Drexel University

College of Computing and Informatics

Department of Computer Science

Look, Don't Tweet

Representation Learning and Social Media

Author

Hunter Heidenreich

Advisor

Dr. Jake Ryland Williams

August 31, 2025

Contents

Chapter 1

Introduction

It is hard to state with precision (at least at this moment in time) exactly how social media is transforming our present-day society. The rate at which information can spread is drastically changing as well as how people interact with and consume it. Embedded in this change are new modes in which people connect, interact, socialize, and conceptualize their roles in society.

Alfred Whitehead once wrote that "the major advances in civilization are processes which all but wreck the societies in which they occur" [1], a statement that feels relevant to any society entrenched in disruptive change. One need not look very far to observe the social problems that social media is helping to call to the surface; problems like misinformation [2–5], disinformation [2, 3, 5], social manipulation [3, 6, 7], filter bubbles [8–11], political radicalization [9, 12], and hate speech [13, 14] have become prominent issues that must be reckoned with, at some point or another.

In considering these issues as societal disruption, one may be reminded of Hans Jonas's writing on technology, ethics, and responsibility. In "Technology and Responsibility: Reflections on the New Tasks of Ethics," Jonas makes an argument for more forethought and caution when producing powerful and disruptive technology [15]. As he notes, in this modern era for the first time ever, technology has the power to destroy the Earth and humanity.

While social media may not invite the destruction of the Earth nor humanity, this mode of caution and responsibility does point towards a more critical lens through which the problems of social media may be analyzed. Furthermore, it offers an assortment of questions one could ask like "is social media changing people and society, and in what ways exactly?" and "if social media is changing individuals, is it aligned with how individuals would like to be changed?" These kinds of questions call back to what media researchers and philosophers like Marshall McLuhan attempted to call attention to through sayings like "the medium is the message" [16], whereby one may understand that often the changes brought on by a shift in a medium may be more subtle than the contents of the message itself.

All of that said, there's no closing Pandora's Box; social media and all of its benefits and harms are here and are unlikely to miraculously fade away with little thought given to their resolutions. Instead, it is of the utmost importance that tools are created to better understand social media and the way it is influencing (and at times, altering) individuals

and societies. This thesis fits in neatly with that need.

Simultaneously with the rapid shifts in communication media, the field of natural language processing (NLP) has seen prominent shifts in the way problems are approached. There has been wide-spread adoption of pre-trained models that learn language representations that may be adapted and tuned to specific, downstream use-cases. These so-called pre-trained models (PTMs) are first trained on a large corpus to learn "universal" language representation so that, when approaching a specific task or problem, one simply needs to adjust these representations at a much lower training cost [17]. This paradigm has enabled the rapid development of NLP applications, increasing performance in a broad cross-section of use-cases.

A critical aspect that has enabled the success of PTMs is the fact that they can be trained in an unsupervised or semi-supervised fashion, allowing for the creation of useful representations simply by having a large corpus to train on. This is, in part, because text has useful structure that can be taken advantage of as a training signal; for example, predicting a word given a surrounding context (either by masking or in isolation) [18–20].

A curious phenomenon has occurred because of the need for more data: Social media and internet-based data has been increasingly prominent in heterogeneous datasets such as those used to train large language models like GPT-2 [21] and RoBERTa [22]. On one hand, this provides an interesting opportunity to investigate the differences in language representation based on the media it was conceptualized in (e.g. print versus a Tweet). On another, it invites difficult questions about the ethics of doing so when considering aspects like privacy, consent, and representation. Additionally, if one is aligned with philosophers like Marshall McLuhan that "the medium is the message" [16] then there is a significant question that must be asked as to whether the strategy of mixing a heterogeneous media is ideal.

If one restricts their focus to social media, there is a variety of additional signals that one might be able to use for pre-training, aligning well with notions of semi-supervision and representation learning. Unlike traditional print text, language on social media occurs in a dynamic and social environment. There are a variety of social signals one may leverage that could yield better representations for studying, analyzing, and predicting events that occur online. This work provides the tools to begin to leverage such additional training signals.

As discussed in the next chapter, social media is an incredibly interesting domain, from a purely NLP perspective. It is noisy and features artifacts of the domain, like URLs, user mentions, and hashtags, that are entirely absent in general text. Additionally, it is a domain of high interest, particularly for understanding an online community, its beliefs, and its social structure. As much of public discourse appears on social media, it would be useful to have tools better equipped to deal with and understand its meaning–particularly in light of the aforementioned evolving social issues aided by social media.

The rest of this work is organized as follows: Chapter 2 calls to attention relevant background and related works, Chapter 3 begins a descriptive analysis of several social media platforms and datasets (culminating in a platform-agnostic data model and Python package for interacting with social conversations), Chapter 4 probes into a set of experiments to investigate the benefit of PTMs crafted for social media-related tasks, and Chapter 5

summarizes the totality of this work.

Chapter 2

Background and Related Work

2.1 Social Media as a Domain

2.1.1 As a Medium

In 1964, Canadian philosopher Marshall McLuhan coined the phrase "the medium is the message" highlighting the inherent influence that the structure of a medium has on how people communicate [16, 23].

Despite frequent misinterpretations of the phrase [24], what McLuhan sought to highlight was the fact that when analyzing media, research frequently focuses on the obvious, ignoring the non-obvious, dynamic components of a medium that ground communication (or otherwise provide context) and can lead to unintended consequences. Specifically, McLuhan emphasized looking beyond the message at the technological medium to better understand how changes affect inter-personal dynamics of communication. An example of this idea is the dissection of the rise of the modern 24/7 news cycle not being a result of the news stories in-and-of-themselves, but rather a shift in attitudes of the public perception of the news (and towards crime, as a by-product).

Social media is still a relatively new medium, but one that is clearly shifting inter-personal dynamics of communication. Developing the tools to not only understand, but to anticipate how social media is affecting society is critical for addressing the detrimental effects of a new media as well as anticipating, mitigating, and avoiding them entirely. As McLuhan advocated "Control over change would seem to consist in moving not with it but ahead of it. Anticipation gives the power to deflect and control force" [16, 23]. It is with such thoughts in mind that this work seeks to better develop the mechanisms that can enable such understanding and mitigation of negative effects.

2.1.2 As a Test Bed

Social media has a long and interesting history of being an environment to empirically study public discourse. This is especially true of political science, considering how public sentiments to the news, politics, and other world events can be extracted, qualified, and quantified [25–27].

Beyond using social media to simply observe discourse in an attempt to understand how

well theory about discourse matches empirical observations, one can also utilize social media as a social diagnostic environment. Previous works highlight such applications like tracking the spread of sickness, memes, and ideologies as all components that can be tracked, quantified, and understood [28]. All such applications, as well as addressing some of the issues social media faces, are further enabled by better representations of the textual content on social media that expose the latent differences that matter for such applications. Such representations are what this work seeks to create.

In a more abstract light, if one were to view society as a type of bio-social organism, the addition of communication technologies aided by artificial intelligence could, perhaps, be viewed under the framing of a techno-bio-social organism. Framed in this manner, it seems a natural step to use communication technologies as a test bed and diagnostic for social health, however that may be defined. Additionally, this view is inline with pushes in the nascent field of machine behavior [29] to characterize, measure, and qualify the effects of automation and artificial intelligence within the spheres of human activity.

2.1.3 As a Language Processing Domain

From a language processing standpoint, text on social media is extremely difficult to work with. It is littered with artifacts like emojis, hashtags, and URLs that wouldn't be present otherwise in mediums like published books. Additionally, people write informally on social media. Text can be very noisy, feature slang terminology, and contain many misspellings. All of this makes for an extremely difficult domain that requires models that are robust to these phenomena.

2.2 Neural Language Representation

2.2.1 Pre-Training

The usage of neural networks for learning continuous, dense vectors for words and sequences goes back to the early 2000s [30]. Neural network PTMs became increasingly popular after the development of inexpensive training of context-independent word representations becoming popular with algorithms like word2vec [18, 19], GloVe [31], and fastText [32]. Since then, context-dependent models have become the focus with the Transformer architecture [33] currently considered as state-of-the-art.

The history of representation learning for NLP is long and branching, but for a scoped survey of neural network-based PTMs and the differences between them, the authors of [17] provide an excellent survey. Across this broad cross-section of PTMs for NLP, a core theme of all of the highlighted models is that they are pre-trained using raw text directly in a semi-supervised fashion, leaning on an ability to exploit the structured signals in a text to develop representations that enable NLP tasks of interest. It is with this exact philosophy in mind that this work seeks to develop better algorithms for training language representation in the social media domain, exploiting additional unused structural signals that are attached with social media posts.

2.2.2 The Transformer Architecture

The Transformer architecture was first introduced in [33] as an architecture that solely relies on a self-attention operation. This modification allows for the typical time recurrence that appears in sequence-to-sequence models to be eliminated, allowing for much larger-scale and parallelizable training. Though first introduced as an encoder-decoder model for neural machine translation, works like GPT [34] and BERT [20] introduced methods for isolating encoding and decoding portions of the original architecture for language modeling-style tasks.

A general Transformer architecture consists of a token embedding layer, a stack of multiple Transformer blocks, and a prediction head (fitted for whatever is the desired output task). If text generation is to be performed (in the case of machine translation or language generation, for example), a Transformer decoder architecture will be paired with the general architecture, but restricted so that it can only attend to previously decoded tokens in an autoregressive fashion.

A Transformer block consists of a multi-head self-attention operation [33], layer normalization [35], and two feed-forward projections. The order of these operations come in two flavors, depending on where the layer normalization occurs. Layer normalization is used in Transformer models to help promote numerical stability and increase generalization capacity. However, there is a current divide within Transformer literature on where to place these computational units: inside the residual blocks or outside?

The original Transformer paper [33] placed the layer normalization unit outside of the residual unit (Post-LN). However, this has been found to be a lest robust choice [36–38]. Authors in [38] suggest that this is, in part, due to a heavier reliance on the residual connections which can destabilize training and lead to divergence. They suggest an adaptive model initialization scheme (admin) to help set a Post-LN model on a convergent path, and demonstrate it attaining higher performance and faster convergence.

In contrast, placing the layer normalization within residual blocks (Pre-LN) has some beneficial theoretical properties, like establishing a structure that preserves an identity mapping for tasks like machine translation [39]. They also avoid unbalanced gradient issues, applying gradients with similar magnitudes to each of its layers [40]. The cost of this choice is that Pre-LN architectures tend to rely less on their residual connections, limiting their capacity by default [38]. [41] presents Pre-LN's with a Switchable-Transformer (ST) block that can be progressively dropped over time, improving training speeds and demonstrating higher performance in some cases. Theoretical evidence also indicates that one can train a Pre-LN without the warm-up stage and it will converge faster anyways [40].

Training Dynamics

Transformer models are notoriously hard to train [36]. For example, stochastic gradient descent (SGD) fails to train Transformer models that converge [42], requiring the usage of more adaptive optimization procedures like the Adam optimizer [43]. This has lead to a number of studies into why this is so as well as ad hoc engineering fixes that help Transformers to find optima and converge, such as different learning rate schemes[1].

[1] [44] suggests that Adam and other adaptive optimizers suffer from problematically large variance in early stages of training and that learning rate warm-up schedules help to mitigate this effect. They also

Additionally, training Transformers take a *long* time to train. This is, in part, due to the usage of low learning rates with warm-up and cool-down training periods. But when combined with the fact that their learning is unstable and they have a high computational cost, it makes the training of a new Transformer model inaccessible to many practitioners and researchers. This is a significant barrier to research of different pre-training procedures.

2.2.3 Transfer Learning

Transfer learning is *hard*; transfer learning in NLP is no different. There are many issues that arise when transferring a model from one domain to another.

For example, an issue that has plagued neural networks is that of catastrophic forgetting [45–47][2]. This is a phenomenon that occurs when a model un-learns a previous task during learning on a new task, thus negating the very benefit that something like pre-training should yield. This has been observed to impact language models like BERT [49].

Another issue that frequently arises when transferring Transformer models to a new domain or task is high variance due to initial random seed and weight initialization schemes for the final output head [50]. As highlighted by authors in [50], changing data order and weight initialization schemes has significant effect on final models, despite fine-tuning layers only accounting for *0.0006%* of model parameterization.

General Versus Domain-Specific

There is a lack of consensus on how one should deal with deep PTMs for varying or specific domains. There has been significant emphasis on general domain pre-training, but whether one should adapt a model to the specific domain by fully training a model on that domain or fine-tuning from a general model remains an open question.

At the core of the issue is the notion from transfer learning literature [51] which indicates that a successful transfer will occur when the target data is scarce but there is abundant source data that is highly relevant to the target. But what happens when the source data is not relevant to the target? Or when one has enough data in their target domain? Is general domain pre-training still as relevant?

Authors in [52] work in the social media domain and find that a pre-trained RoBERTa model [22] (an extension of BERT with better pre-training hyperparameters) tuned on sixty million tweets and then fine-tuned on the downstream task is more effective than pre-training from the start on tweets.

Contradicting those findings, authors in [53] argue that general domain pre-training is not always helpful, and construct a biomedical pre-trained model from a random initialization. One of their hypotheses for the discrepancy is the tokenization scheme; that is, when building on a generally trained model, the vocabulary is fixed, disallowing a model to specialize to the domain-specific vocabulary.

present RAdam which is a rectified Adam optimizer that has theoretically grounded properties that are supposed to help avoid the variance issue.

[2]Of note in [47] is a brief discussion of how catastrophic forgetting appears in biological learning theories [48]. Instead, it seems other processes reinforce tasks not completed recently in a virtual experience capacity. As such, it may not be possible to directly avoid catastrophic forgetting so much as to come up with algorithms that help systems to better recall and generalize performance.

Chapter 3

A Platform-Agnostic Data Model

This work began by highlighting a number of issues that social media platforms are currently experiencing. In line with goals of this thesis, the creation of tools to help mitigate and understand these problems, this section introduces an abstracted data model based on the properties of several platforms, namely Twitter, Facebook, Reddit, and 4chan.

This data model is then implemented as a Python package, **PyConversations**, that offers functionality like the pre-processing of raw social media data into a common data schema, the chunking of posts into disjoint conversations, and the treatment of conversations as Directed Acyclic Graphs (DAGs) for the study of their structural properties. Furthermore, this package is flexible, easily extensible, and is entirely free and open-source. The package may be found at: https://github.com/hunter-heidenreich/PyConversations.

Within the package are all the utilities needed as well as tutorials and examples scripts for basic analysis of the properties of social media data. As an example of the usability of this package, social media data is aggregated across the four platforms and is analyzed using PyConversations. All executables written as part of this demonstration are also made available in the GitHub repository with the hope that they will be reused and serve as an example for other similar analyses.

3.1 Platform Survey

This work considers four social media platforms: Twitter, Facebook, Reddit, and 4Chan. On the surface, it may appear that these sites are drastically different. While it is true that they have differences, this Chapter explores a unified, platform-agnostic data model for representing discourse on social media in an attempt to provide a universal framing for later training and analysis.

3.1.1 Twitter

Twitter is a micro-blogging, directed, social media platform. Each user maintains a timeline of content that they "tweet" out or content that they have "re-tweeted" (re-shared). Twitter is well known for its character limitations (first, 140 characters and then 280 later on) and its extensive usage of hashtags to tagging tweets, linking them throughout the platform.

On Twitter, conversations are one-to-one (one tweet can only directly respond to one other tweet[1]) however, conversational depth is unlimited (in theory). Twitter is pseudo-anonymous: there is no guarantee that someone's name on the platform is their real identity. This affords conversational participants a certain degree of anonymity, but still allows researchers and observers to track conversational participants to understand conversational threads.

On Twitter, there are two primary mechanisms by which users can publicly interact with one another conversationally: *replies* and *quotes*.

Replies

Replies are, perhaps, the most intuitive way for users to have a public dialog. Any user on Twitter may send out a tweet to which any other user (or even the originator themselves) may *reply*. A reply to a tweet introduces a new tweet that branches off of the source material, initiating a thread. Further replies may be tacked onto that reply, creating an increasing depth to a thread. Furthermore, other users may similarly reply to a source tweet, thus introducing a branching factor to the thread. An example of a tweet and a reply to it can be seen in Figure 3.1.

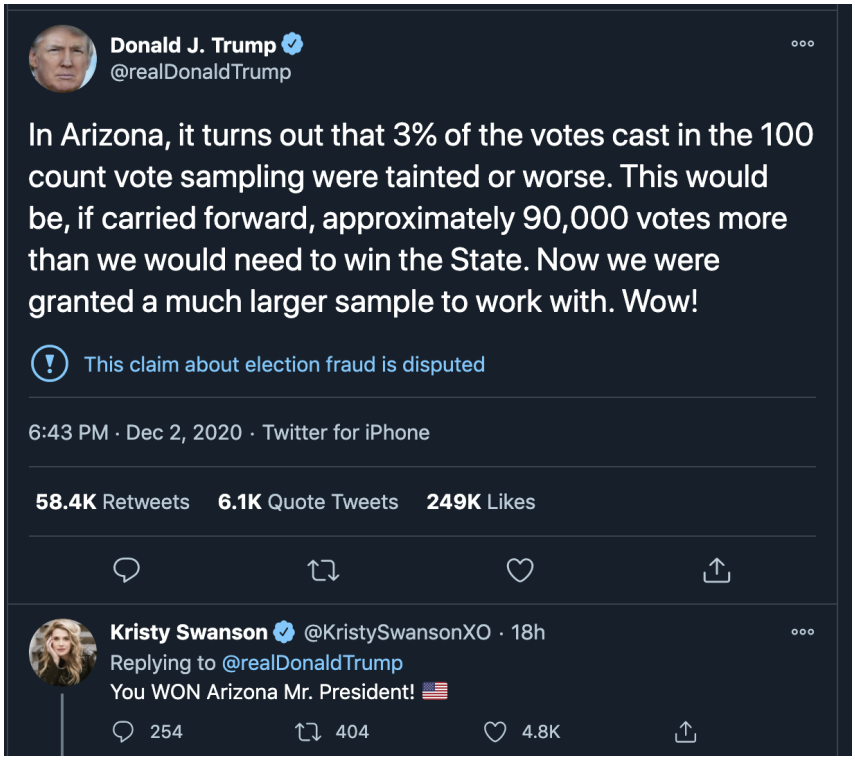

Figure 3.1: An example of a Twitter thread. Users are non-anonymized as they are both verified Twitter users and, as such, assumed to be public figures.

[1]Interestingly enough, though a tweet can only directly reply to one tweet at a time, a tweet can quote tweet a third tweet in a reply, thus technically offering commentary on two tweets at once. This is an extreme case, however. Though physically possible, not many posts feature this in empirical observation herein.

Quotes

In contrast, if a user wishes to comment on a tweet but highlight it to their followers instead of participating in a thread-like dialog, they can perform a *quote* action. A quote is identical to a retweet–in that it projects a source tweet onto their own timeline for their followers to view, except that the quoting user writes a message to provide a new comment or context to their followers. An example of this is shown in Figure 3.2.

Figure 3.2: An example of a quote Tweet action by a verified user on Twitter.

Quote tweets are a particularly interesting mechanism that has been tweaked and adjusted at various points in the platform's lifetime. For example, in an effort to help reduce misinformation and add "conversational friction" during the U.S. 2020 presidential election, Twitter augmented their re-tweet feature prompting users to instead quote tweet with their own comment on a tweet instead of simply broadcasting a tweet to their followers [54].

In Twitter's original blog post, they indicated uncertainty in maintaining this change and ultimately rolled it back after on December 16th, 2020 in a Tweet announcement[2]. In their announcement, they indicated several observations as a result of altering the re-tweet/quote mechanism. Overall, quote/re-tweet actions were reduced by 20%. Furthermore, with this change, 45% of quote tweets were single-word affirmations and 70% featured less than 25 characters of text. While no longer active, this example provides an interesting look at how a platform has adjusted the mechanism of communication in order to curtail and alter human behavior.

Data Privacy

When working with Twitter data, it is standard procedure (and policy) to anonymize user mentions within Tweets. As user mentions are easy to detect with high precision ('@' + the screen name), this work pre-processes all user mentions similarly in an effort to mitigate data privacy concerns. However, where prior works anonymize mentions with a simple `USER` token, this work seeks to apply a similar procedure that retains the ability to track conversational participants. With this in mind, this work performs anonymization of Twitter data at a conversation-level. The first user to be mentioned will be given the token of `USER0` and subsequent users are enumerated and anonymized similarly.

[2] https://twitter.com/TwitterSupport/status/1339350334162890753

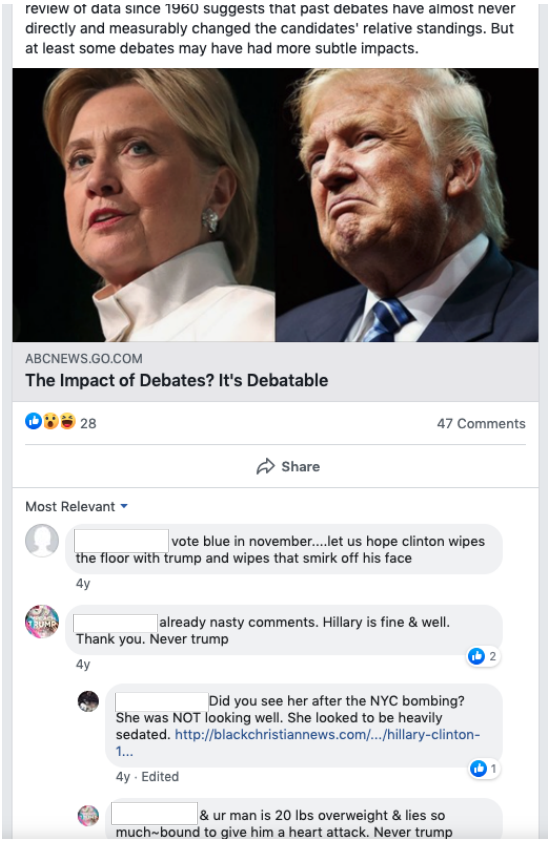

Figure 3.3: An example of a conversation on a public Facebook page.

3.1.2 Facebook

Facebook conversations are one-to-one like Twitter, however, Facebook caps the discussion depth at two levels: comments (top-level comments on a post) and replies (sub-comments, nested under a top-level comment). This has not always been so; initially, Facebook had one level of comments and a system structured around users tagging one another to reply. In an effort to improve on this system, Facebook began to experiment with a reply button and nested comments in late 2012[3]. Additionally, Facebook is unique in that users must use their real names.

Facebook Conversations

Facebook conversations occur in comment sections on private pages, groups, and public pages. This work solely focuses on public pages, particularly pages oriented towards current events and news, though the data model proposed here should still be able to accommodate private page and group data. Given a public page posting, one can comment on that post. Facebook offers one additional level of conversational depth with a reply button. Subsequent replies are concatenated into this nested level as a linear stream, even if a user wishes to reply directly to another nested comment. This is shown in Fig. 3.3.

[3] https://venturebeat.com/2012/11/09/facebook-reply-button/

Data Privacy

Facebook is fairly unique in that its policy is to totally reject anonymity[4]. Despite this, Facebook no longer gives developers access to full user information [55]. Though seemingly well-intentioned, data does not feature name removal prior to restriction. This results in data where true names are contained in the text, but there is no easy way to detect and remove them. Furthermore, it restricts the ability to properly track conversational participants.

3.1.3 Reddit

Reddit is a social media website that functions, largely, as a content aggregation platform. The overall site is split into sub-reddits, individual boards with independent moderation teams typically oriented around a particular topic or interest (e.g., r/pics for pictures). Users submit posts to a board which are then commented on and up- or down-voted by other users. This voting and commenting system allows Reddit to rank the submitted content as well as the comments on individual submissions to help to boost popular and/or relevant content. Sixty days after content is submitted to a board, it and all associated discussion are frozen and archived, though still readily and publicly viewable.

The conversational structure of Reddit is extremely similar to Twitter, with the exception of being organized around boards. Users are psuedo-anonymous, granting a high-degree of anonymity. Replies are one-to-one and can (theoretically) yield infinite depth.

Additionally, Reddit lacks any character limitations on their posts. While seemingly innocuous, this imposes a restriction on the medium that may have an impact on the types of posts written on each site (as well as informs how data must be pre-processed).

Reddit Conversations

Twitter and Facebook are organized by users and pages, but Reddit is not (at least, not primarily). Instead, the site is segmented into independently-run boards based on a mutual interest. Posts are submitted to the subreddit and allow complete branching and infinite depth conversations to branch off of that initial post. This is exemplified in Fig. 3.4.

Data Privacy

As with Twitter, Reddit is pseudo-anonymous. Thus, usernames are identified, enumerated, and replaced with `USERi` tokens.

3.1.4 4Chan

4chan is a very different platform in comparison to the other three. 4chan, like Reddit, is organized in boards centered around a specific topic or set of topics (e.g., technology, paranormal activity, sports). By default, users on 4chan are anonymous (*some* boards attach a unique posting identifier, but not all do). Users have the option to give themselves a more permanent name, but most do not [56].

[4] https://www.facebook.com/help/112146705538576

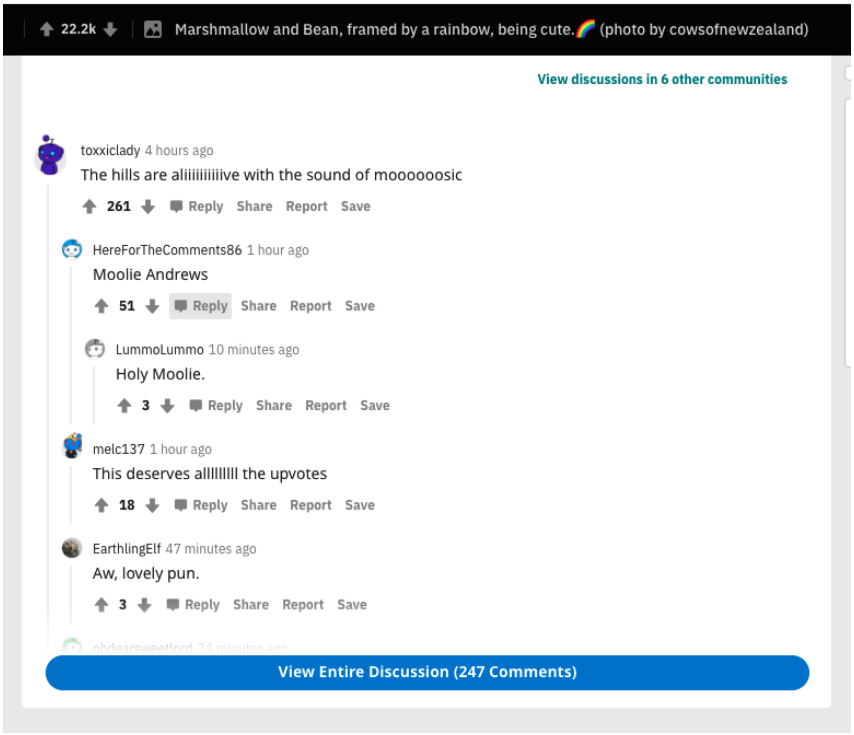

Figure 3.4: An example of a conversation on a subreddit.

Another essential component to how 4chan operates is through ephemerality [57]. When a user replies to a thread, the thread is *bumped* to the top of the board and pushed to the front page. Threads receiving less activity fall from the front page of boards and–if they exceed the last page of the board–are archived. Additionally, boards may enforce a bump limit. Once a thread has exceeded the board's bump limit, its posts can no longer be replied to and eventually they fall off the last page.

A final quirk of 4chan conversations is that users may reply to any other post by referencing its post ID number, preceded by two >>. This has the effect that 4chan sets no limit on the number of posts that any one user may reply to with a single post. This is significantly different from other platforms, which enforce a limit of 1 or 2 out-replies per post. An example of a conversation on 4chan's /x/ board can be seen in Figure 3.5.

Figure 3.5: Example conversation taken from 4chan's Paranormal (/x/) board. Messages are threaded together and linked by their unique post identifiers, all of the form >> $\backslash d+$.

Platform	Depth	Breadth	Anonymity	Identities Detectable?	Boards
Twitter	∞	1	pseudo	yes	no
Reddit	∞	1	pseudo	yes	yes
Facebook	2	1	none	no	no
4Chan	∞	∞	full	no	yes

Table 3.1: A summary of conversational behaviors of the platforms considered in this study.

3.1.5 Platform Abstraction

Although only four platforms have been analyzed, this section formalizes a set of questions asked about the conversational structure of the studied platforms to best draw out the qualities a universal social media data model should encompass. This model is made to be general and with extension in mind–either by adding more platforms that require greater generalizations or by adding more features to the model beyond what has been studied here.

The following are a set of key questions current considered to construct a data model:

- How many levels of conversational depth?
- How many posts can be replied to at once?
- What is the level of anonymization?
- Are conversational participants observable?
- Is the platform organized in boards?

These questions are summarized for the analyzed platforms in Table 3.1

3.2 Data Schema

Abstracting the similarities of posts on social media into a standard format, this work proposes the following structured common features to be extracted into a formal data schema:

- A unique ID
- The text of the post
- The (anonymized) author (if available)
- The creation datetime
- The list of post IDs this post was generated in reply to (or references)
- The platform
- The language
- A collection of project-specific tags and/or task-specific annotations

These features are visualized in Fig. 3.6.

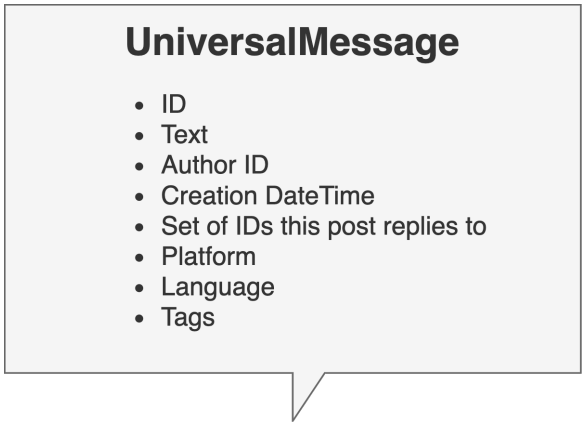

Figure 3.6: The data schema for a "universal" message.

This data schema is designed to be flexible, unifying, and extensible. If more general features are needed, they can be added to this core class. If platform-specific features are needed for analysis, one can easily extended the necessary features with platform-specific sub-classes and still leverage the benefits of this underlying data model. Using this design paradigm, one could imagine the addition of a hashtag index in a Twitter-specific class or the storage of up- and down-votes in a Reddit-specific class as simple extensions to this universal model.

One may wonder why sociometrics are omitted from this model. That is because, for many sociometrics, they are platform dependent. How does a Twitter like map onto a Facebook reaction? The simple answer reached in this work is that they do not, and thus do not currently belong in a universal model (though, they would be great extensions for the platform-specific classes). Instead, this work takes the route of emphasizing the common structural features that are necessary for the alignment of data across diverse social media platforms to help facilitate cross-platform research.

Most important to this work is that this unifying format for social media data sets the stage for self-supervised learning objectives and social media analysis. As an output of this work, this data model and associated conversational analytic tools are made available at: https://github.com/hunter-heidenreich/PyConversations.

3.3 Descriptive Analysis

As an exhibition of the type of analysis one might perform with such a universal data model, this section uses the proposed model to compare several datasets that have been collected from the studied social media platforms. Specifically, this section seeks to highlight how this data model enables the visualization and analysis of the shape of conversations on social media in a fashion that can oddities in both the structural elements of a platform as well as the effects that data collection has had.

For example, Twitter imposes a character limit on Tweets. How does this manifest (or not) within measured statistics? Do characters pack into the tail of the distribution as users attempt to cram their communication into the restrictive limit? Do tokens? Do types? This model has been specifically designed to enable research into questions like

these as a preliminary step towards understanding how structural limitations, rules, and interventions imposed by platforms can shape communication outcomes for social media users.

3.3.1 Datasets

The following datasets are used as samples of their platforms:

Twitter

- **NewsTweet** - A dataset introduced in [58]. This dataset features a collection of tweets that are embedded in news articles. Since these tweets represent a collection of social media posts that were deemed "newsworthy," the conversations branching off of such tweets provide an interesting perspective into the sphere of public discourse on Twitter. The collection of threads branching off of these tweets in this dataset are dubbed **NewsTweetThreads (NTT)**.
- **Coordinated Targeting** - During the Summer of 2020, a work was released that highlighted suspicious activity on a number of high-profile Twitter accounts– particularly, odd behavior in their follower dynamics [59]. In an effort to better characterize the unusual observed behavior, a number of timelines that appeared during these suspicious phenomena were collected. Filtering this collection of the quote tweets originating from the accounts results in a different slice of Twitter conversations that contrasts well with the threaded discusses from NTT. This dataset will be referred to as **Coordinated Targeting Quotes (CTQ)**.

Facebook

- **BuzzFace** - A dataset introduced in [6, 7] which presents the public discourse that was present on the Facebook posts fact-checked by BuzzFeed [60]. This dataset has since been augmented with an auxiliary collection of political/news-oriented Facebook page discourse. The former, BuzzFace collection, will be referred to as **BuzzFace (BF)** and the latter will be referred to as **Outlets (OT)**.

Reddit

- **Change My View** - A dataset introduced in [61] that explores the sub-reddit, r/ChangeMyView. r/ChangeMyView is a particular sub-reddit where users post an opinion they hold and ask for other users to challenge them on it and, as the name suggests, change their view. One aspect of r/ChangeMyView that makes it of interest is the strictly-enforced rules centered around fostering good-faith dialog. The authors of [61] were interested in understanding the dynamics that lead to users violating the sub-reddit's rules by committing an ad hominem attack (an attack against a user in the sub-reddit). The authors made this historical cross-section of sub-reddit data publicly available. This dataset will be referred to as **ChangeMyView (CMV)**.
- **RedditDialog** - In developing an adaptation of GPT-2 [21], the authors of DialoGPT [62] train a language model on 147 M Reddit conversations. Additionally, the authors provide the tools to rebuild a cache of Reddit, which this work uses to

construct a dataset of posts which will be referred to as **Reddit Dialog (RD)**. Specifically, RD consists of threads from 3 sub-reddits: r/news, r/worldnews, and r/politics. The dataset spans these sub-reddits from their creation up to January 2019.

4chan

4Chan has not been widely studied, although there are several publicly available datasets centered around one board in particular: /pol/ [56, 63]. Other datasets exist as well, though not publicly [57, 64, 65].

To attain a bit more board-diversity, this work uses an ad hoc collection of 4chan boards: news (/news/), history (/his/), science (/sci/), technology (/g/), politically incorrect (/pol/), and paranormal (/x/). This dataset is referred to as **4chan (4C)**.

3.3.2 Methods

This work considers the following measurements for its descriptive analysis:

- **Size** – The general size of the social data collected, measured by the number of posts and number of conversations. In this work, conversations are considered as any collection of *at least* 2 reply-linked messages.
- **Temporal Distribution** – The temporal distribution of posting behavior. This is measured both at the day and hour level to characterize both the overall distribution in time as well highlight any regularities in posting behaviors.
- **Language Diversity** – The distribution of detected languages, either as produced by the platform or as detected.
- **Characters per Post** – The size of posts, as measured by the number of characters.
- **Tokens per Post** – The size of posts, as measured by the number of tokens.
- **Types per Post** – The size of posts, as measured by the number of unique tokens (types).
- **Post Innovation Rate** – The innovation rate μ is a fairly regular characteristic associated with human language production [7, 66]. It has also been used to detect social bots [7, 67].
- **Post In-Degree** – The number of replies received by a post.
- **Messages** – The number of messages per conversation.
- **Participants** – The number of detected participants per conversation
- **Languages** – The number of languages being used within a single conversation.
- **Duration** – The length of time of conversations measured in seconds between the first and last collected post.
- **Conversation Innovation Rate** – The innovation rate μ computed at the conversation level. Conversations are likely to be mixtures of language generators, so μ here gives a measure of how "mixed" conversations are [66].

- **Conversation Depth** – The conversational depth as measured as the longest path from a source post to a leaf post.
- **Conversation Width** – The conversation width as measured by the largest number of posts collected at one depth level.

These are not meant to encompass every possible statistic one could measure from the collected data, but rather a useful subset for beginning to characterize the shapes of posts and conversations on social media to begin to reason about how and why they take these forms. Furthermore, it makes use of the previously constructed data model to exhibit the types of analysis one may be able to perform with ease using the **PyConversations** package.

3.3.3 Pre-Processing

In general, posts do not come with language tags. Twitter is the only platform studied that provides a `lang` metadata field. For all other data, language is detected using the GCLD3 package[5].

For text-level analysis, only English and Unknown Language posts are retained. The latter is retained due to the high veracity of short text language classification, especially noisy, short text originating from social media. Furthermore, tokenization is performed in a mass-conserving manner adapted from Text Partitioner [68]. Space is treated as an independent token, reminiscent of more recent approaches to sub-word modeling that specifically mark where spaces occur in a text and attempt to preserve them [69, 70].

When performing conversation-level graph-based analysis, singletons are not considered. This work considers a conversation as a set of at least two, connected posts. Without a connection, there is no dialog. As such, it can be assumed that where conversations are discussed, singletons have been filtered out.

3.3.4 Analysis

Size

The number of posts and conversations are displayed in Table 3.2. Two sizes are displayed; the size of all of the data and the size of the English (en) and Unknown Language (und) subset.

From this table, one can see the varying sizes of these datasets, spanning multiple orders of magnitude. In totality, 308 M posts are considered in this collection and 15.8 M disjoint conversations. As datasets are ordered with respect to the number of collected posts, datasets with many posts but less conversations (OT and BF) are easy to spot.

Another trend is the differences in amount of English data. The Twitter and Facebook datasets contrast with the makeup of Reddit and 4chan conversations, the latter of which are almost entirely written in English or are otherwise unknown.

[5] https://github.com/google/cld3

	All		en & und	
	Posts	Convs.	Posts	Convs.
RD	1.35e+08	4.80e+06	1.30e+08 (96.18%)	4.58e+06 (95.58%)
4C	7.39e+07	2.89e+06	7.17e+07 (97.06%)	2.78e+06 (96.12%)
OT	7.38e+07	7.61e+05	6.03e+07 (81.67%)	6.26e+05 (82.30%)
CTQ	1.78e+07	7.28e+06	1.25e+07 (70.10%)	5.04e+06 (69.25%)
CMV	2.32e+06	3.18e+04	2.29e+06 (98.95%)	3.15e+04 (99.07%)
BF	1.70e+06	2.25e+03	1.36e+06 (80.16%)	1.94e+03 (86.39%)
NTT	1.64e+06	1.38e+04	1.47e+06 (89.55%)	1.28e+04 (92.86%)
Total	3.06e+08	1.58e+07	2.79e+08 (91.27%)	1.31e+07 (82.89%)

Table 3.2: General size for each dataset. For the English (en) and Unknown Language (und) subset, percentage values indicate what percentage of the dataset is retained when filtering to this subset.

Temporal Distribution

An absolute comparison of the concentration of posts over time (per dataset) can be found in 3.7.

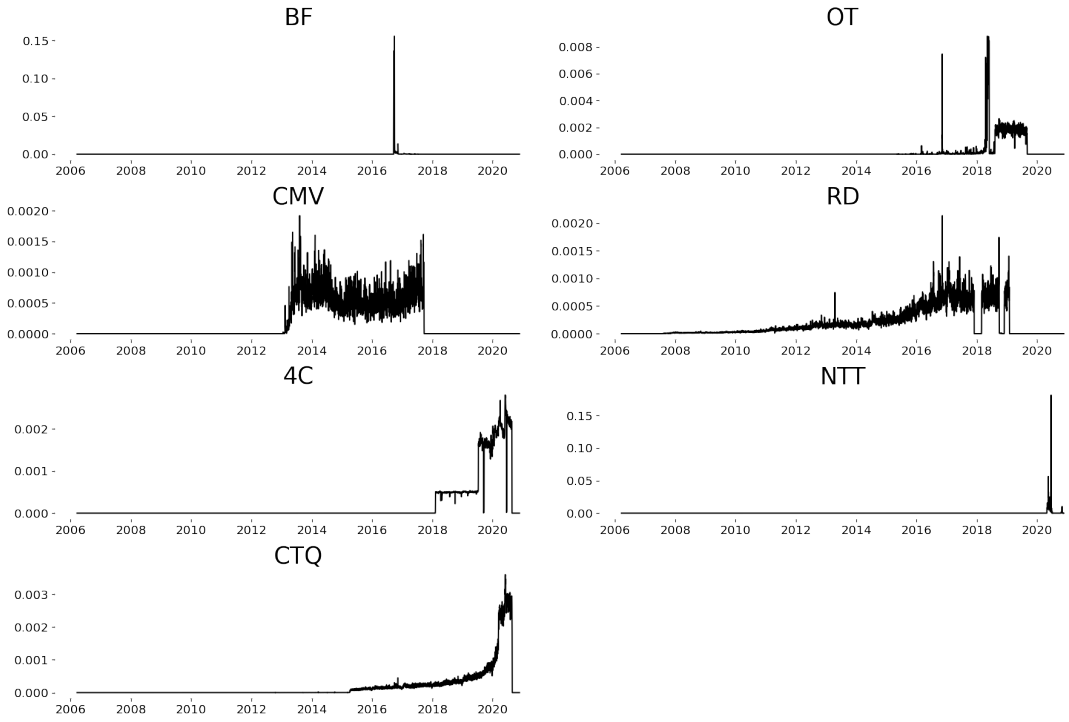

Figure 3.7: Temporal distribution of posts by dataset. Posts are binned by days.

From these small multiples, one can see major differences in datasets that are highly concentrated in time (BF, NTT) and those that are spread well over their collected timespans (CMV, RD, CTQ). That said, many–if not all–offer a rich opportunity for the study of cross-platform alignment, something visually confirmed by this Figure. Additionally, in the details of the sub-figures, collection artifacts can be seen such as skipped collection months (RD) and changes in collection limits and methods (4chan, 2019).

To observe if there are any daily posting time patterns, the normalized counts of posts per hour are displayed in Figure 3.8. Each line within the small multiples corresponds to a particular year of collection, with years with less than 1,000 posts excluded.

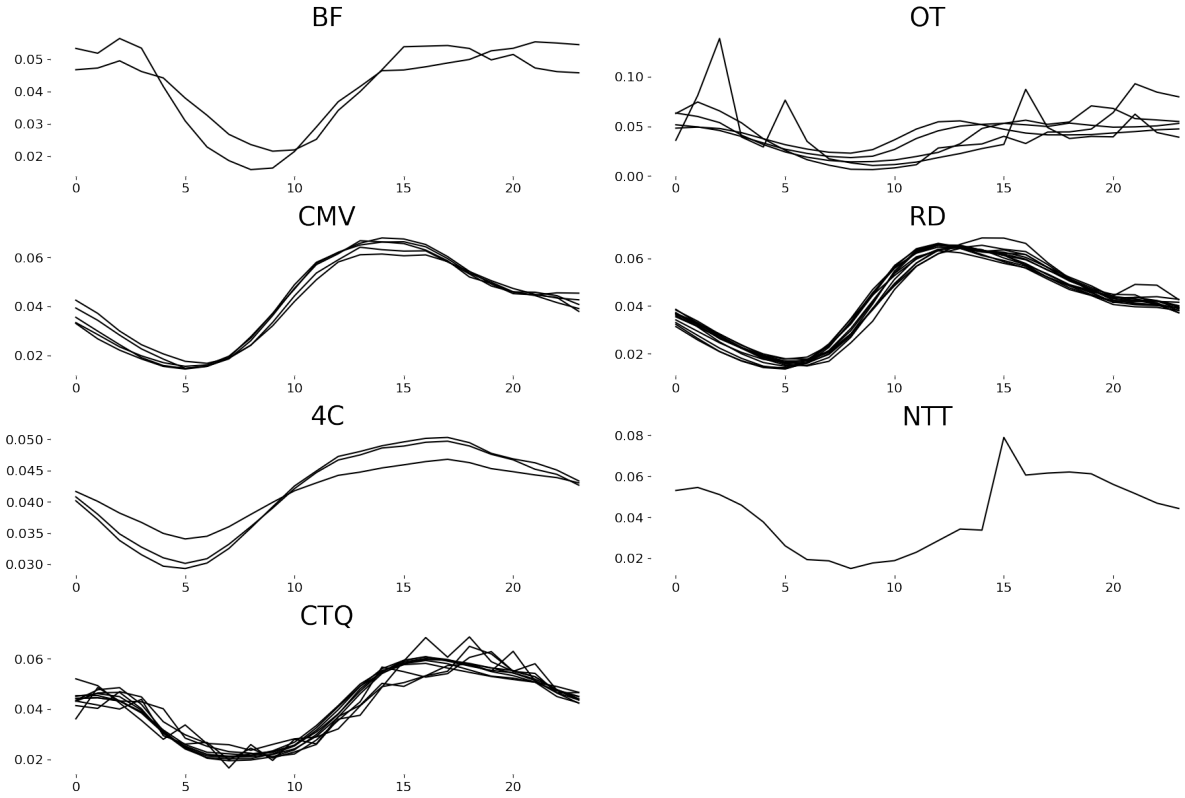

Figure 3.8: Distribution of the posting time behavior within a day, stratified by dataset. Within a single small multiple, each line corresponds to a different year's data.

Almost all the datasets exhibit a similar shape, with a minimum at or around 5 AM. The datsets that fall outside of this characterization include BF, OT, and NTT. Perhaps this is because these datasets are extremely news-centric collections. The news day must occur in order for it to be discussed.

Some of the time graphs are very smooth and consistent (CMV, 4C, RD), others are much noisier (CTQ, BF), and some have odd spiking for some of the collection years (OT, NTT). For the more oddly-shaped datsets, such as OT, further investigation may be required to understand if these spikes are actual anomalies or changes in underlying collection strategy that have lead to greater variance in overall daily shape. For the rest of the datasets, this Figure presents evidence that the daily posting behavior (as collected) has largely remained consistent across years of creation.

Zooming out slightly, the weekly timing regularities of posting behavior are displayed in Figure 3.9. One broad pattern that seems to appear in these Figures is the higher concentration of posts occurring during the weekdays. Perhaps this is a side-effect of several of these datasets focusing on news content, something there is typically more of during the business week. However, CMV does rebuke this explanation for all the data since it doesn't fall within the news-oriented data classification yet appears to exhibit a similar concentration during the weekdays.

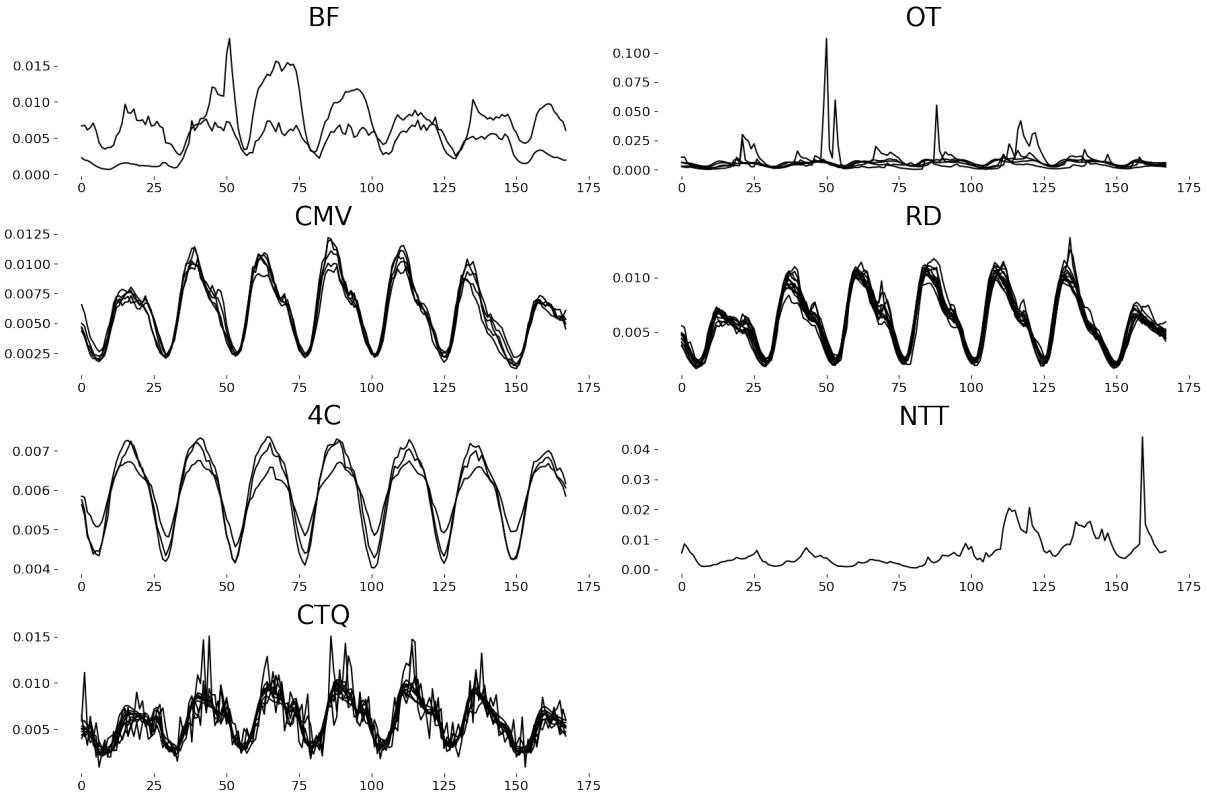

Figure 3.9: Distribution of the posting time behavior within a week, stratified by dataset. Within a single small multiple, each line corresponds to a different year's data. On the x-axis, 0 is Sunday at midnight.

In fact, it appears that yet again, a good portion of the datasets (CMV, RD, 4C, CTQ) have had fairly standard weekly behavior over collection. The general trend holds for these datasets that posting behavior is lower on Saturdays and Sundays. For BF, the more abnormal weekly timing pattern may be partly due to the fact that it is extremely temporally concentrated. One might expect that a year after a set of Facebook posts were posted about particular news events, the posting activity on those posts would be fairly different from how it began when they were first posted. Again, OT stands as an odd contrast to the other datasets.

Language Diversity

Table 3.3 displays the top languages, by post, for each dataset. Languages are listed if at least 1% of posts for any given dataset are written in that language. From this table, one should note how the Twitter (NTT, CTQ) and Facebook (BF, OT) offer much greater language variability, especially compared to the more homogeneous samples from 4chan (4C) and Reddit (CMV, RD). This is further evidence of the high-level observation made when filtering to the English and Unknown Language subset in Table 3.2.

It is worth asking why the Facebook and Twitter samples exhibit more linguistic variability. Is this simply a side-effect of larger, social media platform market share? Is this perhaps an error in the short text language detection module? For data from Facebook, a simplistic explanation could be that the language detection module is classifying names as their linguistic roots; as users frequently tag one another in comment sections, this

	en	und	ja	es	pt	no	fr
RD	90.42	5.77	0.10	0.10	0.11	0.17	0.11
4C	91.23	5.83	0.11	0.07	0.07	0.08	0.09
OT	65.72	15.95	4.45	0.54	0.32	0.60	0.31
CTQ	63.71	6.39	3.48	9.97	4.78	1.10	1.05
CMV	97.27	1.68	0.10	0.03	0.03	0.08	0.02
BF	62.83	17.33	3.84	0.50	0.23	0.68	0.31
NTT	80.47	9.09	0.10	1.04	0.69	0.32	0.35

Table 3.3: Distribution of languages by dataset for languages of posts that make up at least 1% of the entire dataset.

seems reasonable and a side-effect of not being able to properly detect and/or remove true names from the data. For Twitter, it appears that quote tweets (at least within CTQ) are much more linguistically diverse than normal reply tweets (e.g. NTT). This may make sense, particularly if there are individuals taking things and re-phrasing (or translating) them for their own followers, though this could also follow from NTT's focus on the US-centered news ecosystem.

Characters per Post

The amount of characters per post is displayed in Figure 3.10. The the mean and standard deviation of the log-transformed counts are also displayed, alongside the total number of characters in Table 3.4.

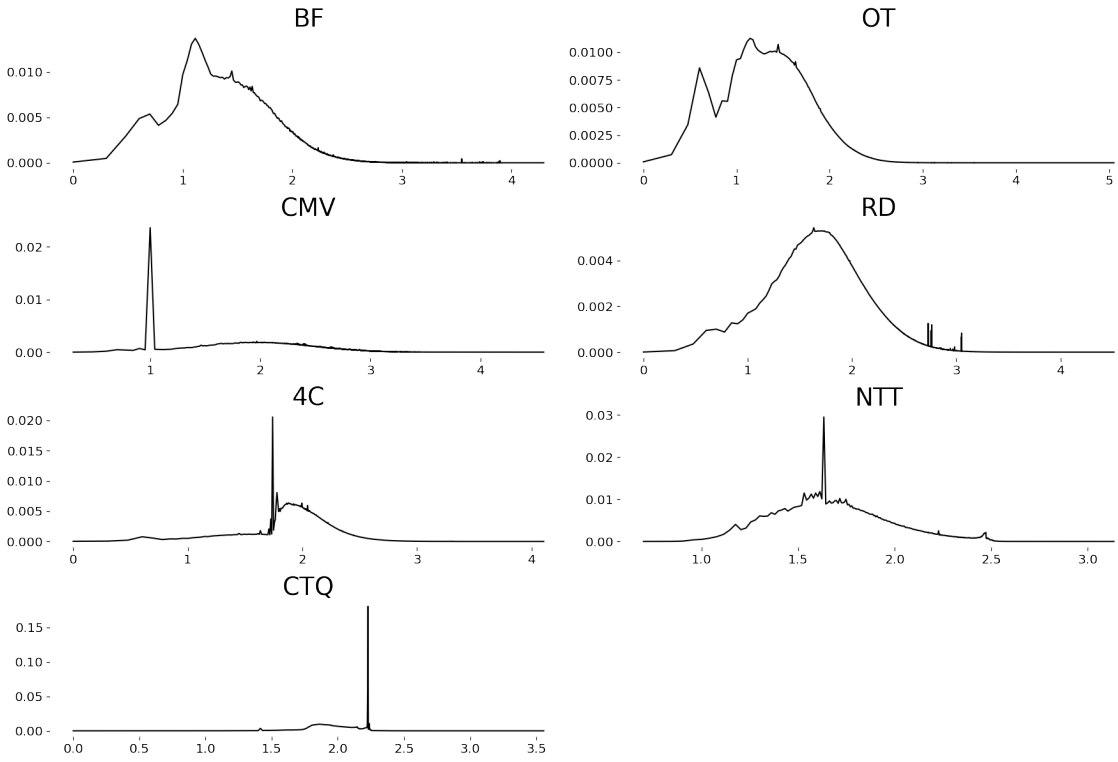

Figure 3.10: Distributions of posts binned by the log of the number of characters within each post.

	Σ	μ	σ
RD	3.02e+10	2.12	0.46
4C	1.47e+10	2.16	0.35
OT	6.42e+09	1.77	0.46
CTQ	1.52e+09	2.05	0.18
CMV	1.33e+09	2.49	0.53
BF	1.92e+08	1.81	0.50
NTT	1.45e+08	1.88	0.32

Table 3.4: Average μ and standard deviation σ of the log-characters per post. Total number of characters Σ are also shown for each dataset.

From these small multiples, some shapes that emerge include point concentrations (over-exaggerated spikes) and bi-modal, log-normal shapes (exhibited in both BF and Outlets). Perhaps the bi-modal shape that appears is a Facebook-specific phenomenon (if so, this, again, could be a direct side-effect of the name tagging phenomena that Facebook users frequently use).

Several of the datasets have spiking point concentrations that are worth examining closer. By far, the dataset with the largest point-mass is CTQ with around 15% of all collected posts having a specific character count (over 10^2 characters). Given the magnitude of mass associated with this point as well as its occurrence at the end of the right-hand side of the distribution, a likely explanation is that this is the maximum character limit associated with posts manifesting as a bunching-up of character-count frequencies as users attempt to cram their message within this artificial limit.

Other datasets have similar point masses, though only consisting of around 2-3% of the total dataset such as CMV at 10^1 characters, 4C at almost 10^2 characters, and NTT around $10^{1.5}$ characters. For the CMV datsets, it seems more than likely that, given the short length, its spike corresponds to some sort of moderation or automated posting behavior. For NTT and 4C, there are less clear explanations, though with 4C it is particularly interesting that the point-mass occurs at the lower limit of the distribution. Is there a typical minimum message size on 4chan?

Several other more subtle phenomena also beg interest. Consider the anomalous spiking in RD, centered around 10^3 characters. Is this some side-effect of a board rule? Perhaps this is related to a URL length?

Tokens per Post

The distributions of post token counts are displayed in Figure 3.11. As before, statistics about the log-transformed token counts are shown in Table 3.5.

This Figure helps to demonstrate how some of the some of the character masses transformed at the token level. For example, CTQ's point concentration appears to still be present in the right-hand side of the distribution. However, it is not as prevalent as it was in the character plots, offering a strong indication that the phenomenon previously observed in the character plots was a result of how character limits effect a platform.

Another example of a character anomaly transferred forward are the small spikes that

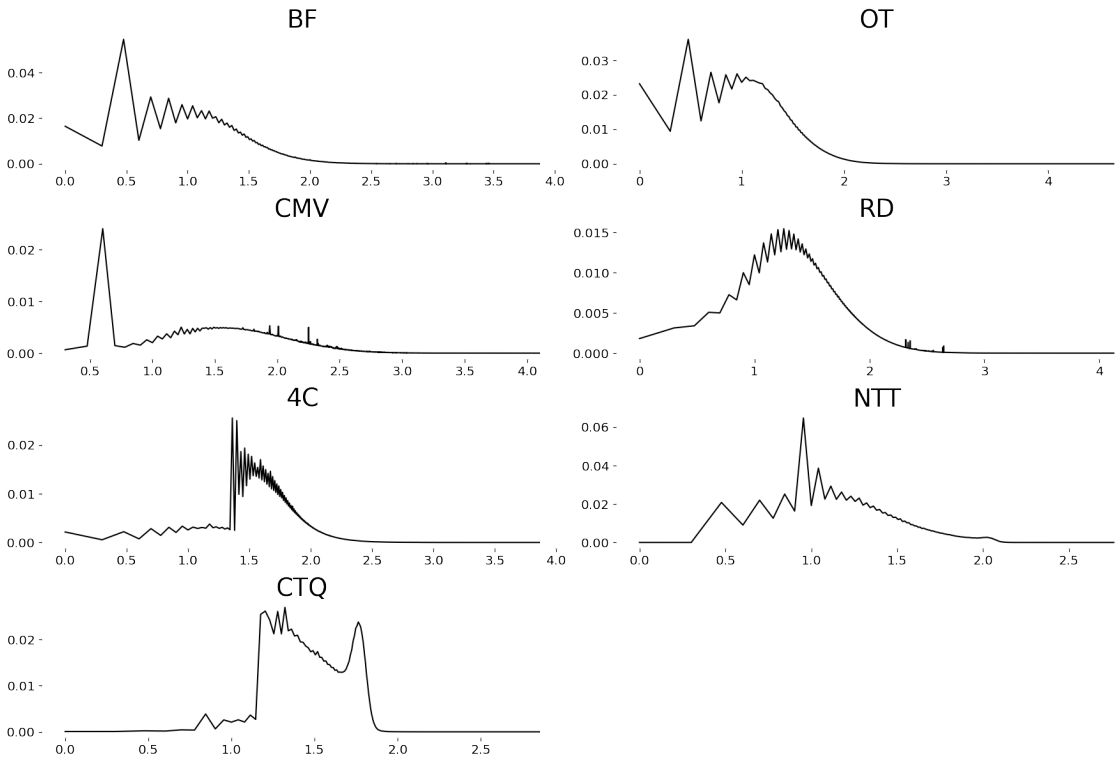

Figure 3.11: Distributions of posts binned by the log of the number of tokens within each post.

were visible in RD's character plots around 10^3 characters. These spikes seem to shift left and shrink, yet this still do appear around 10^2 and 10^3 tokens. This could suggest that this is a result, not of the platform, but rather some agent (or agents) posting in a particular manner that is visible both at the character and token level. Perhaps this is an indication of automated activity.

Types per Post

The distributions of post type counts are displayed in Figure 3.12. As before, statistics about the log-transformed type counts are shown in Table 3.6.

Type distributions appear almost as smoothed and shifted versions of their token-based

	Σ	μ	σ
RD	1.14e+10	1.70	0.46
4C	5.80e+09	1.76	0.35
OT	2.47e+09	1.34	0.49
CMV	5.03e+08	2.07	0.52
CTQ	4.95e+08	1.55	0.21
BF	7.33e+07	1.38	0.53
NTT	4.98e+07	1.37	0.38

Table 3.5: Average μ and standard deviation σ of the log-tokens per post. Total number of tokens Σ are also shown for each dataset.

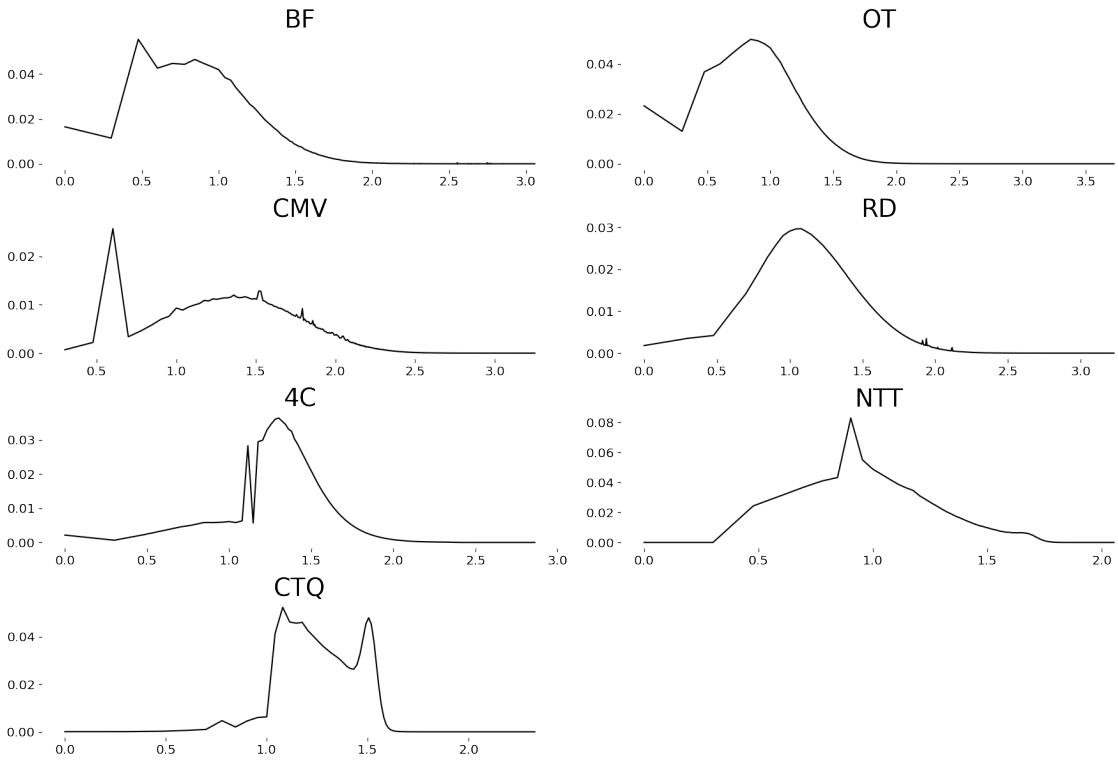

Figure 3.12: Distributions of posts binned by the log of the number of types within each post.

counterparts. Overall, many of the sharp spiking up-and-down (which can be attributed to the discretization of the distribution) are smoothed out almost entirely.

Previous point-masses and anomalous spikes that occurred at the token level have similarly carried over to the type distributions. This further reinforces some of the observations made about RD and CTQ. A new oddity does appear at the type level in 4C, though it is not apparent if this is due to discretization, a pre-processing issues, or some phenomenon unique to 4chan.

	Σ	μ	σ
4C	7.72e+07	1.45	0.28
CTQ	1.34e+07	1.32	0.17
RD	1.11e+07	1.39	0.37
OT	4.85e+06	1.10	0.39
CMV	9.00e+05	1.68	0.41
NTT	6.16e+05	1.15	0.30
BF	3.66e+05	1.13	0.42

Table 3.6: Average μ and standard deviation σ of the log-types per post. Total number of types Σ are also shown for each dataset.

Post Innovation Rate

Regressed post innovation rates are compiled and displayed in Table 3.7 and Figure 3.13. As noted in [7, 66], higher values for innovation rate indicate a higher decay rate and thus a larger degree of text-mixing. For a singular post, this measure should give a sense for the amount of novel text production within a single post.

	Σ	μ	σ
4C	7.13e+07	0.2561	0.4381
CMV	2.29e+06	0.2259	0.3631
BF	1.25e+06	-0.0247	0.3660
OT	5.59e+07	-0.0361	0.3649
NTT	1.44e+06	-0.0380	0.3502
CTQ	1.25e+07	-0.0572	0.2699

Table 3.7: Regressed post innovation rates. The total number of posts Σ that produced valid innovation rates alongside the average μ and standard deviation σ of these values are displayed. Higher innovation rates indicate a higher degree of text mixing occurring or a greater rate of innovation decay [7, 66].

Innovation rate will again be explored at the conversation level, but at the smallest post level, this value could be thought of as how redundant a post is within itself. A high innovation rate (indicating a high novelty decay) would seem to indicate that a post repeats itself or uses many of the same types. As such, it is interesting to observe 4C and CMV appearing to have the highest post innovation rates.

On the other side of the spectrum, NTT and CTQ are two Twitter-based datasets with overall low post innovation rates–in fact, they are negative. This could indicate that, within a single post on Twitter, there is a high degree of novel text production. On a platform like Twitter, where text production is constrained strictly to be within a character limit, this seems fairly intuitive. There simply is not enough character-space to introduce redundancy within a single post.

What would, perhaps, be even more interesting is to consider user's that have self-threads (threads where a user writes multiple tweets in-sequence and in-reply to each other). This would step closer to what past studies have considered from an automaton perspective as a way to detect when a text generating agent is actually an artificial text generator. Using something like innovation rate regression more broadly on a set of posts is also included within PyConversations, allowing for these properties to be evaluated further in future studies.

Reply Out-Degree

Most platforms considered here restrict the number of posts that a given message may reply to. For Facebook and Reddit, this limit is 1 post. For Twitter, one can technically construct a situation where they quote *and* reply to two different messages with one single reply action. For this reason, Twitter's structural reply constraint could be considered as 2. 4chan is the one unique platform studied here where one may reply to an arbitrarily large number of posts (provided they can reference all the posts' associated ID numbers).

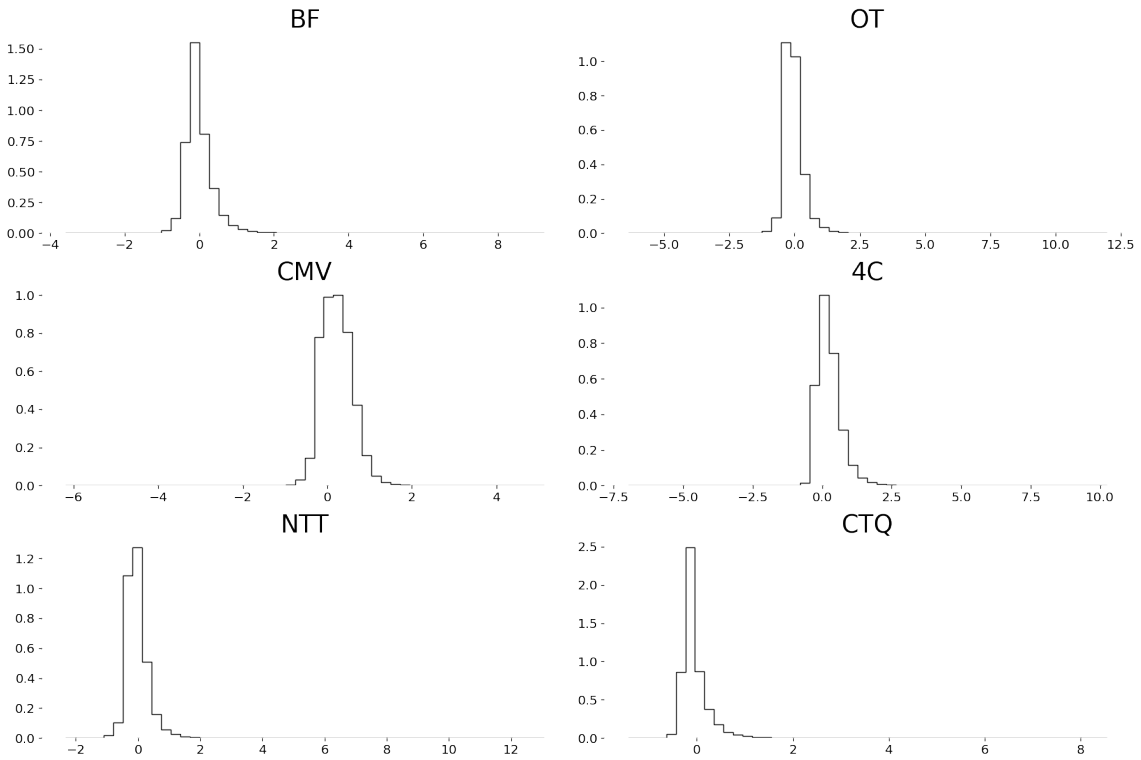

Figure 3.13: Distribution of measured post innovation rates. As in Table 3.7, higher values indicate a higher degree of the text-mixing phenomenon and a higher decay rate of novel text production.

To qualify these assertions, statistics about out-degree of collected posts are stratified by dataset and presented in Table 3.8.

	Σ	μ	σ
4C	7.02e+07	1.70	0.87
CTQ	1.40e+07	1.61	0.49
NTT	1.60e+06	1.01	0.09

Table 3.8: Average μ and standard deviation σ of the number of out-degrees per post. Total out-degree Σ is also shown for each dataset. Posts that aren't replies (source posts) are omitted in computing these stats. Additionally, Facebook and Reddit datasets are omitted as they are either a source or reply to one message.

A couple of observations may be made based on Table 3.8. For one, though it is possible for a Twitter post to reply to more than one post, it is not clear that this is a heavily used feature. For NTT, a dataset of replies, this behavior is all-but-absent. For CTQ, a dataset constructed from quote tweets, the average out degree is higher than 1.5, suggesting that there may be a higher frequency of quote and reply behavior when one is already quoting a tweet to their followers than if one is publicly commenting on a thread.

The 4C dataset is of high-interest when considering reply out-degree. 4chan, providing no structural limitation to the number of posts one may reply to, is an interesting environment to consider what people may do when they have no structural limitations. As expected, there are some collected 4chan posts that reply to many other posts (often

sequencing multiple reply IDs). But unexpectedly, the average reply out-degree, even on 4chan, is less than 2. This may suggest that, for most collected communication, authors do not feel the need to reply to multiple people at once. Perhaps it is too cognitively burdensome.

Either way, the collected 4chan data still features the highest out-degree of any other studied dataset, so it does seem that there is a space beyond current social media platform limitations that users would use, if it were present. That said, from these observations, it does not appear that a multi-reply feature would be used by everyone or even all that frequently.

Post In-Degree

Post in-degree are measured by the number of replies to a post, per the data collection. The distributions of the log-in-degree[6] are displayed in Figure 3.14. Additionally, the maximum in-degree and the average and standard deviation of the log of the in-degree are shown for each dataset in Table 3.9.

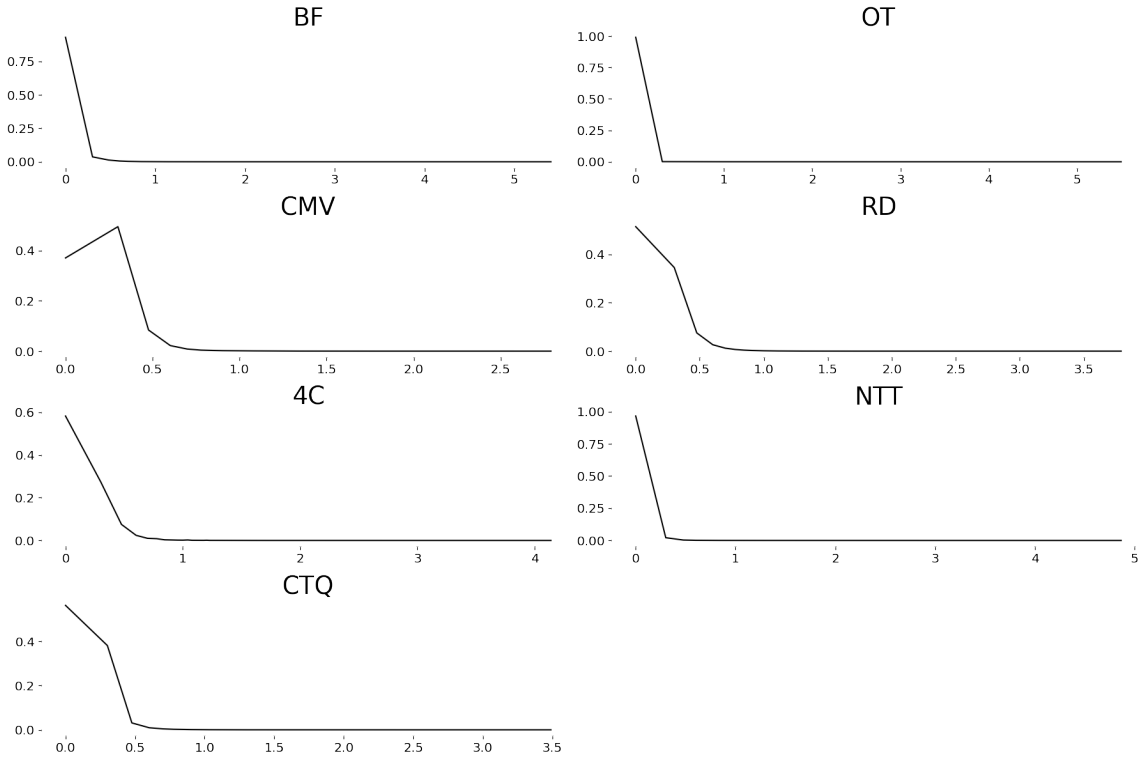

Figure 3.14: Distribution of posts binned by the positively-shifted log-transformation of the post's in-degree.

Consider the two Facebook datasets, BF and OT. By far, these two datasets feature the largest maximum in-degree for a single post, with a magnitude of 10^5 posts replying to a single post. Despite this fact, these two datsets also exhibit the lowest average post in-degrees. Why is that? It seems that there may be a structural influence creating this effect.

[6] shifted +1

	max	μ	σ
OT	316954	0.01	0.15
BF	256689	0.04	0.16
NTT	73304	0.02	0.14
4C	13762	0.18	0.28
RD	6191	0.19	0.23
CTQ	3093	0.15	0.19
CMV	614	0.23	0.21

Table 3.9: For each dataset, the maximum post in-degree is displayed. The log-transformed post in-degree data is also used to calculate an average μ and standard deviation σ.

Facebook only allows two levels of reply nesting. All top level comments reply to the source post. One may nest an additional reply to a top level reply, but no other most may directly reference that inner-nested reply. That does not mean people do not reply to one another *within* this nesting structure, but rather that any users that do so are attempting to do so *against* the structural limitations of the platform. This has the observed effect of an extreme lowering of the possible average in-degree to a post. No nested reply may have an in-degree that is not zero, without some additional conversational disentanglement on top of the already collected data.

Across the rest of the data, there exists a pattern where most posts receive no replies. This is hardly surprising. Human attention is limited and thus some sort of attention filtration should be expected to be observed. Perhaps a more thorough investigation into thread reply count ranks is worth future consideration [71].

Messages per Conversation

Now focusing more closely on conversation-level statistics, consider the number of messages within each collected conversation. The log-transformed average and standard deviation are shown in Table 3.10. The distribution of messages per conversation are displayed in Figure 3.15.

	μ	σ
BF	2.19	0.74
CMV	1.62	0.44
OT	1.31	0.72
NTT	1.24	0.65
4C	0.99	0.52
RD	0.78	0.51
CTQ	0.35	0.14

Table 3.10: Average and standard deviation of the log of the number of messages per conversation.

CTQ is highly biased towards two-turn conversations. As this is a dataset of quote tweets, this seems logical and expected. In order to get more than two messages in a conversation

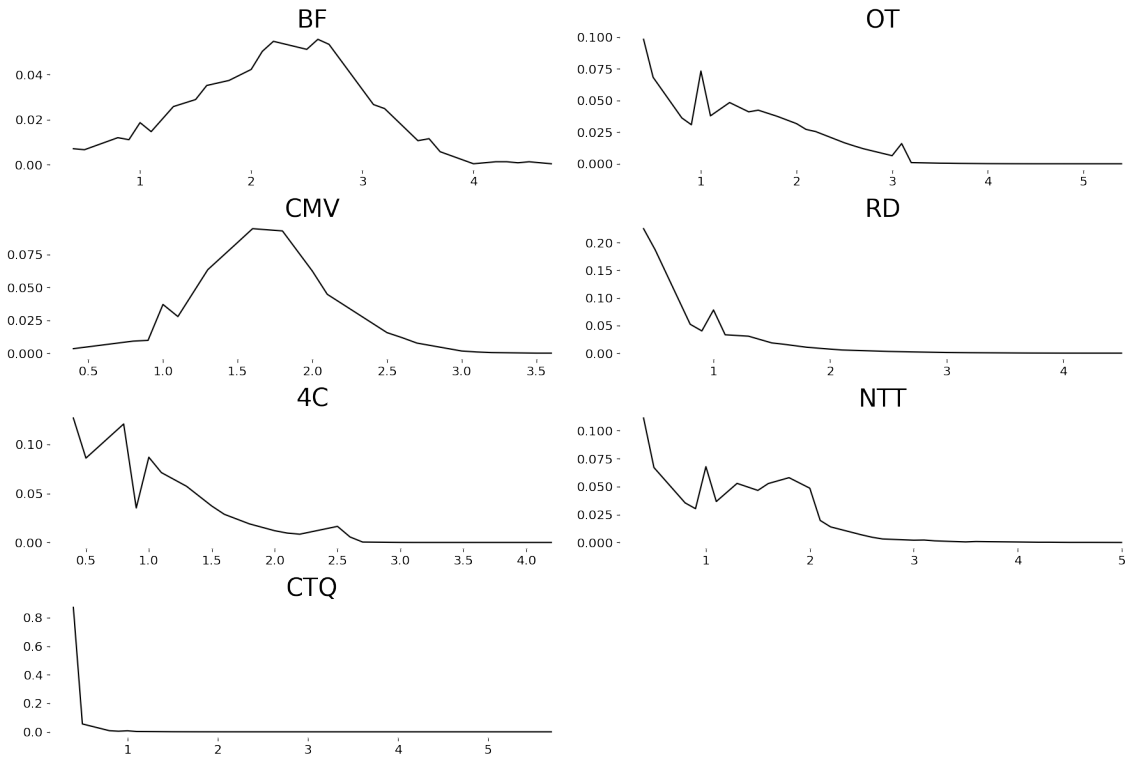

Figure 3.15: Distribution of the log of the number of messages per collected conversations, stratified by dataset.

within this dataset, either multiple Tweets must be quoting the same Tweet or a user was chaining quote Tweets of their own quote Tweets recursively. However, it appears that, based on this plot, two-turn single quote actions are what this dataset primarily contains.

Overall, there is a general trend of sharp attenuation in these distributions. Most conversations receive some activity (per the sampling threshold of two connected posts), but few rise above to garner massive amounts of attention. This is a phenomenon that seems to manifest across datasets, barring some of the more subtle features of each plot.

Why does CMV exhibit a different shape than all other datasets? CMV exhibits a distribution that seems to indicate a "sweet spot" of sorts around $10^1 - 10^{1.5}$ messages. Does this indicate a practice within this sub-reddit that each conversation tends to receive around 10 or so replies? If so, is this an extension of the "good-faith" dialog practice?

The 4C dataset exhibits an odd spiking behavior. It could suggest something artificial is its cause. This could have to do with how conversations are being sampled and burying rate of threads on 4chan, but, nevertheless, is surprising. This phenomenon should be investigated against more complete and alternate 4chan-based datasets to examine if this is a genuine behavior within 4chan or if this is truly a sampling effect.

Finally, it is worth highlighting the odd spike on the OT dataset around 10^3 messages within a conversation. OT has other oddities in previous plots (particularly time-related plots), but this appears to be a different detail altogether. Again, it may be something that is explained through the collection process, but unlike with the 4chan data, this spike is not something that appears to be repeated making it all the more interesting.

	μ	σ
BF	2.00	0.73
CMV	1.23	0.36
NTT	1.20	0.66
RD	0.67	0.46
CTQ	0.33	0.14
4C	0.06	0.15

Table 3.11: The average μ and standard deviation σs of the log of the number of participants per conversation. OT is omitted due to no user information collected.

Participants per Conversation

Next, consider the number of (known) participants per collected conversation. In all likelihood, values based on these measurements are a lower-bound[7] on the actual number of participants due to factors like anonymity and psuedo-anonymity.

Figure 3.16 displays the distribution of the number of users per conversation. OT is omitted from this statistic as user information is absent from this collection. For averages and standard deviations of the log of the number of participants, see Table 3.11.

For several of these datasets, there appears to be a consistent increase in expected participants until around 5-10 users. This is something observed within RD, NTT, CTQ, and CMV (although for CMV, the peak appears to be shifted closer towards 10 users). After this critical point on each of these plots, the number of users appears to attenuate and drop off. This seems to suggest that many of these conversations are exchanges between a limited few in which it may be productive to understand what their conversational exchanges were like. What were their intentions in participating? What were the conversational outcomes? Were opinions exchanged? Were they shifted?

For 4chan, the expectation is that users will be anonymous. Users, by default, will post under the moniker `Anonymous` or `anon` for short. That said, not all users are anonymous in the collected data, as exhibited by the 15% or so of conversations without just a single (anonymous) user–or what appears to be a single user since one cannot disambiguate between two Anonymous users.

To investigate the amount of anonymization on 4chan in its entirety, the amount of Anonymous users are considered and displayed in Table 3.12. From this table, one should note that over 97% of the posts collected where written by `Anonymous`. However, this does differ by board. It is interesting to see that the least amount of anonymized data comes from the /x/ board, a board centered around the discussion of paranormal phenomena. This invites a myriad of questions that ultimately reduce down to a simple "why?" Further study into 4chan and its boards will need to occur before that can begin to be explained, however.

[7]Technically, there could be a scenario in which these counts are actually not a lower-bound. If a single individual (or entity) has made multiple accounts to give the impression that there are more users contributing to the conversation (either through automated or manual means), this could result in a collection that incorrectly over-estimates this lower-bound. However, that would be an extreme adversarial behavior for someone (or someones) to spin up many, fake artificial accounts, so this possibility is neglected here.

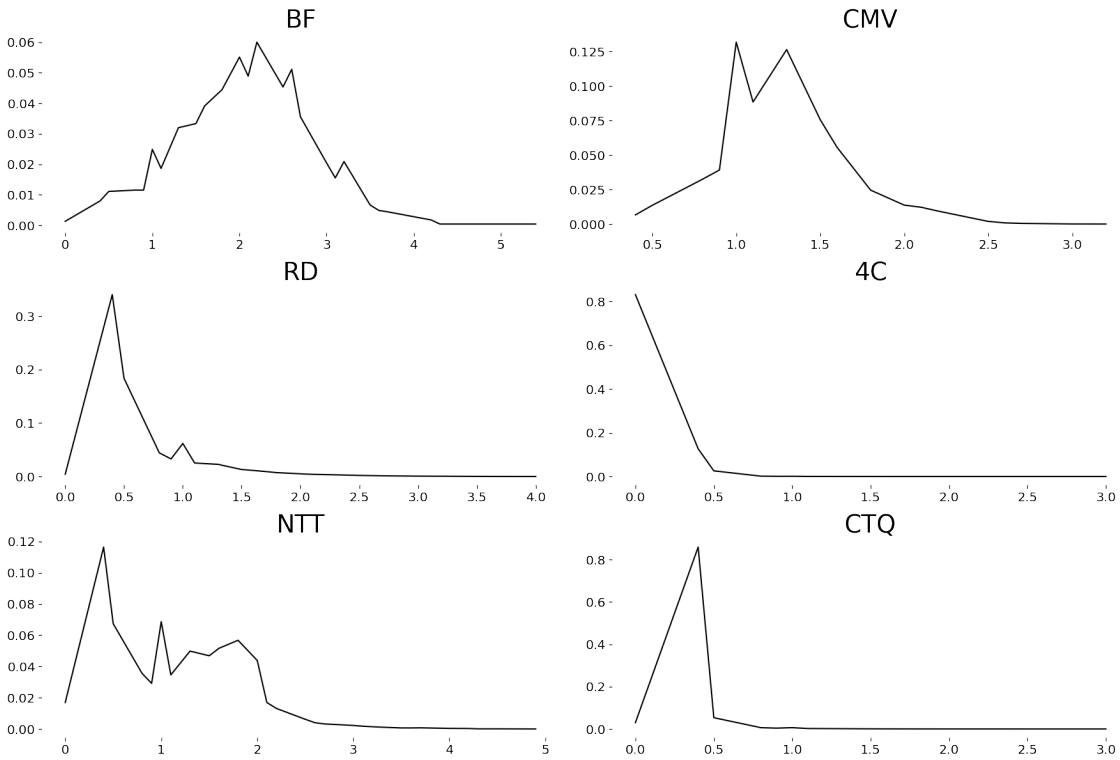

Figure 3.16: Distribution of number of users per collected conversations, stratified by dataset. OT is omitted due to the lack of user information.

Language Diversity per Conversation

For the collected conversations, how diverse are they with respect to the number of different languages that posts are written in? In other words, what is the distribution of the number of languages within a single conversation for each of these datasets?

Figure 3.17 displays the distribution of the log-transformed counts. Table 3.13 shows the average and standard deviation of these values.

Most conversations exhibit a low variety of languages within a single conversation. This makes sense; in order to communicate with one another, users need to post in a common language. The more languages there are discussion in, the less likely it is that participants are to understand one another. Ignoring the Facebook datasets for a moment, Table 3.13

	Anonymous
his	99.01
g	98.78
sci	98.22
pol	98.05
news	96.92
x	91.74
Aggregate	97.81

Table 3.12: Percentage of posts written by Anonymous users on 4chan.

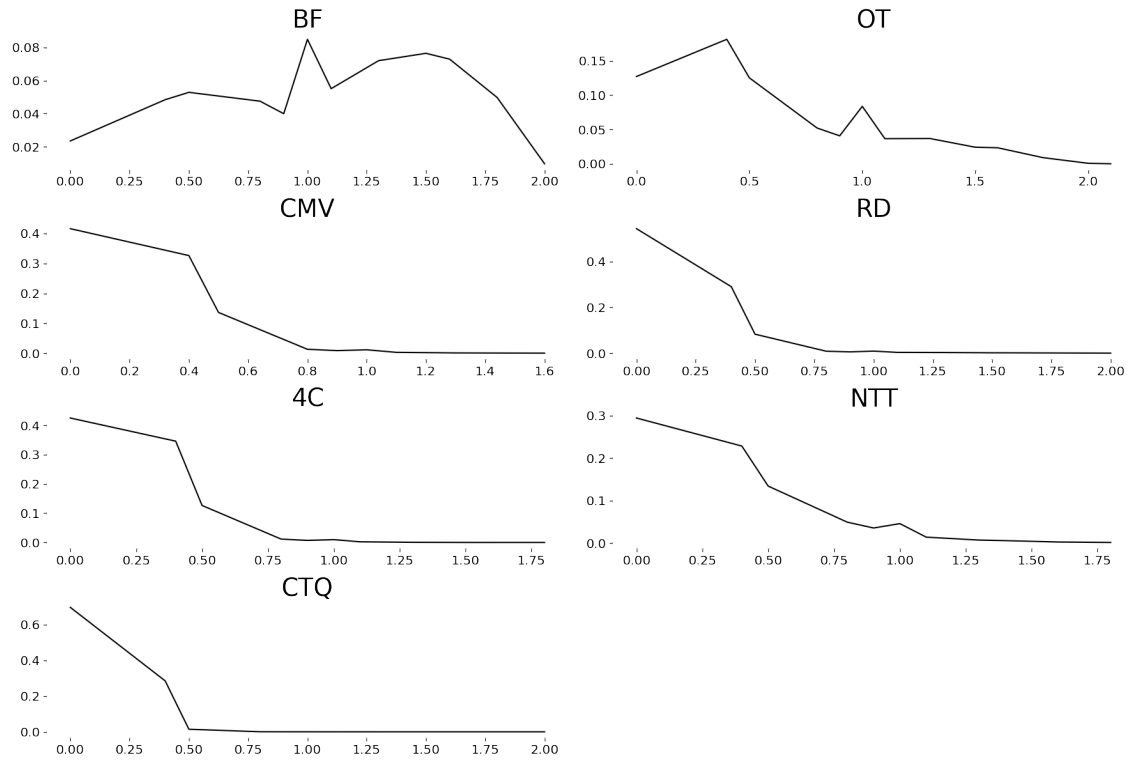

Figure 3.17: Distribution of conversations binned by the log of the number of detected languages within them.

suggests that, on average, conversations exhibit 2-3 languages.

In fact, for all non-Facebook datasets, the single most likely event is that a conversation is entirely within a single language–even controlling for the fact that posts that fail detection (i.e., are classified as Unknown) will increase the language count to 2 if a single other post is correctly detected.

As previously discussed when analyzing language diversity at the post and dataset levels, Facebook appears to have other phenomena effecting the way that languages are being detected and the amount that appear within a single conversations. This seems to offer even more evidence that name tagging (or some otherwise unknown effect) is dramatically altering the way that classification is occurring and counting languages. Although beyond

	μ	σ
BF	1.11	0.46
OT	0.68	0.46
NTT	0.41	0.35
4C	0.24	0.24
CMV	0.25	0.25
RD	0.19	0.26
CTQ	0.10	0.15

Table 3.13: Average μ and standard deviation σ of the log-transformed number of languages detected within a single conversation.

the scope of this work, this further highlights the need for some level of pre-processing to otherwise filter, detect, and remove simple name tag events.

Conversation Duration

What is the lifetime of a conversation on a given social media platform? How does the fast-paced nature of 4chan–where posts that fall off of the last page on a given board are deleted–affect the overall lifespan of a conversation? Or, for Reddit, if posts are locked after 60 days? What shapes form when no restriction is placed by the platform, such as with Facebook and Twitter?

	μ	σ
BF	5.44	0.74
CMV	5.34	0.65
OT	4.91	0.34
NTT	4.76	1.07
RD	3.96	0.94
CTQ	3.90	1.13
4C	3.53	0.77

Table 3.14: The average μ and standard deviation σ of the duration of collected conversations in log-seconds.

Table 3.14 displays the average and standard deviation of the duration of collected conversations as measured in log-seconds. From this Table, one should note that 4chan has the shortest length conversations while Facebook appears to have the longest.

The brevity of the collected 4chan conversations is interesting. 4chan is a platform that does not directly restrict the length of time of a conversation, but has a variety of other structural rules that help to keep each board fresh. For one, bump limits are enforced by some boards which restricts the number of times a conversation may be replied to and thus brought to the top of the board. Additionally, if a post is receiving little attention, it may simply fall of the board's last page and be archived–another mechanism that could encourage shorter conversations on the platform.

A final consideration for this 4chan data is whether or not it encompasses the entirety of conversations or if posts are missed by this collection. If posts are missed due to scraping utilities, critical linking posts could be omitted from the collection, resulting in underestimates in the durations of 4chan conversations. More 4chan datasets should be compared against before asserting that this is true of all 4chan data.

Another observation that can be drawn from Table 3.14 is that the 60-day archive limit that Reddit enforces does not appear to exhibit much of an effect on conversation duration. Though not entirely unexpected–6 months is around 5×10^6 seconds–it is a sanity check that the empirically measured values confirm this.

Figure 3.18 displays the distribution over log-second durations of the collected conversations. Though not much can be abstracted across different dataset, it is interesting to see each dataset exhibit its own temporal time signature that seems to be indicative of how and what was collected. More news oriented datasets like OT, NTT, and BF all seem to

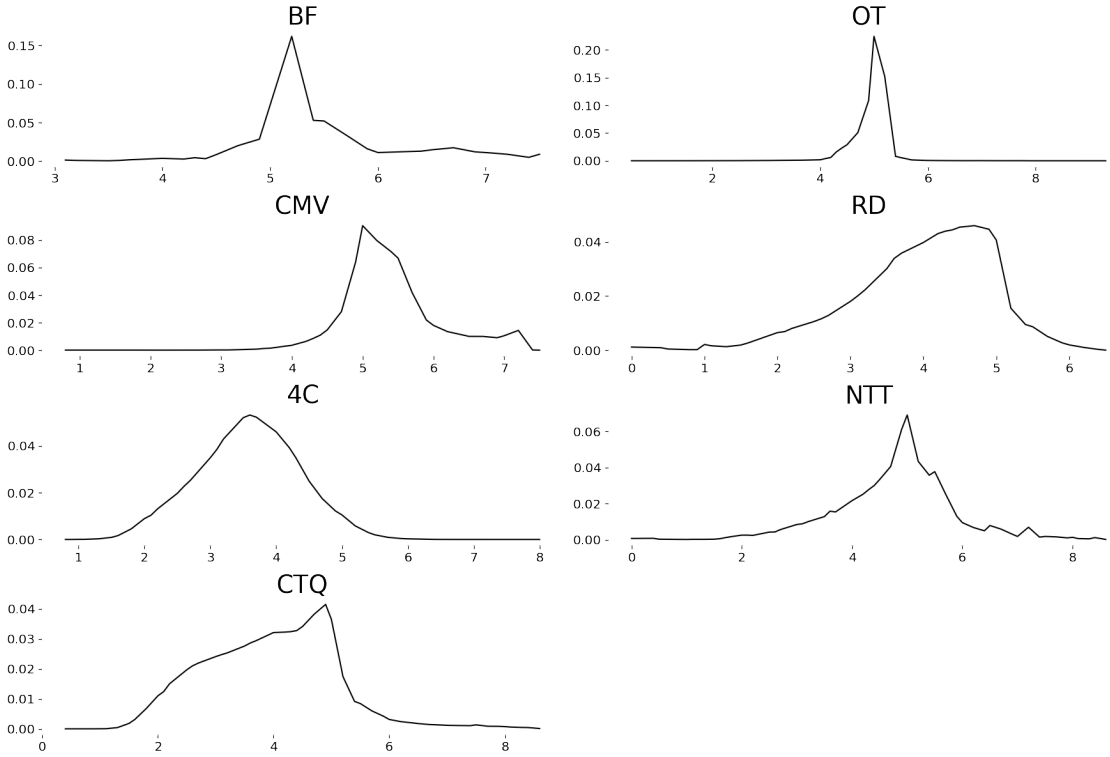

Figure 3.18: Distribution of the temporal length of conversations (in log-seconds), stratified by dataset.

exhibit spiking phenomena that may indicate attenuation of human attention with the news cycle. More socially-oriented datasets such as RD, CMV, 4C, and CTQ, seem to have smoother shapes that are more log-normal.

Conversation Innovation Rate

Regressed conversation innovation rates are compiled and displayed in Table 3.15 and Figure 3.19. As noted in [7, 66], higher values for innovation rate indicate a higher decay rate and thus a larger degree of text-mixing.

From the Figure and Table, several observations may be characterized. For example, CMV experiences the highest degree of average conversational text mixing. This could be indicative of the idea of that conversations start and end around the same topics, thus the innovation rate decays rapidly (e.g., once the topic is established, participants stay concentrated around it). If this is truly what is occurring, this may be another empirical by-product of the board's strict adherence to good-faith discussion and other rules on r/ChangeMyView.

At the other extreme, CTQ has a very low average innovation rate. This should indicate that there is a lower level of mixing occurring within this dataset. With a dataset like CTQ consisting mostly of quote-tweet pairs, it does make sense that this value is lower. For one, there is likely less text to consider in smaller conversations. Additionally, one could imagine that if the intent of a quote-tweeter is to re-phrase or comment on a Tweet to their followers, they may have a higher degree of novel word selection to bridge the gap between the original Tweet and the typical discourse a user cultivates on their own

	Σ	μ	σ
CMV	3.18e+04	0.9330	0.3732
BF	2.25e+03	0.6447	0.4165
4C	2.89e+06	0.4444	0.3225
RD	4.79e+06	0.3599	0.3751
NTT	1.37e+04	0.3560	0.2864
OT	7.60e+05	0.3541	0.2939
CTQ	5.51e+06	0.2349	0.3633

Table 3.15: Regressed conversational innovation rates. The total number of conversations Σ that produced valid innovation rates alongside the average μ and standard deviation σ of these values are displayed. Higher innovation rates indicate a higher degree of text mixing occurring [7, 66].

timeline between themselves and their followers.

Another interesting observation to highlight is how close the average innovation rates are for conversation across several datasets like RD, NTT, and OT. Visually inspecting the distributions, it appears that RD has a higher degree of a left-ward skew towards low innovation rates. OT and NTT look even more similar. As it currently stands, this work offers no explanation as to why these datasets may be close in innovation rate. Further work will need to qualify the degree to which this is a spurious phenomenon or indicative of underlying conversational dynamics that are present across social media platforms (Reddit, Facebook, and Twitter).

Conversation Depth

Conversations on social media can be intuitively represented as directed acyclic graphs (DAGs). When a conversation has a single source (such as a submission post on a subreddit), conversations are simply trees. Structuring conversations as trees (and DAGs) begs the question of "how deep do conversations get?" For platforms like Facebook where conversational depth is limited to two levels, this has a very simple answer. However, for other platforms that feature no such limit, empirical observation should provide evidence as to whether there exists any platform-specific oddities.

Figure 3.20 displays the conversation distribution of tree depths, alongside their numerical statistics in Table 3.16.

As anticipated, sharp cut-offs in depth of conversation are observed. Facebook data lacks true depth due to structural limitations. OT, without any nested comment collection, is totally barren of depth.

4C, somewhat surprisingly, offers a low measured depth value. This, in combination with other previous observations, could be further evidence that critical linking posts are not present to build out full conversational trees. This could stack with any other additional psychological effects like the cognitive burden of dealing with many other anonymous conversation participants to produce shallower trees.

Many of the other datasets, except CMV, exhibit almost identical behavior. There is a higher mass of posts towards the shallow tree regions that sharply declines as larger

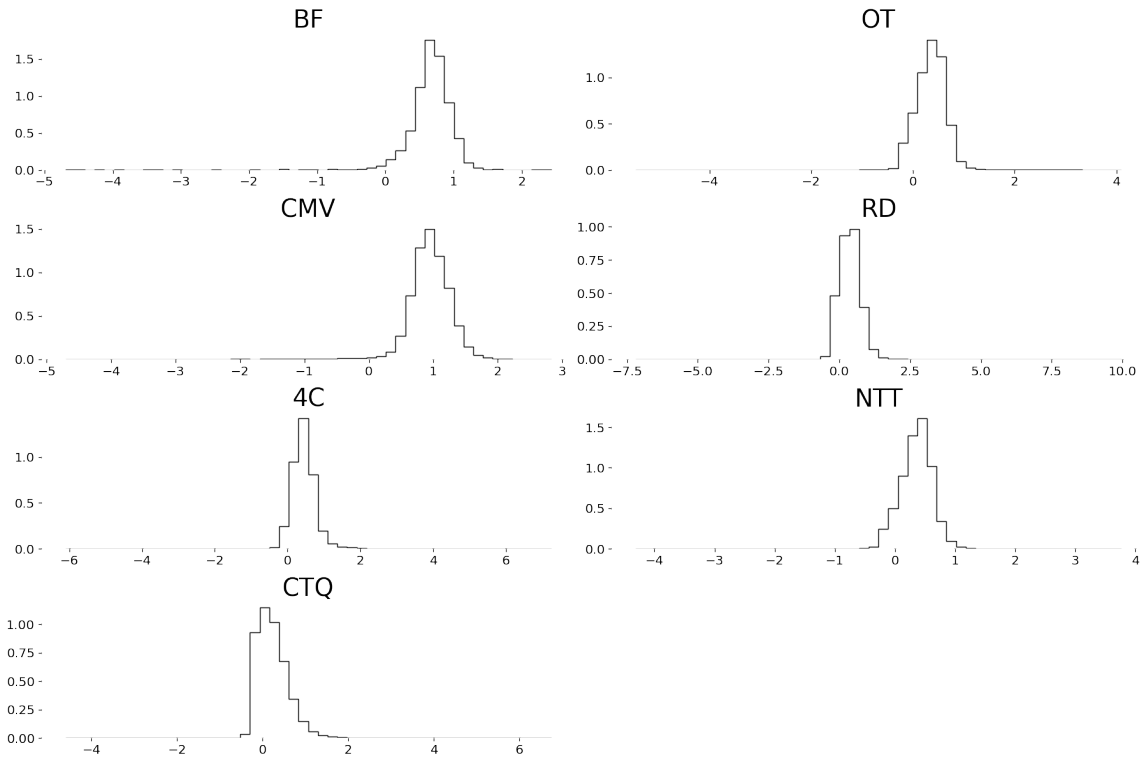

Figure 3.19: Distribution of measured conversational innovation rates. As in Table 3.15, higher values indicate a higher degree of the text-mixing phenomenon and a higher decay rate of novel text production.

and deeper conversations are considered. This occurs without any structural enforcement either on some of these platforms, perhaps indicating a certain threshold that few people engage beyond to continue a discussion deeper and deeper. In fact, this may make the sub-selection of conversations with longer depths even more interesting for further study.

Conversation Width

Another tree-measure that may provide information about the shape of conversations is the width of the conversation. If depth level is determined as the distance from a post to the (or a) source, the width of a depth level is simply the number of posts that fall at that depth level. Similarly, one could say the width of the tree is the maximum number

	μ	σ
CMV	0.91	0.30
RD	0.36	0.36
BF	0.28	0.08
NTT	0.18	0.25
4C	0.02	0.10
CTQ	0.02	0.08
OT	0.00	0.00

Table 3.16: Average μ and standard deviation σ of the log-tree depth of conversations.

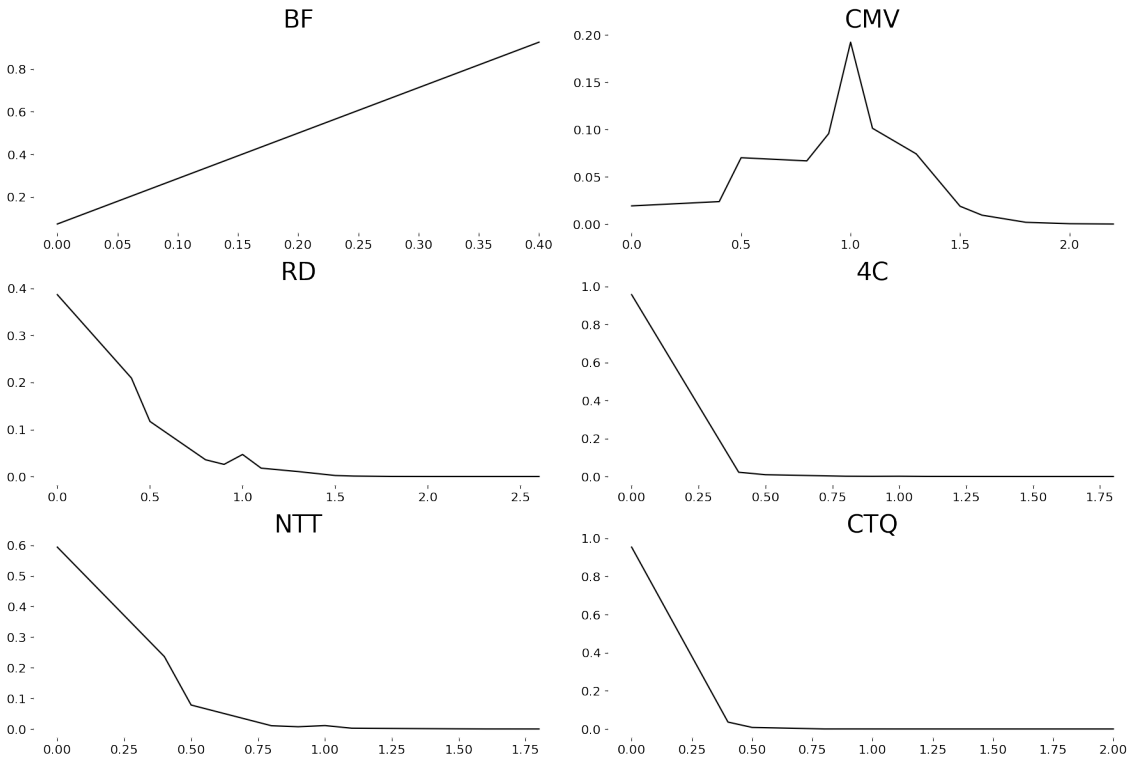

Figure 3.20: Distribution of conversation by the log of their tree depths.

of posts concentrated at a given depth level.

Figure 3.21 displays the distribution of conversation tree widths. Table 3.17 exhibits the average and standard deviation of the log-transformed tree widths.

	μ	σ
BF	2.03	0.75
OT	1.24	0.79
NTT	1.11	0.73
CMV	1.04	0.33
4C	0.88	0.61
RD	0.40	0.45
CTQ	0.06	0.18

Table 3.17: Average μ and standard deviation σ of the log-tree width of conversations in the studied data.

From these characterizations of the data, it becomes even more apparent that Facebook conversations are short and fat. The width levels found on Facebook exceed, greatly, the widths observed on other platforms. This seems to suggest that Facebook, in limiting the depth of conversations, may have created a situation where people desperately attempt to pack their posts into the limited depth levels and ultimately end up expanding the width of conversational trees.

A familiar spiking appears in the 4C dataset. This could be a side-effect of the discretization of the data (log-transforming integer counts). However, the fact that it appeared in

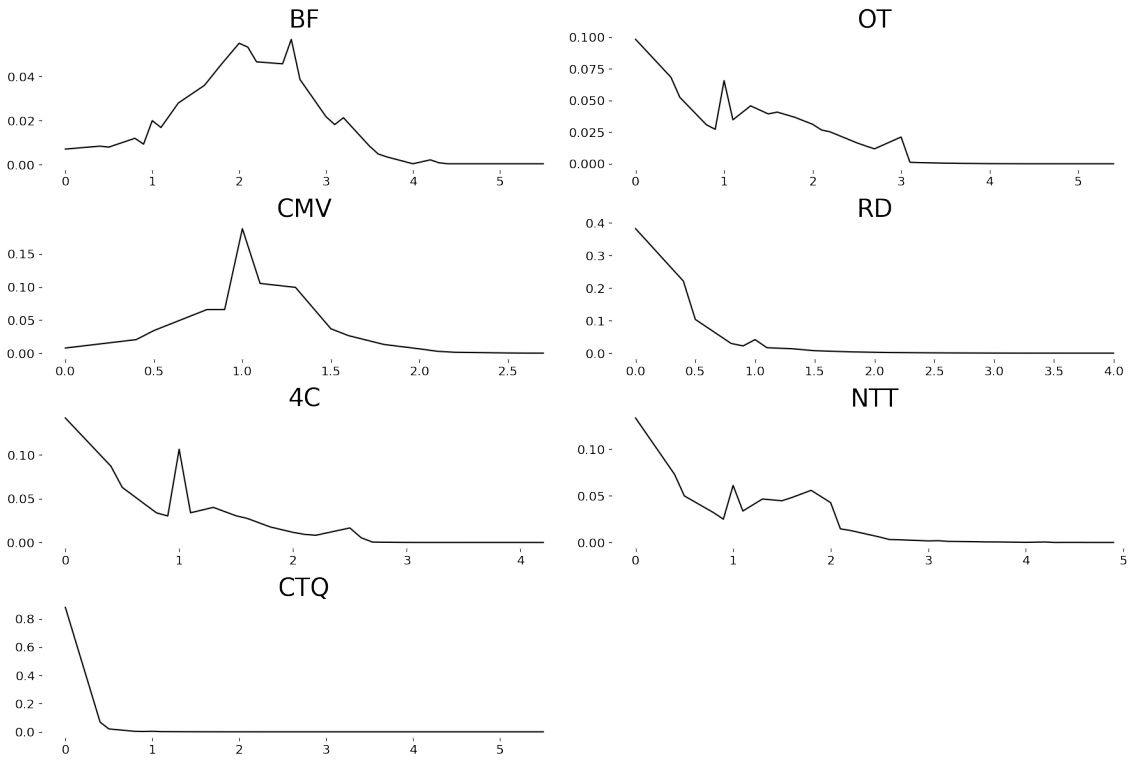

Figure 3.21: Distribution of conversation by log of their tree widths.

both the tree width measure and the number of messages, along with the indication from depth that posts appear to be missing, it seems that we may be able to comfortable assert that the 4chan collection method is missing key pieces of conversations.

The remaining datasets exhibit behavior that is reminiscent of the depth measures. There is typically a high concentration around low conversational widths (small conversations with minimal participants and minimal branching incentive). This begins to attenuate and drop off quickly, implying that only the most salient and hot conversations receive the attention necessary to really inflate the widths of trees.

3.4 Discussion

Having considered a collection of data and discussed the specifics of how they were collected and the observed stats compiled through PyConversations, it is worth stepping back and acknowledging the high-level patterns that one could abstract from this data.

3.4.1 Platform Observations

Linguistic Variability

For data collections from Facebook and Twitter, much greater linguistic variability (in the diversity of languages detected) is observed in comparison to platforms like 4chan and Reddit. While this could be a particular trait that emerges because of the datasets studied here, this does align with previous work that cautions that social media data can present a very limited and biased sample of language and perspective [72].

Character Limits

Twitter is the only platform that has a fairly strict character limit for their posts. This platform rule clearly has significant effect on the data. CTQ, in particular, exhibits a "bunching up" of messages at the upper-limit of this threshold with 15% of all collected posts ending up at this upper limit. While a somewhat more transparent and obvious structural limitation and effect, this empirical observation may serve useful, particularly when aiming to create better representations for social media posts.

Depth Limits

Facebook implements a strict depth-limit to conversations on its platforms. As shown through this analysis, this has an incredibly strong effect that disrupts normal conversation shape, relative to what occurs on other social media sites. While most conversations collected are not too wide nor deep, conversations collected from Facebook are short and fat. This seems to imply that, in lacking the ability to continue to reply to other participants, users simply just bunch their comments into the limited depth levels thus expanding the conversational trees.

Broadcast Limits

A particularly interesting observation from this analysis stems from the amount of posts a single post may reply to. For Facebook and Reddit, this is very strictly set to 1. For Twitter, this limit is technically 2. Only on 4chan is there no limitations to amount of posts that a user may reply to with any one post.

However, as observed in this work, this limitation has pretty minor effects on the data. In general, it does not appear as though users on Twitter use a multi-reference/multi-reply action frequently, as exhibited by NTT. In contrast, CTQ has a higher average out-degree which may suggest that there is a higher chance someone might reply and quote when they are already quote tweeting–in other words, if they're building out a thread of referenced, quoted posts on their own timeline.

Even more interesting is the lack of broadcast on 4chan, despite the fact that the site places no limitations on its users. Despite this extra degree of freedom, the collected data does not lean heavily towards using it. While true that the average out-degree on 4chan is observed to be higher than, say, CTQ or any of the Twitter data, it is only marginally so. Future studies should follow-up on this to confirm this observation, especially in light of some of the other 4chan observations which indicate that there may be signicant gaps in this collection.

3.4.2 Dataset Anomalies

As a final remark, it is worth emphasizing some of the anomalous characteristics of the studied data that this analysis has helped to highlight.

For one, there are several instances that seem to indicate that the 4C collection is missing significant portions of data. Even beginning with the overall temporal distribution, many posts are missing from 2018 to 2020 as indicated by a seemingly artificial ceiling on the amount of data collected in these regions. Additionally, the short time horizons and the

lack of posts, participants, and depth in 4chan conversations could indicate that there is significant data missing as well.

The OT dataset is extremely odd when comparing to the other data. This may be because it is a large Facebook collection or this could be effected by the fact that it is missing several pieces of information, namely the associated conversational depth as well as any information about the users participating in the conversations. Nevertheless, this was not something that was readily apparent until OT was converted into a more universal format through PyConversations and held up in comparison to other social media datasets.

CTQ is incredibly biased towards two-turn quote-tweet "conversations." One could even argue that, given the prevalence of these two-turn exchanges that this dataset may be less-so a "conversation" dataset than just a quote-tweet dataset. However, the quote action is a critical feature for discourse on Twitter and, as exhibited here, CTQ has been a pivotal dataset for comparing the shapes of the other conversational datasets to. If CTQ is what a heavily biased two-turn dataset looks like, then it provides an excellent touch-point for how conversational data can take on more dynamic shapes when extending beyond this limitation.

3.5 Future Work

PyConversations was written with the intent of being an open-source analysis package for cross-platform social media analysis. Ultimately, its construction has in mind many extensions centered around the study of cross-platform phenomena as well as general behaviors of social media conversations. This section highlights several directions of interest as well as the intention (and suggestion) of future lines of work.

3.5.1 Thread Growth & Shape Dynamics

How do conversations grow over time? What shapes do they take over their lifetimes? Do their structural properties indicate anything about their substance? For example, do conversations filled with toxic comments look different in structural shape compared to "good faith" discussions on a sub-reddit like r/ChangeMyView?

Understanding questions like these are of critical importance when considering the implementation of a structural change on a social media platform. If Twitter decides, as it has previously in the past, to change the way that the quote tweet action operates, how will that change the way that conversations unfold on the platform? What properties can be collected to observe the effects that a structural change has on the platform? How might some of these metrics about the shapes of conversations be used to qualify the extent to which a change has had a positive or negative outcome?

Some previous work that aligns well with these questions are works that consider the task of inter-arrival time prediction [73–77]. Consider a conversation that is about to receive a reply. That reply will do one of two things: Add to a depth level that has already existed (branched the conversation out further thus extending its width) or it will add a new depth level (deepening the height of conversation). Perhaps such a process could be modeled with a variational point process, similar to what has been done in [77].

Regardless of whether one seeks to model such processes with point processes or with other

time-series models centered around conversational "burstiness" [78, 79], these studies need a straightforward mechanism for ingesting massive amounts of conversational data and transforming them into input time series. PyConversations makes this transformation easy and straight-forward, making it a strong utility for any researchers seeking to do work similar to these previous works.

3.5.2 Conversational Outcomes

What outcomes occur in a conversation? Is there a productive outcome or is it simply a "slug-fest" of users exchanging ad hominem attacks, hate speech, and other toxic content? These are questions that PyConversation is being actively used to help answer as a utility for structuring and handling social media conversations.

One of the underlying hypotheses explored throughout this work is the idea that through a combination of structural, linguistic, and socially aligned signals, one might be able to better learn representations that may be used to classify and detect phenomena occurring within a conversation.

Furthermore, with better detection, one can consider various structural interventions that could be employed to alter conversational outcomes. For example, past work has strongly considered the idea of comment ranking [80–83]. While not directly considered to alter conversational outcomes (most studies frame discussion around "relevance," similar to how search engines rank their results), it is a simple extension to consider that if comment ranking changes thread dynamics, then it likely also alters the dynamics of conversations and conversational participants.

Along these lines, PyConversations is hoped to help to perform these types of studies that allow for the study and simulation of intervention on conversational outcomes and quality. For example, if conspiratorial or fringe content and discussion can be identified early-on through its dynamics and structure, perhaps it could be down-weighted in relevance rankings, instead of propagating and boosting content that is, at times, inflammatory.

3.5.3 Information Propagation

The flow of information through social networks is of high-interest to this work. Since the introduction of Eli Pariser's notion of filter bubbles [8], much work has gone into quantifying and qualifying the extent to which filter bubbles actually exist [9, 84–86]. Though there is still active debate about filter bubbles, their existence, and their overall influence on social media platforms at large, empirical work like these are critical and the tools to produce such investigations are valuable.

Along different lines, other works have considered information flow through social networks [87, 88]. Though relatively simplistic at this stage, works like this can and should be extended to better understand and model the dynamics of information flow through social media networks. Such works can be empirically grounded in the structures observed in real collected social media data through the usage of packages like PyConversations. Not only can such investigations potentially yield insights into filter bubbles, but also could prove useful in investigating misinformation and disinformation and the ways in which these types of information propagate through social networks.

3.5.4 Bot Detection & Robotic Manipulation

Bots on social platforms are there for a purpose, whether that is transparently described (like Twitter's platform requires, for example) or is hidden by an entity that is misrepresenting itself within a community or social network. Above all, tools that can highlight misleading agents will serve useful, particularly as language generations become increasingly more coherent and believable, and are used to artificially participate within social media conversations [72].

Some works have shown that a useful characteristic for identifying artificial agents within social media ecosystems is their language innovation rate or their novelty rate. As it turns out, humans have a fairly regular measure, so when a social media user differs from this expectation significantly, it can be a positive signal that one is dealing with an artificial agent [7, 67].

Another application of the measure of novelty is to quantify how mixed a text is [66]. This work applied the measure to understand the average value observed for each dataset as a preliminary investigation into how these values differ across collections and social media platforms. A question of interest is how one might be able to use this information to identify conversations that may have social bots embedded in them. Perhaps disruptions to the overall conversational mixture statistics could cue detection systems into the presence of artificial agents.

3.5.5 Simulation

This work has advocated for the study of structural measures of social media platform to better understand, control, and intervene in conversational dynamics. Even more ambitious would be the simulation of these dynamics to obtain better explanations as to why certain phenomena manifests in a particular way. For example, can the simulation of Facebook dynamics allow for better explanations (especially causal explanations) for the shape of their conversational trees? Preliminary work has considered simulation of social media information dynamics through simple quoting models [88], however, future work should extend such ideas further and beyond the explanation of just information flow.

Chapter 4

Representation Learning for Social NLP

In the age of deep learning and NLP, when approaching any task with language data one must ask themselves: to pre-train or to fine-tune? Since the advent of the word2vec algorithm [18, 19], pre-trained language representations have become a prominent feature of many neural architectures.

First came the context-independent word embeddings. These are algorithms like word2vec [18, 19], GloVe [31], and fasttext [32] to name a few popular examples. These representations were often used to initialize embedding layers in a neural architecture and could either be held static (thus expediting the training) or could be dynamically adjusted or tuned (at a higher training cost). For every unique token (type), there is a single corresponding vector representation.

Then came a generation of context-dependent models that used a variety of architectures like Elmo with a bi-LSTM base [89] and OpenAI's GPT [34] that was the first pre-trained model to adopt the new Transformer architecture [33]. The intuition behind these models was that by passing tokens through recurrent or Transformer architectures, one could "contextualize" tokens by having them attend to one another within the neural model. Standard approaches tended to adopt the pre-trained model's weights, optionally fine-tune them on the target domain data, and train a model to perform the desired task. Other approaches also used the representations from these pre-trained models in a manner to static word embeddings.

Since then, there has been widespread adoption of the Transformer architecture and many new models like BERT [20], GPT-2 [21], and more have yielded an entire zoo of Transformer model pre-trained weights to choose from [17]. Yet the question remains, should one pre-train their own model on their own data or would they be better of taking one of the pre-trained Transformer architectures?

Social media-based data invites additional questions about pre-trained language representations. With more specialized information available, information about where, when, and to whom a piece of text is directed to could all add additional context to what is explicitly contained within the written symbols, is it worth exploring a richer set of alternate language representations to exploit this information? Or is it simply enough to incorporate such information shallowly within a final architecture directly for a social

media-based task?

4.1 Methods

This Chapter explores methods for pre-training language representations for a social media-context. While the different approaches explore different training procedures and architectures, several features remain constant across the approaches: the pre-training data, the learning objectives (or rather the subset selected for each approach), and the evaluation methods.

4.1.1 Pre-Training Data

The data used for pre-training are sampled across several different platforms by a variety of different methodologies. Before use, datasets are first converted into a universal format via the PyConversations package. From there, they may be used to generate the signals necessary for the pre-training objectives described in the next subsection. For a more in-depth discussion about the datasets and their properties, see the previous Chapter.

The datasets used, by platform, are:

Twitter

- **NewsTweet** - A dataset introduced in [58]. This dataset features a collection of tweets that are embedded in news articles. Since these tweets represent a collection of social media posts that were deemed "newsworthy," the conversations branching off of such tweets provide an interesting perspective into the sphere of public discourse on Twitter, which this dataset samples.
- **Coordinated Targeting** - During the Summer of 2020, a work was released that highlighted suspicious activity on a number of high-profile Twitter accounts– particularly, odd behavior in their follower dynamics [59]. In an effort to better characterize the unusual observed behavior, a number of timelines that appeared during these suspicious phenomena were collected. Filtering this collection of the quote tweets originating from the accounts results in a different slice of Twitter conversations that contrasts well with the threaded discusses in the previous dataset.

Facebook

- **BuzzFace** - A dataset introduced in [6, 7] which presents the public discourse that was present on the Facebook posts fact-checked by BuzzFeed [60]. This dataset has since been augmented with an auxiliary collection of political/news-oriented Facebook page discourse.

Reddit

- **Change My View** - A dataset introduced in [61] that explores the sub-reddit, r/ChangeMyView. r/ChangeMyView is a particular sub-reddit where users post an opinion they hold and ask for other users to challenge them on it and, as the name suggests, change their view. One aspect of r/ChangeMyView that makes it of interest is the strictly-enforced rules centered around fostering good-faith dialog.

The authors of [61] were interested in understanding the dynamics that lead to users violating the sub-reddit's rules by committing an ad hominem attack (an attack against a user in the sub-reddit). The authors made this historical cross-section of sub-reddit data publicly available.

- **RedditDialog** - In developing an adaptation of GPT-2 [21], the authors of DialoGPT [62] train a language model on 147 M Reddit conversations. Additionally, the authors provide the tools to rebuild a cache of Reddit, which this work uses to construct a dataset of posts. Specifically, the dataset consists of threads from 3 sub-reddits: r/news, r/worldnews, and r/politics. The dataset spans these sub-reddits from their creation up to January 2019.

4chan

4Chan has not been widely studied, although there are several publicly available datasets centered around one board in particular: /pol/ [56, 63]. Other datasets exist as well, though not publicly [57, 64, 65].

To attain a bit more board-diversity, this work uses an ad hoc collection of 4chan boards: news (/news/), history (/his/), science (/sci/), technology (/g/), politically incorrect (/pol/), and paranormal (/x/).

4.1.2 Learning Objectives

This work follows many previous works [20, 22, 90] and attempts to tune models based on two types of training objectives:

Masked Language Modeling (MLM)

Given a sequence of tokens $t_1, ..., t_n$, $p\%$ are sampled at random and either randomly replaced, masked, or left unaltered. The MLM objective tasks a model with un-corrupting the corrupted input by predicting the original tokens. It is important to emphasize that this objective does not consider prediction on the $(1-p)\%$ of the un-touched tokens, so this objective only ever actively trains on $p\%$ of any given sequence.

Specifically, this work uses a variant of MLM which performs Whole-Word Masking (WMM) [53]. The distinction between WMM and the original MLM objective is that when a token is sampled in WMM, all of the non-whitespace separated tokens are also sampled with it. In other words, full "words" are masked this way as opposed to word fragments.

Binary Objectives

Given any two posts that are connected in some manner (or not if, for example, they're a noise contrastive example or negative sample), maximum likelihood estimation can be used to optimize a binary cross-entropy loss objective. Any of the following binary properties could be modeled as a binary prediction between two posts:

- **Reply-Order Prediction (ROP)**: Predict whether the first or second post is the parent post.

- **Same Author (SA)**: Predict whether two posts are written by the same user (within a conversation).
- **Same Recipient (SR)**: Predict whether two posts were directed to the same user (within a conversation).
- **Same Parent (SP)**: Predict whether two posts share a parent post.
- **Same Conversation (SC)**: Predict whether two posts came from the same conversation.

ROP is similar in intention to Next Sentence Prediction [20] and Sentence Order Prediction [90] in that ROP is designed to optimize models towards encoding information about longer-range dependencies and semantics.

The remaining "Same" objectives (SA, SR, SP, SC) are all designed to encourage optimization towards encoding information about what makes posts similar, each by a different criterion. These objectives are similar to what has been tried with SBERT [91] where training seeks to predict a similarity between representations of text.

In this initial work, only ROP is considered with all "same" objectives mentioned, but saved for future consideration.

4.1.3 Evaluation Methods

For evaluation, this work takes a similar approach to previous works [22, 52] and tests model downstream performance on a set of benchmark tasks. Specifically, this work focuses on the TweetEval benchmark [52] as this is a set of 10 sequence classification tasks for the Twitter/social media domain.

TweetEval

TweetEval is a heterogeneous benchmark of Twitter multi-class classification tasks introduced in [52]. TweetEval was assembled through a collection of classification-based SemEval tasks to promote the development and consolidated evaluation of NLU for Twitter (and social media at large). As such, this work uses this pre-constructed benchmark which consists of the following sub-tasks:

- **Emotion Recognition** (SemEval 2018, Task 1 [92]) - Specifically, an adaptation of the original Sub-Task 2. Instead of an eleven class multi-label classification task, TweetEval takes a subset of the data to focus on a four class classification of the emotions *anger*, *joy*, *sadness*, and *optimism*.
- **Emoji Prediction** (SemEval 2018, Task 2 [93]) - Given a tweet's text, this task requires the prediction of one of twenty possible emojis that was paired with the text. Due to Twitter data-sharing rules, the authors of [52] artificially limit the training data originally used with this task to fit the maximal size requirement for public sharing of raw data. This change results in a training set that is a tenth of the original training data size.
- **Irony Detection** (SemEval 2018, Task 3 [94]) - Specifically just Sub-Task A that targets the binary classification of ironic or not.

- **Hate Speech Detection** (SemEval 2019, Task 5 [95]) - This task, also known as Hateval, seeks to evaluate the ability to discriminate and identify hate speech against women and immigrants. Only Sub-Task A (the binary classification of detection) is included in this benchmark.
- **Offensive Language Identification** (SemEval 2019, Task 6 [96]) - A restriction to Sub-Task A, centered around detection of offensive language (and not sub-classification of the types).
- **Sentiment Analysis** (SemEval 2017, Task 4 [97]) - Specifically Sub-Task A, targeting general sentiment classification (without restriction to any given topic).
- **Stance Detection** (SemEval 2016, Task 6 [98]) - Specifically Sub-Task A, exploring supervised classification of Tweets spanning 5 key topics: abortion, atheism, climate change, feminism, and Hillary Clinton.

As can be surmised from these descriptions, TweetEval is a benchmark of multi-class, sequence classification tasks for Tweets. This leaves room for the construction of further benchmarks that can explore performance in other areas.

Fine Tuning

Following previous works [22, 52], fine-tuning for a classification dataset is performed by adding a dense feed-forward layer that projects from the final model output size down to the classification label set.

A limited hyperparameter search is performed (mirroring the fine-tuning parameters from [22]), selecting a batch size $\in \{8, 16, 32\}$ and a learning rate $\in \{1\times10^{-5}, 2\times10^{-5}, 3\times10^{-5}\}$. A maximum of 10 epochs of training are performed on a fine-tuning dataset, with early stopping implemented to select the tuned model that performs best on the dev set.

4.2 Tuning RoBERTa for Social Media

Models like GPT-2 [21] and RoBERTa [22] are not unfamiliar with internet data. In fact, 38 GB of RoBERTa's training is known to be Reddit data. While some work advocates that general language pre-training is beneficial no matter the domain [52], other works question this assumption and investigate whether one can gain more performance using a domain-specific language pre-training [53].

Here, this section considers the performance of the former approach and follows the following procedure:

- Select a general pre-trained Transformer model
- "Pre-tune" on a general training objective in the new domain
- Fine-tune the adjusted general model on any downstream, evaluative dataset

While this approach restricts models to the architectural hyperparameters decided by previous work (including tokenization scheme), this provides an approach that allows one to test a hypothesis without completely re-training a Transformer from the ground-up.

Parameter	Value
Layers (L)	12
Hidden size (H)	768
Attn. Heads (A)	12
FF Size ($4H$)	3072
Sequence Length (S)	512
Vocabulary Size (V)	50,257

Table 4.1: Rough overview of the architectural hyperparameters for RoBERTa-base.

For example, if training on multiple platform's data, is it helpful to learn some sort of platform embedding to help a model differentiate between platforms? Such an embedding could be structured off of the questions considered in Chapter 3 like whether the platform has a board and whether the platform is pseudo-anonymous. To test this idea, a model could be warm-started with general, pre-trained weights (potentially freezing the weights as a feature extraction), augmented with additional weights for processing platform embedding / representation information.

The downside of this approach is that all pre-training and structural biases that are present in the general model will likely be inherited by the domain-specific adaptation. Additionally, it cannot adopt a new tokenization scheme–something that has been observed to be a limiting factor in the biomedical domain resulting in drug and compound names being split across multiple tokens [53]. A common practice in NLP for social media data is to strip URLs and perform special pre-processing to collapse other social media artifacts into the same type of token. Perhaps this is less beneficial to do than retaining these artifacts and having a domain-specific tokenizer that can efficiently tokenize these elements?

Since this work is focused on autoencoding models, a pre-trained RoBERTa model [22] is selected from Hugging Face's *Transformer* library [99]: RoBERTa-base[1].

4.2.1 RoBERTa

RoBERTa is a BERT model [20] with more optimally tuned pre-training hyperparameters and a larger byte-pair encoding (BPE) vocabulary size. A summary of the pertinent architectural hyperparameters can be seen in Table 4.1.

In adopting RoBERTa's weights, one also adopts RoBERTa's tokenizers. As RoBERTa has a fixed, max-sequence length of 512 sub-word tokens and is a general language tokenizer, Table 4.1 considers the size of the posts in each platform's datasets to understand how well pairs of posts will fit within this sequence length. Additionally, the distribution of post sizes by social media domain can be seen in Figure 4.1.

Table 4.2 shows that posts fit well within a length of 256, leaving each post half of a sequence length. Even the least clipped and covered dataset (Reddit) is still well covered, with 97.27% of posts naturally fully fitting within the sequence length of 256. Additionally, not a single platform has less than 90% coverage even when considering a max length of 128 tokens.

[1] https://huggingface.co/roberta-base

Platform	Sub-Words	Max Length			
		32	64	128	256
Twitter	40.51 ± 14.56	31.85 / 79.64	96.30 / 99.42	99.94 / 99.99	100 / 100
Facebook	32.06 ± 59.33	72.92 / 89.53	91.40 / 96.99	97.72 / 99.18	99.34 / 99.75
Reddit	60.71 ± 86.09	45.92 / 73.42	72.96 / 88.94	90.00 / 96.51	97.27 / 99.18
4chan	39.17 ± 54.42	64.95 / 84.68	85.80 / 94.53	95.33 / 98.36	98.65 / 99.64

Table 4.2: Average post size as determined by RoBERTa's tokenizer. Values are average post size with respect to platform ± the standard deviation. For size thresholds, left-most value is indicative of the percent of posts that are less than or equal to the maximum length and the right-most value is the percent of post's tokens that are observed (as a fuzzy-coverage metric).

An additional interesting note is the average sub-word sizes of each platform. For all but one platform, the sizes of posts are around the same number of tokens: 30-40 tokens. Twitter features far less variance in post size, but both 4chan and Facebook appear to be fairly similar in lengths. For Reddit, the average is much larger with a lot higher standard deviation–likely indicative of the longer, more essay-like posts that some users write to one another (particularly on a subreddit like r/ChangeMyView).

4.2.2 Warm-Start Tuning Strategies

For all warm-start tuning experiments (experiments where models are first adapted to the social media data domain, using pre-training objectives), the same general procedure will be followed. An Adam [43] optimizer is used with a standard slanted triangular learning rate schedule with a warm-up of 10% of the steps and a cool-down for the other 90%. The peak learning rate is set to be 1×10^{-4}. Furthermore, a batch size of 2,048 is selected with a maximum training step limit of 62,500. Model performance is evaluated against a dev set every 25 steps, and tuning is halted if dev perplexity is not decreased after 50 consecutive dev evaluations (1,250 steps). Tuned models are always checkpointed at their best evaluated dev perplexity.

Beyond these fixed hyperparameters, two choices are considered for tuned models: tuning data and tuning objective.

Data

As discussed in Chapter 3, data has been collected from four social media domains: Twitter, Facebook, Reddit, and 4Chan. Any of these platforms' data can be used individually or in combination with one another. For example, if a model is to be deployed with specific open-source use-cases in mind, it might be better to tune on Twitter and Reddit, but drop Facebook entirely. 4Chan is a questionable domain. Its inclusion in a tuning set could allow an NLU model to become aware of many slurs, which may be useful in the detection of hate speech. In contrast, it may just toxify a model with storing un-helpful biases. If 4Chan is to be used, extreme caution and testing is advised.

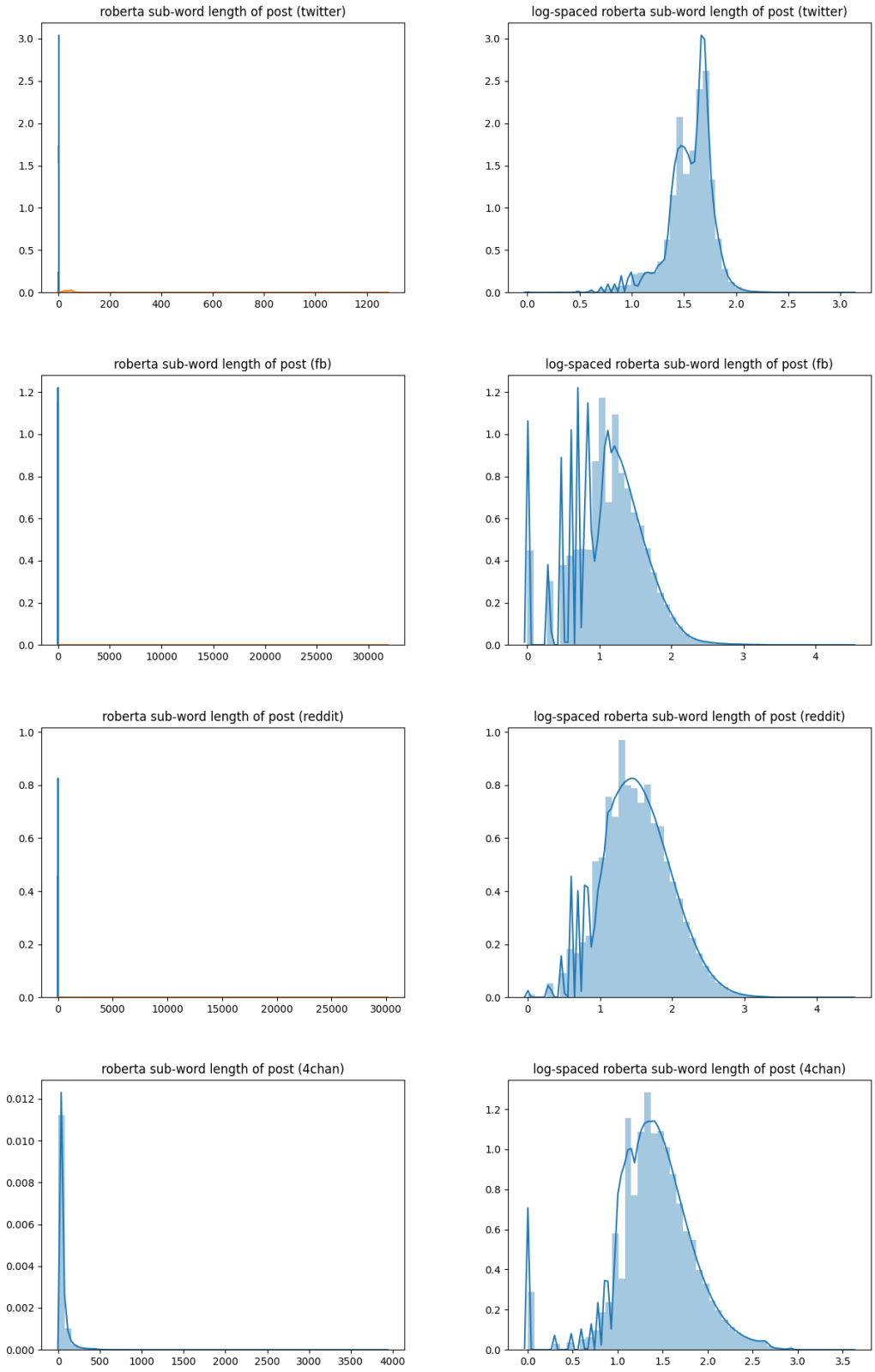

Figure 4.1: Post sub-word length (per RoBERTa tokenization) across platforms. Left column features raw sizes on x-axis, right column features log sizes on x-axis.

Input Format

In previous works like BERT and ALBERT, two text sequences were concatenated with the goal of learning long-range dependencies via cross-attention. This work similarly selects two sequences to concatenate, but does so in a structured way: a post and a reply pair. The hope is the pre-trained base will transfer some of its long-range self-attention operations to extract useful features between parent-child texts.

Pre-Tuning Objectives

Since RoBERTa was first published, whole-word masking (WWM) has been found to be a more efficient method of applying masking to a text [53]. Therefore, this work uses WWM as its masking strategy at a rate of 15%.

Additionally, this work seeks to understand if auxiliary learning objectives can influence a better transfer to the social media domain. Similar to the sentence order prediction objective (SOP) from ALBERT [90], this work experiments with feeding a model with pairs of posts and replies in random order and explicitly tasking the model with predicting if they are in the correct order. This is a binary task that this work refers to as *reply order prediction* (ROP). Just as the SOP objective was designed to encourage long range attention operations, this objective makes explicit that a model should encode information useful for understanding parent-child dependencies.

4.3 Full Pre-Training

A full pre-training of a Transformer model, while yielding much greater flexibility in hyperparamter selection, is incredibly expensive. They are notoriously hard to train from scratch and are known to exhibit unstable training dynamics. Additionally, it is not entirely clear, at this stage, whether it is beneficial to fully train one's own model if they have enough in-domain data. Perhaps it is simply more beneficial to start with the weights of a model that already exhibits decent performance on a task like masked language modeling?

For example, authors in [52] explore the idea of comparing a full pre-training of a Transformer model versus simply fine-tuning a pre-trained model. Their results suggest that, at least for Twitter tasks, a full pre-training is unnecessary. Instead, they recommend to start with a pre-trained model, like RoBERTa, and spend some time pre-tuning it to the target domain.

In contrast to the findings of [52], the authors of PubMedBERT [53] find the opposite to be true of the biomedical domain. Their findings highlight the mismatch of tokenization strategies and advocate that a more domain specific tokenization allows for improvements on linguistic tasks in the biomedical domain. This suggestion about tokenization, especially sub-word tokenization for Transformers, is not a new remark as other works have similarly expressed the importance of domain and obtaining a proper sub-word tokenization encoding [36].

Another significant barrier to creating one's own pre-trained model like BERT or RoBERTa is the necessary hardware and pre-training time cost. When scoping a fully pre-trained model with the available lab hardware, it was calculated that a single experimentation

run (on a 4 GPU setup) would take upwards of 180 days to complete. As this is far beyond the scope of this current work, this option was omitted. Instead, this work looks to highlight different architectural considerations, initialization schemes, and pre-training algorithms to increase the speed and efficiency of the pre-training, yielding these insights in one place for future work and development.

4.3.1 Architecture

The architectural adjustments here are made with autoencoders like BERT and RoBERTa in mind. However, instead of inheriting the architecture of BERT as the base of this model, this work highlights the following adaptations:

Embedding Factorization

In the original BERT paper, the context-independent, sub-word embedding size is tied to the internal, hidden layer size of a model. ALBERT [90] advocates for splitting this size selection into two parameters: an embedding size, E, and a hidden layer size, H. This factorization allows for the internal representation size to vary without increasing the embedding size.

Furthermore, the authors of [90] make the argument that the dimensionality required to learn context-independent sub-word embeddings is much smaller than what is necessary to learn context-dependent internal representations. Acting under this assumption, one can see that it is much more efficient to factor these two sizes and reduce the number of multiplications that are necessary and would only be updated sparsely during training.

For example, consider a vocab size $V = 50,000$, a hidden representation size of $H = 768$, and an embedding size $E = 128$. The cost of computing $O(V \times H) \leq 38.4 \times 10^6$ is far greater than $O(V \times E + E \times H) \leq 6.4 \times 10^6 + 98.3 \times 10^3$. As one can see from this example, the factorization has produced an operation count that is approximately one sixth of the original cost.

Reduced Sequence Length

As noted when warm-starting RoBERTa in the socia domain, social media posts are short with respect to the general text that models like BERT and RoBERTa aim to model. As one of the major bottlenecks of computational cost in a Transformer model is its self-attention cost (quadratic in the sequence length L), reducing the input sequence length would further reduce the cost-per-step of Transformer training and inference.

Here, we train domain-specific Transformer vocabularies with a SentencePiece tokenization to understand the amount a sequence length can be reduced beyond $L = 512$. For each of these domain-specific models, $V = 50,257$, equivalent to the vocabulary sizes for both RoBERTa and GPT-2. A presentation of the amount of truncated posts at various sequence length can be seen in Table 4.3.

Unsurprisingly, Table 4.3 shows that a specialized tokenization tuned for the data from a specific platform does benefit the sub-word token coverage. Arguably more interesting is the fact that training a tokenizer on all of the social media data obtains most of the benefits seen from a platform specific tokenizer, while having generality as a social media

Platform	Sub-Words	Max Length			
		32	64	128	256
TWT-Specific	36.04 ± 15.65	44.84/83.76	96.55/99.48	99.96/99.99	100/100
TWT-Social	36.06 ± 15.67	44.41/83.88	96.47/99.50	99.97/99.99	100/100
TWT-RoBERTa	40.11 ± 18.04	36.55/79.83	93.37/98.90	99.77/99.95	99.98/100
FB-Specific	24.29 ± 51.80	80.79/92.30	93.35/97.70	98.32/99.41	99.53/99.82
FB-Social	24.32 ± 51.92	80.77/92.29	93.36/97.70	98.31/99.41	99.53/99.82
FB-RoBERTa	25.58 ± 56.12	79.52/91.75	92.86/97.50	98.16/99.36	99.49/99.80
RDT-Specific	130.79 ± 181.55	23.12/51.62	45.05/71.40	69.57/86.88	87.57/95.50
RDT-Social	134.35 ± 187.08	22.52/50.29	44.16/70.71	68.72/86.41	87.07/95.28
RDT-RoBERTa	133.87 ± 187.31	23.01/51.25	44.61/70.88	68.86/86.42	87.07/95.28

Table 4.3: Sequence length for domain-specific SentencePiece tokenization models. For platform specific coverages, percentages are given as a hard percentage (percent of posts that are not truncated) and a soft percentage (percent of tokens that are covered, prior to truncation).

domain general tokenizer. The only data source that seems to deviate slightly from this observation is the Reddit data. This does makes sense, however, when one considers the fact that Reddit posts are typically longer-form and more formally written than something like a Twitter post.

Weight Sharing

As noted in works like SBERT [91], BERT's strategy of concatenating two sentences to compute cross-attention between sentences via self-attention is extremely inefficient. SBERT highlights how finding the pair of most similar sentences in a collection of 10,000 sentences using this method would take 50 million inference computations and over 65 hours. Instead, by processing sequences using the same weights (sometimes called Siamese networks or triplet network structures) one can perform the same pair-wise computations in around 5 seconds.

It is not readily obvious that there is significant benefit to pre-training with cross-attention between two social media posts (a post and reply). In fact, by avoiding this one can shorten the max sequence length and increase pre-training speed. Instead, this work considers the explicit modeling of post vectors that can be combined with other post vectors through more computationally efficient operations like cosine similarity during auxiliary learning objectives, similar to sentence order pair.

Interestingly enough, this line of work leads back to sentence representation modeling like the Universal Sentence Encoder [100–102]. Leaning heavily upon an idea, presented in other works [103], that the representation learning problem is easier than the language prediction problem.

Model	Objective	Max Step	Train Loss		Dev Loss	
			Start	End	Start	End
RoBERTa-FB	MLM	21.6 K	4.32	1.40	4.71	1.68
RoBERTa-reddit	MLM	19.3 K	2.95	1.95	2.85	1.63
RoBERTa-4chan	MLM	19.9 K	3.99	2.29	4.29	1.29
SocBERTa-FB	MLM+ROP	23.7 K	17.93	1.50	17.95	1.75
SocBERTa-reddit	MLM+ROP	12.6 K	18.06	5.75	17.98	4.84
SocBERTa-4chan	MLM+ROP	21.9 K	17.89	5.09	17.89	3.29

Table 4.4: Results of pre-tuning models from a general RoBERTa-base checkpoint to various social media domains. The maximum number of steps are included because some models are trained for longer based on when early stopping criteria is reached. Additionally, final training and dev losses are presented to contextualize how well the model fit the tuning objective.

4.4 Experiments

4.4.1 Effects of In-Domain Tuning

As previously highlighted, there is significant uncertainty within the field as to whether one should (when possible) pre-train their own model from scratch in-domain or should warm-start their model with weights from general domain pre-training. In this experiment, the social media domain is considered to understand which is more beneficial: training from scratch, warm-starting, or directly fine-tuning from general to domain specific task.

Models are tuned in-domain as outlined above. Only three domains are selected[2]; combinations of domains are left to future work. Since early stopping is implemented, some models are trained for more steps than others based on how easily the model adapts to the new domain. Table 4.4 indicates the maximum number of steps each domain-specific model were trained as well as its final train and dev loss.

4.4.2 Evaluation

Model	Emoji	Emotion	Hate	Irony	Offensive	Sentiment	Stance	All
RoBERTa-base	33.1 ± 0.2 (33.3)	78.8 ± 1.0 (80.0)	45.6 ± 3.2 (50.9)	65.1 ± 1.7 (67.9)	80.5 ± 0.3 (81.0)	72.6 ± 1.0 (73.4)	69.3 ± 0.6 (70.0)	63.6 ± 0.7 (64.6)
RoBERTa-FB	33.4 ± 0.2 (33.6)	80.6 ± 0.6 (**81.3**)	48.4 ± 3.1 (53.8)	70.7 ± 1.2 (**72.2**)	80.7 ± 1.1 (81.9)	72.5 ± 0.5 (73.1)	70.3 ± 0.4 (70.9)	65.2 ± 0.7 (**66.5**)
SocBERTa-FB	33.9 ± 0.4 (**34.3**)	79.4 ± 0.8 (80.4)	47.1 ± 1.9 (49.0)	71.0 ± 1.2 (72.1)	80.3 ± 0.5 (80.9)	72.7 ± 0.1 (72.8)	71.2 ± 1.5 (**73.3**)	65.1 ± 0.2 (65.4)
RoBERTa-4chan	33.2 ± 0.2 (33.4)	78.6 ± 0.6 (79.0)	51.4 ± 2.6 (**56.0**)	66.9 ± 1.5 (68.6)	81.1 ± 0.7 (81.7)	72.2 ± 0.4 (72.9)	70.1 ± 0.9 (71.5)	64.8 ± 0.2 (65.0)
SocBERTa-4chan	33.2 ± 0.3 (33.5)	78.1 ± 0.3 (78.3)	49.8 ± 2.1 (53.1)	65.4 ± 1.9 (67.2)	79.9 ± 0.5 (80.5)	72.0 ± 0.4 (72.6)	70.6 ± 0.6 (71.3)	64.2 ± 0.5 (64.9)
RoBERTa-reddit	33.0 ± 0.2 (33.4)	79.3 ± 0.9 (80.7)	48.2 ± 2.3 (51.3)	65.6 ± 3.1 (69.1)	80.8 ± 0.7 (**82.0**)	73.1 ± 0.4 (**73.7**)	70.0 ± 0.6 (70.8)	64.3 ± 0.8 (65.3)
SocBERTa-reddit	33.2 ± 0.2 (33.5)	78.4 ± 0.7 (79.1)	48.2 ± 1.3 (49.4)	68.0 ± 0.4 (68.6)	80.8 ± 1.0 (81.7)	72.5 ± 0.3 (73.0)	69.9 ± 1.4 (70.9)	64.4 ± 0.2 (64.7)
SotA	36.0	79.8	65.1	70.5	82.9	72.9	72.6	
Metric	M-F_1	M-F_1	M-F_1	$F^{(i)}$	M-F_1	M-Rec	AVG($F^{(a)}$, $F^{(f)}$)	TE

Table 4.5: Model comparison on TweetEval benchmark. All results are displayed for the test split. Results are averaged ± standard deviation over 5 random seeds. Parentheses indicates the maximal performance obtained in a single run. Bolded values are the maximal observed during experimentation.

Table 4.5 displays the final test performance of tuned models on the TweetEval benchmark. For each model, fine-tuning is performed with 5 random seeds and averaged over the trials

[2]At time of model development, Twitter data did not have size parity with respect to other data domains.

to help control for randomness in model performance. Additionally, maximal performance for each model is recorded in parentheses.

From this evaluation, one should note a few things:

- Tuning in the desired domain is extremely beneficial. This is somewhat trivial and obvious from past work, but is a reassuring sanity check that the warm-start tuning experiments were not for naught. In all tasks (and overall, when averaging scores for the final TE metric), one of the tuned models outperforms the baseline RoBERTa model.
- Tuning on Facebook data seems to be beneficial for Twitter tasks in TweetEval. This is somewhat surprising, although out of the 3 domains attempted, perhaps Facebook is most aligned with the style of Twitter posts. The Facebook trained models are the best in all but the sentiment, offensive, and hate-speech detection sub-tasks.
- 4chan data can actually be extremely helpful—when seeking to detect hate speech, that is. Although not immensely helpful for many of the other tasks, 4chan does provide significant boosts to performance of the models trained for hate speech detection.
- ROP-tuned models do not experience too much of a difference and for many of the models, the auxiliary objective appears to harm performance. Future work should investigate this further, but this aligns with other objections to the importance of auxiliary objectives [90].

4.4.3 Discussion

As is apparent from past works [52, 53] as well as the results on the TweetEval benchmark, it is highly beneficial to tune generally pre-trained models on the domain of interest. One of the most successful experiments resulted in more than a 5 point increase in F1 score on hate speech detection when exposing the RoBERTa model to 4chan data. This is interesting and insightful. Future work should consider such an approach going forward, with caution, as it is still uncertain what negative effects could arise from the exposure of generally pre-trained models to particularly toxic data domains (beyond simply being able to detect hateful/toxic content more accurately).

Additionally, it is worth remarking on the somewhat *underwhelming* performance increases attained by this line of experimentation. For warm-start fine-tuning experiments, where a single domain tuning takes upwards of a week on the hardware available, is it truly worth this adjustment if the result is just a moderate 1-3 point increase in most settings? Such questions and reflection requires another look at whether the pre-training is really what is necessary to adjust for these types of tasks or if there is a different methodology or approach that would be more successful.

Concurrent with the conclusion of this work is the publication of another work which suggests that Transformer-based autoencoding models may not be doing as much as previous researchers have claimed [104]. Specifically, the authors in [104] explore questions of word order and whether it truly even matters in pre-training. Their results are surprising; many of them seem to suggest that word order does not impact a model that much, and

that what is truly important is the ability for a model to be able to fit higher-order co-occurrence statistics and to incorporate some notion of position into pre-training, which can then be recovered or mapped to the actual notion of order and position present in the downstream, target task.

Regardless of the precise interpretation one takes from this work, it does offer some explanatory power for why certain results are observed in the experiments here: Transformer models simply may not be doing, internally, what is assumed and different tuning objectives are needed to tackle the more socially-oriented tasks that are sought after for the social media domain. That's not to even say that tuning, in the fashion presented here, cannot yield even higher performance with more data, but rather that, perhaps, this should indicate that this is not the most fruitful line of approach.

In line with these recent observations and experimental outcomes, this work is actively exploring new methodologies for initializing language representations using a data-free method.

Chapter 5

Conclusion

Thus concludes this undergraduate thesis exploration into representation learning in the social media domain. This work has covered both a universal data model that will be utilized when tackling new social media, cross-platform studies as well as for the compilation of training data for machine learning models in this ever-growing social domain. This work also explored the tuning of Transformer-based language models, finding moderate yet underwhelming success when transferring models to the proper domain (social media) prior to tuning for the choice task. While already apparent from previous, this gives further empirical support to these notions as well as providing an interesting perspective of the exposure of pre-trained models to toxic data domains when one seeks to classify and detect hate speech.

The most significant outcome from this work is an abstract and concrete one. In the abstract, this work has introduced an underlying data model for cross-platform social media analysis. While by no means does this work proclaim the model is optimal or that it will not be surpassed in a future update to it, it does present itself as an initial starting point that future works can revise, adjust, or completely deviate from. Much more concretely, this work also produces a Python package to facilitate such cross-platform social media analysis. To exhibit this kind of exploration and what the data model and package enable, this work considered several social media datasets, highlighted the effects of structural decisions, and outlined a number of paths that can be pursued using this data model from information propagation to filter bubbles to the growth and evolution of singular posts and entire social communities.

This work also explored the warm-start fine-tuning approach for large Transformer-based language models, a technique where one begins with the weights of a pre-trained model, uses a pre-training objective to adapt to a new data domain, and then finally fine-tunes to whatever target task is desired (within that new domain). Exploring a couple of pre-training objectives and several data domains, this work found benefit (albeit, minimal benefit) from taking this approach, managing to surpass baseline models in most combinations. That said, the results in this work along with more far-reaching results like that of [104] indicate that there is a fundamental disconnect between what it is that these models are fitting during pre-training and what exactly makes them so successful. In line with these glaring gaps in what is understood and what is hypothesized, future work that stems from this project is currently focused on whether or not pre-training is just an extremely computationally expensive method for weight initialization and if data-free,

theory-driven approaches can be developed that circumvent such a high temporal and computational cost.

Bibliography

[1] Alfred North Whitehead. "Symbolism: Its Meaning and Effect". In: (1928) (page 1).

[2] Liang Wu, Fred Morstatter, Kathleen M Carley, and Huan Liu. "Misinformation in social media: definition, manipulation, and detection". In: *ACM SIGKDD Explorations Newsletter* 21.2 (2019), pp. 80–90 (page 1).

[3] Caroline Jack. "Lexicon of lies: Terms for problematic information". In: *Data & Society* 3 (2017), p. 22 (page 1).

[4] Chengcheng Shao, Giovanni Luca Ciampaglia, Alessandro Flammini, and Filippo Menczer. "Hoaxy: A platform for tracking online misinformation". In: *Proceedings of the 25th international conference companion on world wide web*. 2016, pp. 745–750 (page 1).

[5] Christine Geeng, Savanna Yee, and Franziska Roesner. "Fake News on Facebook and Twitter: Investigating How People (Don't) Investigate". In: *Proceedings of the 2020 CHI conference on human factors in computing systems*. 2020, pp. 1–14 (page 1).

[6] Giovanni C Santia and Jake Ryland Williams. "Buzzface: A news veracity dataset with facebook user commentary and egos". In: *Twelfth International AAAI Conference on Web and Social Media*. 2018 (pages 1, 16, 45).

[7] Giovanni C Santia, Munif Ishad Mujib, and Jake Ryland Williams. "Detecting Social Bots on Facebook in an Information Veracity Context". In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 13. 2019, pp. 463–472 (pages 1, 16, 17, 26, 35, 36, 43, 45).

[8] Eli Pariser. *The filter bubble: How the new personalized web is changing what we read and how we think*. Penguin, 2011 (pages 1, 42).

[9] Uthsav Chitra and Christopher Musco. "Analyzing the impact of filter bubbles on social network polarization". In: *Proceedings of the 13th International Conference on Web Search and Data Mining*. 2020, pp. 115–123 (pages 1, 42).

[10] Seth Flaxman, Sharad Goel, and Justin M Rao. "Filter bubbles, echo chambers, and online news consumption". In: *Public opinion quarterly* 80.S1 (2016), pp. 298–320 (page 1).

[11] Dominic DiFranzo and Kristine Gloria-Garcia. "Filter bubbles and fake news". In: *XRDS: Crossroads, The ACM Magazine for Students* 23.3 (2017), pp. 32–35 (page 1).

[12] 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*. Vol. 5. 1. 2011 (page 1).

[13] Mainack Mondal, Leandro Araújo Silva, and Fabrício Benevenuto. "A measurement study of hate speech in social media". In: *Proceedings of the 28th acm conference on hypertext and social media*. 2017, pp. 85–94 (page 1).

[14] Binny Mathew, Ritam Dutt, Pawan Goyal, and Animesh Mukherjee. "Spread of hate speech in online social media". In: *Proceedings of the 10th ACM conference on web science*. 2019, pp. 173–182 (page 1).

[15] Hans Jonas. "Technology and responsibility: Reflections on the new tasks of ethics". In: *Social Research* (1973), pp. 31–54 (page 1).

[16] Marshall McLuhan. *The medium is the message*. 1964 (pages 1, 2, 4).

[17] Xipeng Qiu, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang. "Pre-trained models for natural language processing: A survey". In: *arXiv preprint arXiv:2003.08271* (2020) (pages 2, 5, 44).

[18] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. "Distributed representations of words and phrases and their compositionality". In: *Advances in neural information processing systems*. 2013, pp. 3111–3119 (pages 2, 5, 44).

[19] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. "Efficient estimation of word representations in vector space". In: *arXiv preprint arXiv:1301.3781* (2013) (pages 2, 5, 44).

[20] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. 2019. arXiv: 1810.04805 [cs.CL] (pages 2, 6, 44, 46, 47, 49).

[21] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. "Language models are unsupervised multitask learners". In: *OpenAI blog* 1.8 (2019), p. 9 (pages 2, 16, 44, 46, 48).

[22] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. "Roberta: A robustly optimized bert pretraining approach". In: *arXiv preprint arXiv:1907.11692* (2019) (pages 2, 7, 46–49).

[23] Marshall McLuhan. *Understanding media: The extensions of man*. MIT press, 1994 (page 4).

[24] Mark Federman. "What is the Meaning of the Medium is the Message". In: *Preuzeto sa http://individual. utoronto. ca/markfederman/MeaningTheMediumistheMessage. pdf. Uticaj interneta na tradicionalne medije* (2004) (page 4).

[25] Danah Boyd, Scott Golder, and Gilad Lotan. "Tweet, tweet, retweet: Conversational aspects of retweeting on twitter". In: *2010 43rd Hawaii international conference on system sciences*. IEEE. 2010, pp. 1–10 (page 4).

[26] Kiran Garimella, Ingmar Weber, and Munmun De Choudhury. "Quote RTs on Twitter: usage of the new feature for political discourse". In: *Proceedings of the 8th ACM Conference on Web Science*. 2016, pp. 200–204 (page 4).

[27] Ramon Villa-Cox, Sumeet Kumar, Matthew Babcock, and Kathleen M Carley. "Stance in Replies and Quotes (SRQ): A New Dataset For Learning Stance in Twitter Conversations". In: *arXiv preprint arXiv:2006.00691* (2020) (page 4).

[28] Jiue-An Yang, Ming-Hsiang Tsou, Chin-Te Jung, Christopher Allen, Brian H Spitzberg, Jean Mark Gawron, and Su-Yeon Han. "Social media analytics and research testbed (SMART): Exploring spatiotemporal patterns of human dynamics with geo-targeted social media messages". In: *Big Data & Society* 3.1 (2016), p. 2053951716652914 (page 5).

[29] Iyad Rahwan, Manuel Cebrian, Nick Obradovich, Josh Bongard, Jean-François Bonnefon, Cynthia Breazeal, Jacob W. Crandall, Nicholas A. Christakis, Iain D. Couzin, Matthew O. Jackson, Nicholas R. Jennings, Ece Kamar, Isabel M. Kloumann, Hugo Larochelle, David Lazer, Richard McElreath, Alan Mislove, David C. Parkes, Alex 'Sandy' Pentland, Margaret E. Roberts, Azim Shariff, Joshua B. Tenenbaum, and Michael Wellman. "Machine behaviour". In: *Nature* 568.7753 (2019), pp. 477–486. DOI: 10.1038/s41586-019-1138-y. URL: https://doi.org/10.1038/s41586-019-1138-y (page 5).

[30] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. "A neural probabilistic language model". In: *Journal of machine learning research* 3.Feb (2003), pp. 1137–1155 (page 5).

[31] Jeffrey Pennington, Richard Socher, and Christopher D Manning. "Glove: Global vectors for word representation". In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 2014, pp. 1532–1543 (pages 5, 44).

[32] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. "Enriching Word Vectors with Subword Information". In: *Transactions of the Association for Computational Linguistics* 5 (2017), pp. 135–146. ISSN: 2307-387X (pages 5, 44).

[33] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention is all you need". In: *Advances in neural information processing systems*. 2017, pp. 5998–6008 (pages 5, 6, 44).

[34] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. *Improving language understanding by generative pre-training*. 2018 (pages 6, 44).

[35] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. "Layer normalization". In: *arXiv preprint arXiv:1607.06450* (2016) (page 6).

[36] Martin Popel and Ondřej Bojar. "Training Tips for the Transformer Model". In: *arXiv preprint arXiv:1804.00247* (2018) (pages 6, 52).

[37] Alexei Baevski and Michael Auli. "Adaptive Input Representations for Neural Language Modeling". In: *International Conference on Learning Representations*. 2018 (page 6).

[38] Liyuan Liu, Xiaodong Liu, Jianfeng Gao, Weizhu Chen, and Jiawei Han. "Understanding the Difficulty of Training Transformers". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020)*. 2020 (page 6).

[39] Qiang Wang, Bei Li, Tong Xiao, Jingbo Zhu, Changliang Li, Derek F Wong, and Lidia S Chao. "Learning Deep Transformer Models for Machine Translation". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 2019, pp. 1810–1822 (page 6).

[40] Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tie-Yan Liu. "On layer normalization in the transformer architecture". In: *arXiv preprint arXiv:2002.04745* (2020) (page 6).

[41] Minjia Zhang and Yuxiong He. "Accelerating Training of Transformer-Based Language Models with Progressive Layer Dropping". In: *Advances in Neural Information Processing Systems* 33 (2020) (page 6).

[42] Jingzhao Zhang, Sai Praneeth Karimireddy, Andreas Veit, Seungyeon Kim, Sashank J Reddi, Sanjiv Kumar, and Suvrit Sra. "Why ADAM beats SGD for attention models". In: *arXiv preprint arXiv:1912.03194* (2019) (page 6).

[43] Diederik P Kingma and Jimmy Ba. "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980* (2014) (pages 6, 50).

[44] Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. "On the variance of the adaptive learning rate and beyond". In: *arXiv preprint arXiv:1908.03265* (2019) (page 6).

[45] Michael McCloskey and Neal J Cohen. "Catastrophic interference in connectionist networks: The sequential learning problem". In: *Psychology of learning and motivation*. Vol. 24. Elsevier, 1989, pp. 109–165 (page 7).

[46] Roger Ratcliff. "Connectionist models of recognition memory: constraints imposed by learning and forgetting functions." In: *Psychological review* 97.2 (1990), p. 285 (page 7).

[47] Ian J Goodfellow, Mehdi Mirza, Da Xiao, Aaron Courville, and Yoshua Bengio. "An empirical investigation of catastrophic forgetting in gradient-based neural networks". In: *arXiv preprint arXiv:1312.6211* (2013) (page 7).

[48] James L McClelland, Bruce L McNaughton, and Randall C O'Reilly. "Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory." In: *Psychological review* 102.3 (1995), p. 419 (page 7).

[49] Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. "How to fine-tune bert for text classification?" In: *China National Conference on Chinese Computational Linguistics*. Springer. 2019, pp. 194–206 (page 7).

[50] Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith. "Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping". In: *arXiv preprint arXiv:2002.06305* (2020) (page 7).

[51] Sinno Jialin Pan and Qiang Yang. "A survey on transfer learning". In: *IEEE Transactions on knowledge and data engineering* 22.10 (2009), pp. 1345–1359 (page 7).

[52] Francesco Barbieri, Jose Camacho-Collados, Leonardo Neves, and Luis Espinosa-Anke. "TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification". In: *arXiv preprint arXiv:2010.12421* (2020) (pages 7, 47, 48, 52, 56).

[53] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. "Domain-specific language model pretraining for biomedical natural language processing". In: *arXiv preprint arXiv:2007.15779* (2020) (pages 7, 46, 48, 49, 52, 56).

[54] Twitter. *Additional steps we're taking ahead of the 2020 US Election*. 2020. URL: https://blog.twitter.com/en_us/topics/company/2020/2020-election-changes.html (page 10).

[55] Facebook Business. *Restricting Data Access and Protecting People's Information on Facebook*. 2018. URL: https://www.facebook.com/business/news/restricting-data-access-and-protecting-peoples-information-on-facebook (visited on 04/04/2018) (page 12).

[56] Antonis Papasavva, Savvas Zannettou, Emiliano De Cristofaro, Gianluca Stringhini, and Jeremy Blackburn. "Raiders of the lost kek: 3.5 years of augmented 4chan posts from the politically incorrect board". In: *Proceedings of the International*

AAAI Conference on Web and Social Media. Vol. 14. 2020, pp. 885–894 (pages 12, 17, 46).

[57] Michael Bernstein, Andrés Monroy-Hernández, Drew Harry, Paul André, Katrina Panovich, and Greg Vargas. "4chan and/b: An Analysis of Anonymity and Ephemerality in a Large Online Community". In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 5. 1. 2011 (pages 13, 17, 46).

[58] Munif Ishad Mujib, Hunter Scott Heidenreich, Colin J Murphy, Giovanni C Santia, Asta Zelenkauskaite, and Jake Ryland Williams. "NewsTweet: A Dataset of Social Media Embedding in Online Journalism". In: *arXiv preprint arXiv:2008.02870* (2020) (pages 16, 45).

[59] Hunter Scott Heidenreich, Munif Ishad Mujib, and Jake Ryland Williams. "Investigating Coordinated'Social'Targeting of High-Profile Twitter Accounts". In: *arXiv preprint arXiv:2008.02874* (2020) (pages 16, 45).

[60] Craig Silverman, Lauren Strapagiel, Hamza Shaban, Ellie Hall, and Jeremy Singer-Vine. *Hyperpartisan Facebook Pages Are Publishing False And Misleading Information At An Alarming Rate*. 2016. URL: https://www.buzzfeednews.com/article/craigsilverman/partisan-fb-pages-analysis (visited on 10/20/2016) (pages 16, 45).

[61] Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. "Before Name-Calling: Dynamics and Triggers of Ad Hominem Fallacies in Web Argumentation". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. 2018, pp. 386–396 (pages 16, 45, 46).

[62] Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. "DialoGPT: Large-Scale Generative Pre-training for Conversational Response Generation". In: *ACL, system demonstration*. 2020 (pages 16, 46).

[63] Emilija Jokubauskaitė and Stijn Peeters. "Generally curious: Thematically distinct datasets of general threads on 4chan/pol". In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 14. 2020, pp. 863–867 (pages 17, 46).

[64] Gabriel Hine, Jeremiah Onaolapo, Emiliano De Cristofaro, Nicolas Kourtellis, Ilias Leontiadis, Riginos Samaras, Gianluca Stringhini, and Jeremy Blackburn. "Kek, cucks, and god emperor trump: A measurement study of 4chan's politically incorrect forum and its effects on the web". In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 11. 1. 2017 (pages 17, 46).

[65] Marc Tuters and Sal Hagen. "(((They))) rule: Memetic antagonism and nebulous othering on 4chan". In: *new media & society* 22.12 (2020), pp. 2218–2237 (pages 17, 46).

[66] Jake Ryland Williams, James P Bagrow, Christopher M Danforth, and Peter Sheridan Dodds. "Text mixing shapes the anatomy of rank-frequency distributions". In: *Physical Review E* 91.5 (2015), p. 052811 (pages 17, 26, 35, 36, 43).

[67] Eric M Clark, Jake Ryland Williams, Chris A Jones, Richard A Galbraith, Christopher M Danforth, and Peter Sheridan Dodds. "Sifting robotic from organic text: a natural language approach for detecting automation on Twitter". In: *Journal of computational science* 16 (2016), pp. 1–7 (pages 17, 43).

[68] Jake Williams. "Boundary-based MWE segmentation with text partitioning". In: *Proceedings of the 3rd Workshop on Noisy User-generated Text*. 2017, pp. 1–10 (page 18).

[69] Rico Sennrich, Barry Haddow, and Alexandra Birch. "Neural Machine Translation of Rare Words with Subword Units". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2016, pp. 1715–1725 (page 18).

[70] Taku Kudo and John Richardson. "Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing". In: *arXiv preprint arXiv:1808.06226* (2018) (page 18).

[71] Jake Ryland Williams, Paul R Lessard, Suma Desu, Eric M Clark, James P Bagrow, Christopher M Danforth, and Peter Sheridan Dodds. "Zipf's law holds for phrases, not words". In: *Scientific reports* 5.1 (2015), pp. 1–7 (page 29).

[72] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 2021, pp. 610–623 (pages 39, 43).

[73] Michal Lukasik, Trevor Cohn, and Kalina Bontcheva. "Point process modelling of rumour dynamics in social media". In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. 2015, pp. 518–523 (page 41).

[74] R. Krohn and T. Weninger. "Modelling Online Comment Threads from their Start". In: *2019 IEEE International Conference on Big Data (Big Data)*. 2019, pp. 820–829 (page 41).

[75] Mohammad Akbari, Alberto Cetoli, Stefano Bragaglia, Andrew D O'Harney, Marc Sloan, and Jun Wang. "Modeling User Return Time Using Inhomogeneous Poisson Process". In: *European Conference on Information Retrieval*. Springer. 2019, pp. 37–44 (page 41).

[76] S. Dutta, D. Das, and T. Chakraborty. "Modeling Engagement Dynamics of Online Discussions using Relativistic Gravitational Theory". In: *2019 IEEE International Conference on Data Mining (ICDM)*. 2019, pp. 180–189 (page 41).

[77] Zhen Pan, Zhenya Huang, Defu Lian, and Enhong Chen. "A Variational Point Process Model for Social Event Sequences". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 01. 2020, pp. 173–180 (page 41).

[78] Sejeong Kwon and Meeyoung Cha. "Modeling bursty temporal pattern of rumors". In: *Eighth International AAAI Conference on Weblogs and Social Media*. 2014 (page 42).

[79] Sejeong Kwon, Meeyoung Cha, and Kyomin Jung. "Rumor detection over varying time windows". In: *PloS one* 12.1 (2017), e0168344 (page 42).

[80] Aaron Jaech, Victoria Zayats, Hao Fang, Mari Ostendorf, and Hannaneh Hajishirzi. "Talking to the crowd: What do people react to in online discussions?" In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. 2015, pp. 2026–2031 (page 42).

[81] Greg Stoddard. "Popularity dynamics and intrinsic quality in reddit and hacker news". In: *Ninth International AAAI Conference on Web and Social Media*. 2015 (page 42).

[82] Maria Glenski and Tim Weninger. "Rating effects on social news posts and comments". In: *ACM Transactions on Intelligent Systems and Technology (TIST)* 8.6 (2017), pp. 1–19 (page 42).

[83] Maria Glenski, Greg Stoddard, Paul Resnick, and Tim Weninger. "Guessthekarma: a game to assess social rating systems". In: *Proceedings of the ACM on Human-Computer Interaction* 2.CSCW (2018), pp. 1–15 (page 42).

[84] Tien T. Nguyen, Pik-Mai Hui, F. Maxwell Harper, Loren Terveen, and Joseph A. Konstan. "Exploring the Filter Bubble: The Effect of Using Recommender Systems on Content Diversity". In: *Proceedings of the 23rd International Conference on World Wide Web*. WWW '14. Seoul, Korea: Association for Computing Machinery, 2014, pp. 677–686. ISBN: 9781450327442. DOI: 10.1145/2566486.2568012. URL: https://doi.org/10.1145/2566486.2568012 (page 42).

[85] Frederik Zuiderveen Borgesius, Damian Trilling, Judith Möller, Balázs Bodó, Claes H De Vreese, and Natali Helberger. "Should we worry about filter bubbles?" In: *Internet Policy Review. Journal on Internet Regulation* 5.1 (2016) (page 42).

[86] Cristian Vaccari, Augusto Valeriani, Pablo Barberá, John T. Jost, Jonathan Nagler, and Joshua A. Tucker. "Of Echo Chambers and Contrarian Clubs: Exposure to Political Disagreement Among German and Italian Users of Twitter". In: *Social Media + Society* 2.3 (2016), p. 2056305116664221. DOI: 10.1177/2056305116664221. eprint: https://doi.org/10.1177/2056305116664221. URL: https://doi.org/10.1177/2056305116664221 (page 42).

[87] Matthew S Weber and Peter Monge. "The flow of digital news in a network of sources, authorities, and hubs". In: *Journal of Communication* 61.6 (2011), pp. 1062–1081 (page 42).

[88] James P Bagrow and Lewis Mitchell. "The quoter model: A paradigmatic model of the social flow of written information". In: *Chaos: An Interdisciplinary Journal of Nonlinear Science* 28.7 (2018), p. 075304 (pages 42, 43).

[89] Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. "Deep Contextualized Word Representations". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. 2018, pp. 2227–2237 (page 44).

[90] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. *ALBERT: A Lite BERT for Self-supervised Learning of Language Representations*. 2020. arXiv: 1909.11942 [cs.CL] (pages 46, 47, 52, 53, 56).

[91] Nils Reimers and Iryna Gurevych. "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 2019, pp. 3973–3983 (pages 47, 54).

[92] Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. "Semeval-2018 task 1: Affect in tweets". In: *Proceedings of the 12th international workshop on semantic evaluation*. 2018, pp. 1–17 (page 47).

[93] Francesco Barbieri, Jose Camacho-Collados, Francesco Ronzano, Luis Espinosa Anke, Miguel Ballesteros, Valerio Basile, Viviana Patti, and Horacio Saggion. "Semeval 2018 task 2: Multilingual emoji prediction". In: *Proceedings of The 12th International Workshop on Semantic Evaluation*. 2018, pp. 24–33 (page 47).

[94] Cynthia Van Hee, Els Lefever, and Véronique Hoste. "Semeval-2018 task 3: Irony detection in english tweets". In: *Proceedings of The 12th International Workshop on Semantic Evaluation*. 2018, pp. 39–50 (page 47).

[95] Valerio Basile, Cristina Bosco, Elisabetta Fersini, Nozza Debora, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, Manuela Sanguinetti, et al. "Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter". In: *13th International Workshop on Semantic Evaluation*. Association for Computational Linguistics. 2019, pp. 54–63 (page 48).

[96] Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. "SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval)". In: *Proceedings of the 13th International Workshop on Semantic Evaluation*. 2019, pp. 75–86 (page 48).

[97] Sara Rosenthal, Noura Farra, and Preslav Nakov. "SemEval-2017 task 4: Sentiment analysis in Twitter". In: *Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017)*. 2017, pp. 502–518 (page 48).

[98] Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. "Semeval-2016 task 6: Detecting stance in tweets". In: *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. 2016, pp. 31–41 (page 48).

[99] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. "HuggingFace's Transformers: State-of-the-art Natural Language Processing". In: *ArXiv* abs/1910.03771 (2019) (page 49).

[100] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. "Universal Sentence Encoder for English". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. 2018, pp. 169–174 (page 54).

[101] Matthew Henderson, Rami Al-Rfou, Brian Strope, Yun-Hsuan Sung, László Lukács, Ruiqi Guo, Sanjiv Kumar, Balint Miklos, and Ray Kurzweil. "Efficient natural language response suggestion for smart reply". In: *arXiv preprint arXiv:1705.00652* (2017) (page 54).

[102] Matthew Henderson, Iñigo Casanueva, Nikola Mrkšić, Pei-Hao Su, Tsung-Hsien Wen, and Ivan Vulić. "Convert: Efficient and accurate conversational representations from transformers". In: *arXiv preprint arXiv:1911.03688* (2019) (page 54).

[103] Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. "Generalization through Memorization: Nearest Neighbor Language Models". In: *International Conference on Learning Representations*. 2019 (page 54).

[104] Koustuv Sinha, Robin Jia, Dieuwke Hupkes, Joelle Pineau, Adina Williams, and Douwe Kiela. "Masked language modeling and the distributional hypothesis: Order word matters pre-training for little". In: *arXiv preprint arXiv:2104.06644* (2021) (pages 56, 58).