiment? The Case for the Technology Sector. *Big Data Research*, 4:70–74, 2016.

[20] Sauvik Das and Adam Kramer. Self-censorship on Facebook. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 7(1), pages 120–127, 2013.

[21] Vasant Dhar and Elaine A Chang. Does Chatter Matter? The Impact of User-Generated Content on Music Sales. *Journal of Interactive Marketing*, 23(4):300–307, 2009.

[22] Mustafa Doğan, Ömer Metin, Elif Tek, Semih Yumuşak, and Kasım Öztoprak. Speculator and Influencer Evaluation in Stock Market by Using Social Media. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 4559–4566, 2020.

[23] Mark Dredze, Prabhanjan Kambadur, Gary Kazantsev, Gideon Mann, and Miles Osborne. How Twitter is Changing the Nature of Financial News Discovery. In *Proceedings of the Second International Workshop on Data Science for Macro-Modeling*, pages 1–5, 2016.

[24] Clare Duffy. Twitter hit with one of the biggest outages since Elon Musk took over. *CNN*, 2023.

[25] Abdelmalek El Kadoussi. The perception of self-censorship among Moroccan journalists. *The Journal of North African Studies*, 27(2):231–263, 2022.

[26] Chao Fan, Yucheng Jiang, Yang Yang, Cheng Zhang, and Ali Mostafavi. Crowd or Hubs: Information diffusion patterns in online social networks in disasters. *International Journal of Disaster Risk Reduction*, 46:101498, 2020.

[27] Matthew Feeney and Will Duffield. Trump's Truth Social Rejects Free Speech, For Good Reason. 2022.

[28] Casey Fiesler and Nicholas Proferes. "Participant" Perceptions of Twitter Research Ethics. *Social Media + Society*, 4(1):2056305118763366, 2018.

[29] Rudolph Flesch. A New Readability Yardstick. *Journal of Applied Psychology*, 32(3):221, 1948.

[30] Deen Freelon and Tetyana Lokot. Russian Twitter disinformation campaigns reach across the American political spectrum. *Misinformation Review*, 2020.

[31] Anna Lena Füllsack. Guidelines 8/2020 on the targeting of social media users - Version 1.0. *Frontex: European Border and Coast Guard Agency*, 2020.

[32] Amira Ghenai. *Health Misinformation in Search and Social Media*. PhD thesis, University of Waterloo (Canada), 2019. pp. 95.

[33] James L Gibson and Joseph L Sutherland. Keeping Your Mouth Shut: Spiraling Self-Censorship in the United States. *Available at SSRN 3647099*, 2020.

[34] Kristina Gligorić, Ashton Anderson, and Robert West. How constraints affect content: The case of Twitter's switch from 140 to 280 characters. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12, 2018.

[35] Wyatte Grantham-Philips. The Reddit blackout, explained: Why thousands of subreddits are protesting third-party app charges. *Associated Press*, 2023.

[36] Jacob Groshek and Chelsea Cutino. Meaner on Mobile: Incivility and Impoliteness in Communicating Contentious Politics on Sociotechnical Networks. In *Proceedings of the 7th 2016 International Conference on Social Media & Society*, pages 1–7, 2016.

[37] Mahmud Hasan, Mehmet A Orgun, and Rolf Schwitter. Real-time event detection from the Twitter data stream using the TwitterNews+ Framework. *Information Processing & Management*, 56(3):1146–1165, 2019.

[38] Wei Hu and Diogo Barradas. Work in Progress: A Glance at Social Media Self-Censorship in North America. In *Proceedings of the 2023 Workshop on Socio-Technical Aspects in Security and Trust (To Appear)*. IEEE, 2023.

[39] Abraham Israeli and Oren Tsur. Free speech or Free Hate Speech? Analyzing the Proliferation of Hate Speech in Parler. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 109–121, Seattle, Washington (Hybrid), July 2022. Association for Computational Linguistics.

[40] Swati Jain, Suraj Prakash Narayan, Rupesh Kumar Dewang, Utkarsh Bhartiya, Nalini Meena, and Varun Kumar. A Machine Learning based Depression Analysis and Suicidal Ideation Detection System using Questionnaires and Twitter. In *2019 IEEE Students Conference on Engineering and Systems (SCES)*, pages 1–6. IEEE, 2019.

[41] Maximilian Jenders, Gjergji Kasneci, and Felix Naumann. Analyzing and Predicting Viral Tweets. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 657–664, 2013.

[42] Ujun Jeong, Paras Sheth, Anique Tahir, Faisal Alatawi, H Russell Bernard, and Huan Liu. Exploring Platform Migration Patterns between Twitter and Mastodon: A User Behavior Study. *arXiv preprint arXiv:2305.09196*, 2023.

[43] Qingwen Jia and Suqi Xu. An Overall Analysis of Twitter and Elon Musk M&A Deal. *Highlights in Business, Economics and Management*, 2:436–441, 2022.

[44] Maria Renee Jimenez-Sotomayor, Carolina Gomez-Moreno, and Enrique Soto-Perez-de Celis. Coronavirus, ageism, and Twitter: An evaluation of tweets about older adults and COVID-19. *Journal of the American Geriatrics Society*, 68(8):1661–1665, 2020.

[45] Andreas M Kaplan and Michael Haenlein. The early bird catches the news: Nine things you should know about micro-blogging. *Business Horizons*, 54(2):105–113, 2011.

[46] Ben Kepes. How To Kill Your Ecosystem. Twitter Pulls An Evil Move With Its Firehose. *Forbes*, 2015.

[47] Jihie Kim and Jaebong Yoo. Role of Sentiment in Message Propagation: Reply vs. Retweet Behavior in Political Communication. In *Proceedings of the 2012 International Conference on Social Informatics*, pages 131–136. IEEE, 2012.

[48] Kai Kupferschmidt. As Musk reshapes Twitter, academics ponder taking flight. *Science (New York, NY)*, 378(6620):583–584, 2022.

[49] Rabindra Lamsal. Coronavirus (COVID-19) Tweets Dataset. IEEE Dataport. 2020.

[50] Anna Grøndahl Larsen, Ingrid Fadnes, and Roy Krøvel. *Journalist Safety and Self-Censorship.* Routledge, 2021.

[51] James M Leonhardt and Igor Makienko. Keep It Simple, Readability Increases Engagement on Twitter: An Abstract. In *Back to the Future: Using Marketing Basics to Provide Customer Value: Proceedings of the 2017 Academy of Marketing Science (AMS) Annual Conference*, pages 333–334. Springer, 2018.

[52] Gilad Lotan, Erhardt Graeff, Mike Ananny, Devin Gaffney, Ian Pearce, et al. The Arab Spring— The Revolutions Were Tweeted: Information Flows during the 2011 Tunisian and Egyptian Revolutions. *International journal of communication*, 5:31, 2011.

[53] Yao Lu, Peng Zhang, Yanan Cao, Yue Hu, and Li Guo. On the Frequency Distribution of Retweets. *Procedia Computer Science*, 31:747–753, 2014.

[54] Brandon Lwowski, Paul Rad, and Kim-Kwang Raymond Choo. Geospatial Event Detection by Grouping Emotion Contagion in Social Media. *IEEE Transactions on Big Data*, 6(1):159–170, 2018.

[55] Renkai Ma and Yubo Kou. "How advertiser-friendly is my video?": YouTuber's Socioeconomic Interactions with Algorithmic Content Moderation. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–25, 2021.

[56] Lindsey Meeks. Tweeted, deleted: theoretical, methodological, and ethical considerations for examining politicians' deleted tweets. *Information, Communication & Society*, 21(1):1–13, 2018.

[57] Mohsen Mosleh, Qi Yang, Tauhid Zaman, Gordon Pennycook, and David G Rand. Analysis: Trade-offs between reducing misinformation and politically-balanced enforcement on social media, 2022.

[58] Samuel David Mueller and Marius Saeltzer. Twitter made me do it! Twitter's tonal platform incentive and its effect on online campaigning. *Information, Communication & Society*, 25(9):1247–1272, 2022.

[59] Rishab Nithyanand, Brian Schaffner, and Phillipa Gill. Measuring Offensive Speech in Online Political Discourse. In *Proceedings of the 7th USENIX Workshop on Free and Open Communications on the Internet (FOCI 17)*, 2017.

[60] Pippa Norris. Closed Minds? Is a 'Cancel Culture' Stifling Academic Freedom and Intellectual Debate in Political Science? Technical report, HKS Working Paper No. RWP20-025, Available at SSRN: https://ssrn.com/abstract=3671026, 2020.

[61] Petra Kralj Novak, Luisa De Amicis, and Igor Mozetič. Impact investing market on Twitter: influential users and communities. *Applied network science*, 3:1–20, 2018.

[62] Elvin Ong. Online repression and self-censorship: Evidence from southeast Asia. *Government and Opposition*, 56(1):141–162, 2021.

[63] Miles Osborne and Mark Dredze. Facebook, Twitter and Google Plus for breaking news: Is there a winner? In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 8, pages 611–614, 2014.

[64] Yu Ouyang and Richard W Waterman. *Trump, Twitter, and the American Democracy: Political Communication in the Digital Age (The Evolving American Presidency).* Springer Nature, 2020.

[65] Indraneil Paul, Abhinav Khattar, Ponnurangam Kumaraguru, Manish Gupta, and Shaan Chopra. Elites Tweet? Characterizing the Twitter Verified User Network. In *2019 IEEE 35th International Conference on Data Engineering Workshops (ICDEW)*, pages 278–285. IEEE, 2019.

[66] Katie Paul and Sheila Dang. Exclusive: Twitter leans on automation to moderate content as harmful speech surges. *Reuters*, 2022.

[67] Elia Powers, Michael Koliska, and Pallavi Guha. "Shouting Matches and Echo Chambers": Perceived Identity Threats and Political Self-Censorship on Social Media. *International Journal of Communication*, 13:20, 2019.

[68] Andrew Prokop. Why the Twitter Files actually matter. *Vox*, 2022.

[69] Ankita Rane and Anand Kumar. Sentiment Classification System of Twitter Data for US Airline Service Analysis. In *2018 IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC)*, volume 1, pages 769–773. IEEE, 2018.

[70] Hans Rosenberg, Shahbaz Syed, and Salim Rezaie. The Twitter pandemic: The critical role of Twitter in the dissemination of medical information and misinformation during the COVID-19 pandemic. *Canadian Journal of Emergency Medicine*, 22(4):418–421, 2020.

[71] Alesia Rudnik. Why do bloggers keep silent? Self-censorship in social media: cases of Belarus and Russia. Master's thesis, Södertörn University (Sweden), 2020.

[72] Dominic Rushe, Gloria Oladipo, Johana Bhuiyan, Dan Milmo, and Joe Middleton. Twitter slashes nearly half its workforce as Musk admits 'massive drop' in revenue. *The Guardian*, 2022.

[73] K Sathiyakumari and MS Vijaya. Community Detection Based on Girvan Newman Algorithm and Link Analysis of Social Media. In *Digital Connectivity–Social Impact: 51st Annual Convention of the Computer Society of India, CSI 2016, Coimbatore, India, December 8-9, 2016, Proceedings 51*, pages 223–234. Springer, 2016.

[74] Adam Schunk. An Analysis on The Network Structure of Influential Communities in Twitter. Master's thesis, University of Waterloo (Waterloo), 2019.

[75] Filipo Sharevski, Alice Huff, Peter Jachim, and Emma Pieroni. (Mis) perceptions and engagement on Twitter: COVID-19 vaccine rumors on efficacy and mass immunization effort. *International Journal of Information Management Data Insights*, 2(1):100059, 2022.

[76] Anil Kumar Singh and Pratya Goyal. A Language Identification Method Applied to Twitter Data. *TweetLID@ SEPLN*, 2014:26–29, 2014.

[77] Manya Sleeper, Rebecca Balebako, Sauvik Das, Amber Lynn McConahy, Jason Wiese, and Lorrie Faith Cranor. The Post that Wasn't: Exploring Self-Censorship on Facebook. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, pages 793–802, 2013.

[78] Manya Sleeper, Justin Cranshaw, Patrick Gage Kelley, Blase Ur, Alessandro Acquisti, Lorrie Faith Cranor, and Norman Sadeh. "I read my Twitter the next morning and was astonished" A Conversational Perspective on Twitter Regrets. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 3277–3286, 2013.

[79] Bongwon Suh, Lichan Hong, Peter Pirolli, and Ed H. Chi. Want to be Retweeted? Large Scale Analytics on Factors Impacting Retweet in Twitter Network. In *2010 IEEE Second International Conference on Social Computing*, pages 177–184, 2010.

[80] John Suler. The Online Disinhibition Effect. *Cyberpsychology & Behavior*, 7(3):321–326, 2004.

[81] Alecia Swasy. *How Journalists Use Twitter: The Changing Landscape of U.S. Newsrooms*. Lexington Books, 2016.

[82] Viriya Taecharungroj. Starbucks' marketing communications strategy on Twitter. *Journal of Marketing Communications*, 23(6):552–571, 2017.

[83] Rima Tanash, Zhouhan Chen, Dan Wallach, and Melissa Marschall. The Decline of Social Media Censorship and the Rise of Self-Censorship after the 2016 Failed Turkish Coup. In *Proceedings of the 7th USENIX Workshop on Free and Open Communications on the Internet*, 2017.

[84] Edson C Tandoc Jr and Erika Johnson. Most students get breaking news first from Twitter. *Newspaper Research Journal*, 37(2):153–166, 2016.

[85] Io Taxidou, Peter M Fischer, Tom De Nies, Erik Mannens, and Rik Van de Walle. Information Diffusion and Provenance of Interactions in Twitter: Is it only about Retweets? In *Proceedings of the 25th International Conference Companion on World Wide Web*, pages 113–114, 2016.

[86] Onur Varol. Should we agree to disagree about Twitter's bot problem? *arXiv preprint arXiv:2209.10006*, 2022.

[87] Ethan Ward. Parlez-vous le hate?: Examining topics and hate speech in the alternative social network Parler. Master's thesis, University of Waterloo, 2021.

[88] Mark Warner and Victoria Wang. Self-censorship in social networking sites (SNSs)–privacy concerns, privacy awareness, perceived vulnerability and information management. *Journal of Information, Communication and Ethics in Society*, 2019.

[89] Stefan Wojcik and Adam Hughes. Sizing Up Twitter Users. *PEW research center*, 24:1–23, 2019.

[90] Yiping Xia, Josephine Lukito, Yini Zhang, Chris Wells, Sang Jung Kim, and Chau Tong. Disinformation, performed: Self-presentation of a Russian IRA account on Twitter. *Information, Communication & Society*, 22(11):1646–1664, 2019.

[91] Kai-Cheng Yang, Francesco Pierri, Pik-Mai Hui, David Axelrod, Christopher Torres-Lugo, John Bryden, and Filippo Menczer. The COVID-19 Infodemic: Twitter versus Facebook. *Big Data & Society*, 8(1):20539517211013861, 2021.

[92] Savvas Zannettou. "I Won the Election!": An Empirical Analysis of Soft Moderation Interventions on Twitter. In *Proceedings of the 15th International AAAI Conference on Web and Social Media (ICWSM 2021)*, pages 865–876, 2021.

[93] Savvas Zannettou, Barry Bradlyn, Emiliano De Cristofaro, Haewoon Kwak, Michael Sirivianos, Gianluca Stringini, and Jeremy Blackburn. What is Gab? A Bastion of Free Speech or an Alt-Right Echo Chamber? In *Proceedings of the The Web Conference 2018*, pages 1007–1014, 2018.

[94] Lu Zhou, Wenbo Wang, and Keke Chen. Tweet Properly: Analyzing Deleted Tweets to Understand and Identify Regrettable Ones. In *Proceedings of the 25th International Conference on World Wide Web*, pages 603–612, 2016.

[95] Haris Bin Zia, Jiahui He, Aravindh Raman, Ignacio Castro, Nishanth Sastry, and Gareth Tyson. Flocking to Mastodon: Tracking the Great Twitter Migration. *arXiv preprint arXiv:2302.14294*, 2023.

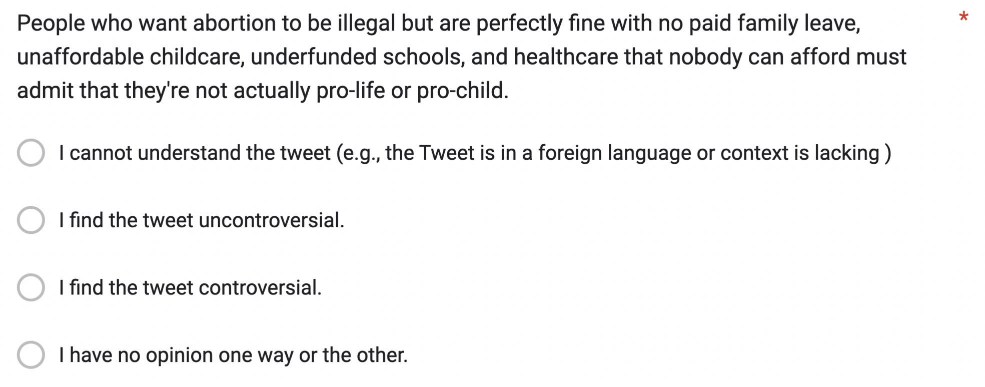
# APPENDICES

# Appendix A

# Tweet Contentiousness User Survey

In this section, we set out to validate the search queries chosen for our *Custom-Winter* dataset. To form that dataset, we had two sets of search terms. One of the sets was meant to lead to contentious Tweets that make up $D_c$, while the other set was supposed to yield $D_u$, a dataset of uncontentious Tweets.

To verify that $D_c$ and $D_u$ indeed differed meaningfully in their level of contention, we created a Google Form questionnaire consisting of 100 items, where each item is a randomly sampled Tweet (50 comes from each of $D_u$ and $D_c$). Respondents were asked to label each Tweet as either controversial or uncontroversial.



People who want abortion to be illegal but are perfectly fine with no paid family leave, unaffordable childcare, underfunded schools, and healthcare that nobody can afford must admit that they're not actually pro-life or pro-child. *

○ I cannot understand the tweet (e.g., the Tweet is in a foreign language or context is lacking )

○ I find the tweet uncontroversial.

○ I find the tweet controversial.

○ I have no opinion one way or the other.

Figure A.1: Respondents are shown 100 randomly sampled Tweets, 50 from the contentious dataset and 50 from the uncontentious dataset. For each Tweet, users are presented with four options: cannot understand the Tweet, the Tweet is controversial, the Tweet is uncontroversial, no opinion.

Respondents are also given the options to not express an opinion, or to indicate that they could not understand the Tweet. The four options available to the respondents are shown in Figure A.1. Respondents are prompted to view controversiality as being likely to lead to contentious and emotional invested arguments. An innocuous topic however divisive (e.g., do pineapples belong on pizza) would not be considered controversial.

For this sanity check, we gathered n=17 participants. Of the 17 x 100 = 1700 responses, there were 356 answers that were either "cannot understand the Tweet" or "no opinion." $D_c$ and $D_u$ saw a roughly equal number of "no opinions" (115 vs. 120 respectively), but the number of "cannot understands" were much higher for uncontentious Tweets (89 vs. 32). Removing these, we are left with 703 opinionated responses for $D_c$ and 641 for $D_u$.

Of the opinionated responses, we saw that for $D_c$ that 54.8% of the responses were for the "controversial" option, while this value is 23.4% for $D_u$. We are reassured in seeing that both point estimates are located in the 95% confidence interval from *Custom-Autumn* in Section 2.3.1. In the bar graphs of Figure A.2, we show the raw number of votes for each of the four options, aggregated across all items in the questionnaire.

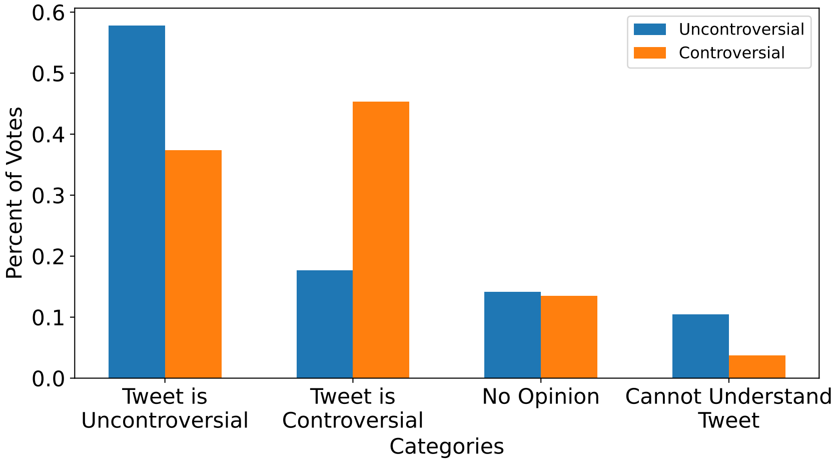


Figure A.2: The percentage of votes across the four multiple choice options in the questionnaire, for $D_u$ and $D_c$.

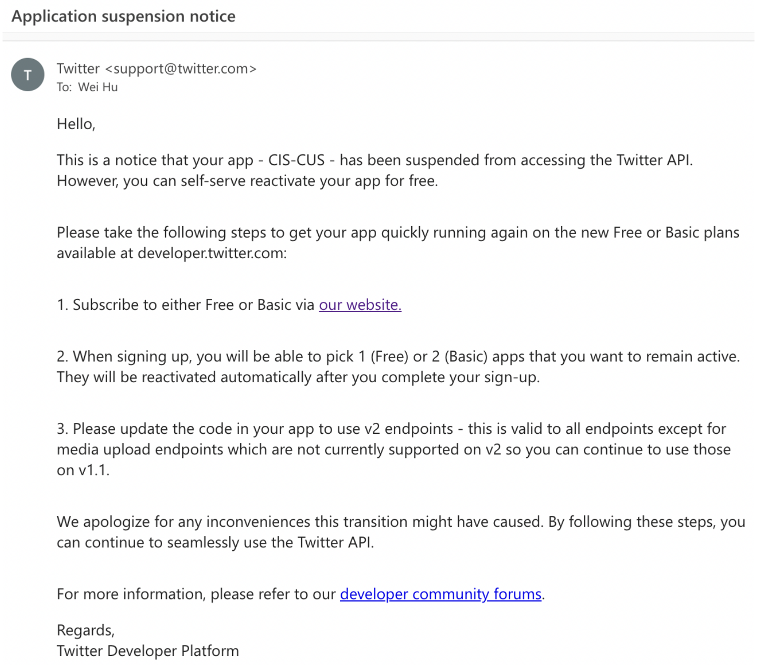
For more rigorous analysis, we examined the votes on each individual Tweet, rather than all the votes in aggregate. For each Tweet, if there are more "controversial" votes than "uncontroversial" votes, then we label the Tweet "controversial". We form a binary vector of length 100 where, where index $i$ is set to 1 if Tweet $i$ is "controversial". We create another vector of length 100 where index $i$ is set 1 if Tweet $i$ is from $D_c$. From there, we take the dot product of the two vector.

Under the null hypothesis that is no difference in contention between the $D_u$ and $D_c$, we would expect the value of the dot product to be 50, with $\sigma$ of 5 (we essentially have a $Binomial(100, 0.5)$ distribution). The true value we obtained was 73, which corresponds to a p-value of $\approx$2E-6.

# Appendix B

# Twitter Revokes Academic APIs

Beginning in early February, Twitter made access to its APIs much more restrictive [8]. Free-tier access was eliminated, and the academic research access to Twitter was also gradually eliminated. Our access to the APIs was revoked on April 27, 2023 after receiving the email shown in Figure B.1.



Application suspension notice

Twitter <support@twitter.com>
To: Wei Hu

Hello,

This is a notice that your app - CIS-CUS - has been suspended from accessing the Twitter API. However, you can self-serve reactivate your app for free.

Please take the following steps to get your app quickly running again on the new Free or Basic plans available at developer.twitter.com:

1. Subscribe to either Free or Basic via our website.

2. When signing up, you will be able to pick 1 (Free) or 2 (Basic) apps that you want to remain active. They will be reactivated automatically after you complete your sign-up.

3. Please update the code in your app to use v2 endpoints - this is valid to all endpoints except for media upload endpoints which are not currently supported on v2 so you can continue to use those on v1.1.

We apologize for any inconveniences this transition might have caused. By following these steps, you can continue to seamlessly use the Twitter API.

For more information, please refer to our developer community forums.

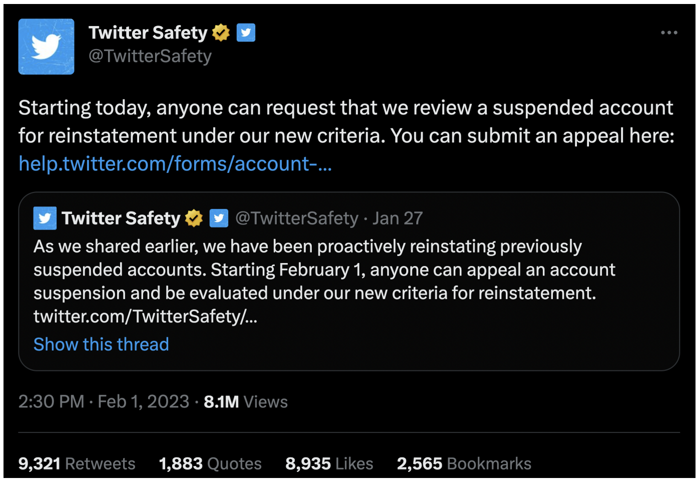
Regards,
Twitter Developer Platform

Figure B.1: Email from Twitter, as part of eliminating academic access to its APIs.

# Appendix C

# Twitter to Reinstate Suspended Accounts

In late January, Twitter announced its intent to reinstate previously suspended accounts. It also allowed users to appeal their account suspensions.



Twitter Safety
@TwitterSafety

Starting today, anyone can request that we review a suspended account for reinstatement under our new criteria. You can submit an appeal here: help.twitter.com/forms/account-...

Twitter Safety @TwitterSafety · Jan 27
As we shared earlier, we have been proactively reinstating previously suspended accounts. Starting February 1, anyone can appeal an account suspension and be evaluated under our new criteria for reinstatement. twitter.com/TwitterSafety/...
Show this thread

2:30 PM · Feb 1, 2023 · 8.1M Views

9,321 Retweets 1,883 Quotes 8,935 Likes 2,565 Bookmarks

Figure C.1: Twitter announced accounts suspended under its previous moderation policies would be reinstated, and that users could appeal their suspensions.

In our custom datasets, we can see the effect of this reinstatement. The shaded areas in Figure 3.16 and Figure 3.7 show a meaningful return in users that have previously unretweeted.