Novice Language Adaption in Social Media Forums

By

Alexander Brooks

A dissertation submitted in partial fulfillment of

the requirements for the degree of

Doctor of Philosophy

(Computer Sciences)

at the

University of Wisconsin–Madison

2021

Date of final oral examination: August 30, 2021

The dissertation is approved by the following members of the Final Oral Committee:

Matthew Berland, Associate Professor, Computer Sciences
Xiaojin Zhu, Professor, Computer Sciences
Eric Bach, Professor, Computer Sciences
Mitchell Nathan, Professor, Educational Psychology

Contents

Abstract

A new generation is growing up who have known the internet for their entire lives. More people incorporate social media into more aspects of our lives, including how we learn and negotiate our identities. While formal online classrooms such as MOOCs are a part of the picture, learning also occurs in informal settings such as affinity groups' forums. In the latter spaces, veteran and novice members may interact and engage in ad hoc teaching and learning practices. It is increasingly important to understand the mechanics by which learning functions in these social media forums. One lens we can use is the changing language use practices of novices as they learn. I examine how novices in several Reddit forums change their language production over time to adapt to their communities, focusing on novices' word choice and particularly their use of local jargon. I first demonstrate conditions necessary to separate novices from veterans automatically using machine learning. Through a second statistical analysis of word choice distributions, I find that some novices appear to follow distinct phases of jargon appropriation followed by winnowing. My findings collectively suggest a possible trajectory of language learning in informal online communities.

Chapter 1

Introduction

The research presented in this document focuses on conversational language used by participants in social media and how individuals' language use changes over time. Its goal is twofold: First, using the lens of vocabulary, to shed some light on the mechanics which govern how people change their social media writing patterns. Second, to develop a proof of concept for computational analysis of informal learning and language change in this setting, which can then be extended in future work to other elements of conversational language use. It is my hope that this dissertation will be accessible and useful to colleagues from a variety of backgrounds.

To my fellow computer scientists, I present my work as an application of ML and NLP tools to social media text data, grounding my analysis through the lens of the education research on informal learning. If you are already looking at CS applications in the analysis of social media or informal learning, I urge you to look closely at my experimental design for novel approaches to addressing common challenges. If my results can provide grounding for your analysis, so much the better. If you are not yet considering applications in this area, I hope this work offers you the inspiration to begin.

To my fellow learning scientists, I present my work as a new approach to an old question. "How do we learn?" is one of the foundational topics of our field. I look specifically at learning informal language—dialect—in social media communities. If you are already familiar with this body of research, I think you may find my results directly useful, and I hope I demonstrate the value of incorporating computational methods into work in this area. If you are not yet engaged with this subfield, I hope again to offer inspiration. Perhaps you may find in my methods and results elements that apply to work in more traditional parts of the learning sciences.

To my colleagues in the computational sociolinguistics community, I present my work as a continuation of our collective project. I think most of what I write here will be to some degree familiar. For my chief contributions to the community, I call your attention to my use of ordinal time, and to how this interacts with my use of learning and learning sciences as a lens for considering linguistic changes.

1.1 Research Questions

Central motivating research question: How do novices learn to speak the informal languages of their online communities?

While directly answering this central motivating question would encompass a lifetime of work and is as such outside the scope of this dissertation, I include it here as a reminder of the goal that I am working toward. I address a series of related sub-questions, each designed to directly expand our collective scope of knowledge in pursuit of an understanding of this central question.

In my first classification study, I demonstrate that the social media communities I examine are valid domains for examining language change. I show detectable differences in conversational word choice are present, cannot be purely attributed to topics

of discussion, and appear between "novice" and "veteran" members of individual communities.

> **Research question 1:** Is vocabulary used within different Subreddits sufficiently distinct for a simple classifier to differentiate?
>
> **Research question 2:** Do detectable differences in vocabulary arise from differences in informal language use, or from the dominant topic of discussion?
>
> **Research question 3:** Do detectable differences in vocabulary appear between novice and veteran posters within a single Subreddit?

In my second study, I consider individual trajectories from the narrow lens of vocabulary size. Through an exploratory analysis, I reveal a common pattern and two related sub-patterns that describe novice behavior in using community jargon. I further investigate and provide evidence that novices engage in an authentic learning process, rather than novices being weeded out on the basis of their initial fit with the community.

> **Research question 4:** As a novice becomes an experienced member of the community, does their generated vocabulary size for posts within the community grow?
>
> **Research question 5:** Does the generated vocabulary size of a user's initial posts within a community predict their longevity?

In my third study, I again consider individual trajectories of word choice over time. In this study, I measure statistical divergence between a prototypical veteran word choice distribution and both novice and veteran posts. I find two conflicting patterns as novices gain experience within their communities: convergence toward the

veteran distribution in a low-jargon setting, and divergence away from the veteran distribution in a high-jargon setting.

> **Research question 6:** How do novices' word choices in posts change over time with respect to a prototypical distribution of veterans' word choices?

1.2 Positionality Statement

These studies investigate phenomena in human society, and so I begin with a reflexive analysis of myself in relation to this work. Following the guidelines set out by Holmes, I primarily examine my positionality as an *insider* or *outsider* with respect to "(1) the subject under investigation, (2) the research participants, and (3) the research context and process" [26, 28].

This work sits in an interdisciplinary space between computer science, learning science, social science, and linguistics. Each of these fields brings its own history of racial bias in research, with Black, Indigenous, women, and other marginalized scholars' viewpoints systematically devalued. As a White man, I benefit from a system that has boosted perspectives of people like myself. In mixed-methods work, it is important to acknowledge this privilege and to directly engage with racialized and other "-ized" issues related to the study. Color-blindness risks reinforcing the dominant narrative and perspective to the detriment of views expressed by minoritized scholars[37]. While usage rates of social media platforms in general appears to be largely similar between American men and women, and between Black, White, and Hispanic Americans as of 2015, the same survey notes distinct differences by level of formal education, by age, and by socio-economic status[46]. Moreover, similar rates of use do not correlate to similar *experiences*. This variation can take the form of differences in activity, as in a 2014 study of college student Facebook use which

found that women students used Facebook more frequently than their men student counterparts, making changes such as updating their profile more often[36]. It can also take the form of differences in passive quality of experience, such as the varying experience of witnessing microaggressions present in text. There has been recent progress in identifying microaggressions in social media which suggests that they are present and unevenly distributed, and thus could be a significant factor in some domains[6]. Most users of Reddit and StackOverflow do so anonymously, and so I offer a caveat to my studies: This work examines learning and language processes experienced by largely anonymous members of social media communities. I do not have the ability to breach their anonymity, and I choose not to attempt to infer demographics from their behavior on the grounds that inferred demographics using the techniques available to me would be worse than my current lack of demographic data. Rather than highlighting the experiences and views of minoritized community members, inferred demographics risk highlighting stereotypes associated with them. I believe that it is valid and necessary to include communities with anonymous members as subjects of study, and I will emphasize within my discussion that patterns visible within the quantitative results of this work can only be patterns inherent to the dominant group(s) within a given community and do not represent patterns that are universal to all community members.

I consider users of social media websites (variously indicated as "posters" or "community members" of different types) as human individuals. I myself am a regular member of some Reddit communities. I have taken pains to avoid analyzing Reddit communities with which I am already involved, as well as adjacent communities, but I have gained some familiarity with each of the communities I analyze over the course of my research. I nevertheless remain an outsider, drawing on a combination of internal statements by community leaders (e.g. forum moderators) and statistical

analysis for any judgments I make about qualities of these communities. Regarding my secondary dataset on StackOverflow, I have written several StackOverflow questions and am a frequent reader of existing questions; here I would position myself as a *tourist* according to the terms I define in my studies. I do not believe I hold significant understanding from personal experience of how StackOverflow works, despite a large number of passing interactions with the community. As in the case of the Reddit communities, I conduct my analysis of StackOverflow from this outsider perspective. I make a number of simplifying assumptions about the identities of individual posters within both Reddit and StackOverflow communities. I make these assumptions in full knowledge of exceptions that exist; I myself have multiple separate accounts on Reddit and have directly interacted with robot users, including users that I learned were automated after the fact. Similarly, I am very aware of how our social media personas do not always match our offline personas. I justify each simplifying assumption in detail in Chapter 3.

I began my study of learning within social media as a proxy for learning in other domains. Over time, I have also come to study it for its own sake as the centrality of social media in many of our lives continues to grow. I view learning and education generally through the lens of identity-building, as a collaborative process involving both the individual learner as an individual and their surrounding community. In my studies of language production, I am looking for evidence of learning as an intermediate step toward the goal of analyzing characteristics of informal learning processes. I try to maintain an awareness throughout my work that "learning is happening" is not a guarantee within my social media communities. Research questions 1-3 and 5 are each designed in part to address this concern by specifically investigating possible alternatives to a novice language learning process.

1.3 Learning Framework

I take a socio-cultural approach to learning, drawing on Vygotsky's sociocultural theory of learning and development in which the learning process is negotiated between the learner and their environment[14]. Social context is critical, as it constitutes all of the scaffolding supporting the learning process.

Outside the classroom, learning is at once individual—we can consider a singular learner and reason about their experiences—and collective, arising from community practices and the individual's experiences within them. I draw heavily on Lave and Wenger's notion of *legitimate peripheral participation* as the cornerstone of this learning process[35]. In legitimate peripheral participation, the learner engages authentically with the shared purpose of the community, interacting with and receiving feedback from established core community members. Learning is the process of moving through the boundaries of the community from outside it to within it, intertwined inextricably with changes to the learner's identity. I apply the framework of a community of practice to my social media communities.

Wenger draws a distinction between learning as *belonging* (a purely community-focused approach) and learning as *becoming* (an individualistic approach)[60]. Although Wenger acknowledges these as separate categories, he points out that they are deeply linked. I refer to both concepts together, and I highlight that an individual's relationship with their communities is negotiated between the individual and members of their community. It is dynamic and multiple, as an individual may become a core member of several communities at once or drift away from any or all of them.

I do not claim that these communities are communities of practice in the specific way that term has been used historically; rather, I claim that it is a useful frame

to consider. In a community of practice, legitimate peripheral participation involves authentic engagement with a shared project or practice. It involves action that is recognizable by other members of the community. A novice's reward for their learning is to be accorded, informally and by the community itself, identity as a member of the community's in-group. From this identity they then derive some authority within the community itself. In the context of social media, the shared project is sometimes nebulous. Each Subreddit within the overall Reddit community holds a shared topic of discussion, but whether this constitutes a shared project or practice in the sense of a community of practice is very much up for debate. The StackOverflow community, with its explicit question and answer format, arguably sets a shared practice of asking and answering questions, but again whether this truly constitutes a shared project or practice is debatable. While we might reasonably disagree about whether they count as communities of practice, I use the traditional definition of legitimate peripheral participation to reinforce my decision to focus on posting and commenting in my social media communities. If we accept participation in the shared discussion of relevant topics as the shared community project, posting and commenting both constitute unambiguous participation in that project.

This approach reinforces Papert's notion of students learning *to do* rather than learning *about*[44]. Papert discussed this in the context of mathematics—learning to do math and be mathematicians versus learning about mathematics. In my case, the novice learns to do discourse on the community's relevant topic. There may be authentic practices related to a social media community in addition to discourse. For example, the *use* of productivity tips is likely a shared activity of the productivity-tip-sharing community r/productivity, but examination of language production within that Subreddit can give at best indirect evidence of the offline activities of community members. In each community, I look only at the discourse, without access to any other

relevant practices. As such, all of the content I consider is based in text, and in fact all interaction within a Subreddit consists of sharing text, sharing content in the form of links, videos, or images, and reading or voting up or down said content.

I approach the mechanics of language use change itself using Bakhtin's theory of appropriation[2]. Novices engage with their community through observation of the discourse around them, appropriation of statements and ideas, and ultimately by remixing to produce novel language. I look for evidence of appropriation within online informal learning spaces. While there is rarely explicit teaching going on, conversation between novices and veterans provides a feedback loop that may enable the same or similar learning processes to occur as if explicit teaching is present. If they occur, they should be observable in my data. Discourse does not stop at word choice or even styles of speech. Gee describes discourses as "ways of being in the world", a holistic notion of the interaction between the individual and the community[11]. I concur with this definition, and emphasize that my investigation of vocabulary acts as a starting point, not an end point. To see the appropriation of discourse in action, we should be able to see appropriation of vocabulary as a visible symptom.

1.4 Metrics

In this section I outline the quantitative and computational tools that I employ within my following studies. Each method is explained in detail within the appropriate chapter(s); this section presents a general overview of my methods.

1.4.1 Natural language processing

We can analyze how people produce language by looking at how they produced language in the past, and use natural language processing (NLP) tools to build statistical

predictive models around patterns within that language production. I am not referring to models that can produce novel instances of language, although there are NLP projects that aim to do this as well. Rather, we can develop the ability to recognize patterns by identifying the key subsets of those patterns ("features") that combine to best predict the relevant language use. What is relevant varies from study to study; in this work, I am generally interested in patterns of language use that are characteristic of different populations. That is, can I identify features that help me distinguish between speech by novice members of a community and veteran members of that community? Can I identify features that help me distinguish between speech by members of two different communities? Sometimes, although not always, these are features that bear human-comprehensible descriptions.

I subdivide an individual's language production into a series of *utterances*, equivalently their series of posts and comments in a social media forum. The notion of an utterance is familiar in the qualitative learning sciences community and in the linguistics community. In both settings, an utterance is a complete thought, often viewed as an 'object' constructed from language and handed to the interlocutor. Half of an utterance is unlikely to make sense, but a person may make several utterances back-to-back. While in spoken conversation individuals' utterances may overlap, the asynchronous communication pattern of an online forum cleanly separates each individual's post or comment from replies to it, and while it is common in an active community for several community members to reply to the same utterance at the same time, forum metadata cleanly separate these parallel statements and link them to the appropriate prior post or comment.

I further subdivide each post or comment into its vocabulary *tokens*. In NLP literature generally, a token is defined as the basic building block of text with respect to the study or tool at hand. The token marks the smallest level of granularity that

will be considered, and as such is different for various tools and research projects. In my work, the relevant level of granularity is that of the individual word. Tokenization then decomposes an utterance into its component words, each of which is considered immutable. For completeness, note that while I will often use *token* and *word* as synonyms, the latter is not actually formally defined. Contractions, hyphenated or compound words, and similar edge cases complicate the picture. Tokenization tools have established answers for each question of what counts as a word-level token, and I opt to stick to these defaults.

So which tokens did the author use to make their utterance? Common structural tokens (such as "and", "the", etc) might appear everywhere, but variations in usage of these structural tokens could indicate a Subreddit community dialect, just as much as the appearance of unusual terms. I conduct analysis without removing common tokens, with the exception of several experiments where I deliberately focus on "topic-relevant" vocabulary which naturally excludes common structural tokens along with many other terms. I explicitly note in my methods where this occurs.

I use vocabulary as my primary lens into individuals' and communities' language use patterns. In Chapter 2, I explain the theoretical justification for this approach. Briefly summarized, vocabulary is one plausible angle for investigation, and it is particularly compatible with the quantity of data in my studies. My work follows individual language production over time at scale, with dozens of utterances from each individual and thousands across each community that I examine. As such, I require methods to rapidly describe and compare the language produced by different individuals and in different communities. These descriptions and comparisons are straightforward to conduct through the lens of vocabulary. More computationally intensive methods (e.g. word embeddings or parsing) and more human-intensive

qualitative methods (e.g. discourse analysis) are valuable but largely out of scope for my present studies.

1.4.2 Classification

When presented with a sample containing intermingled elements from multiple distinct populations, we can examine the key differences between those populations through neural network classification. As long as we can identify a set of possible features, as long as a distinction between the populations exists, and as long as the setting is not designed to confound the neural net, the neural net can with some likelihood discover a set of features which together distinguish between them. The neural network's effectiveness will vary based on the scenario; in particular populations that are more distinct from each other will be naturally easier to separate. This approach shines when addressing settings with a large number of possible features, and when it is unknown which subset of those features might be relevant. Both characteristics apply to my investigation of language change through vocabulary, as each word/token is a possibly relevant feature.

The specific populations of interest vary between my studies, but I generally look at veterans, novices, and tourists within a given social media community. These populations are distinct, although the boundaries between them can be blurry. For example, a new active member of a community begins as a novice and eventually becomes a veteran, but there is no single moment that encompasses the entire transition from one to the other. Similarly, while tourists are characterized by never sticking around long enough to achieve veterancy, there are tourists who came close to doing so and whose language patterns may well exhibit some degree of influence from the community. These possible imperfections in the data actually provide another reason

to apply classification: like most machine learning (ML) tools, a neural net seamlessly handles outliers in messy input data. A classifier is judged by its ability to correctly predict an unlabeled sampling of inputs, and the presence of occasional novices that look like veterans, veterans that look like tourists, etc simply result in slightly lower rates of correct predictions.

I use classification as a tool for demonstrating the separability of informal languages while minimizing my prior assumptions about the features that distinguish them. With high noise levels in my data, I am wary of repeatedly hypothesizing possible sets of features. Qualitative discourse analysis research suggests a variety of features that might indicate learning processes. Unfortunately, a basic law of statistics states that if I try enough of them I will, through chance, find one that seems to work—even if in reality none work. A classifier avoids this issue, although my approach does also mean that the sets of distinguishing features I find are not readily human-comprehensible.

1.4.3 Ordinal time

In this work, I examine the language production by members of several online communities and how their writing habits change over time. Since these social media communications do not occur at regular intervals, and are entirely asynchronous, it is important to address how I think about time in these spaces.

Several prior studies have addressed time in social media by subdividing their data into chunks one month in duration[17, 33]. They can then consider changes in language production between these month-long segments. These are examples of a real time framing of language change in social media. Real time framing is a powerful approach, but requires large quantities of data. Kershaw et al's study selected one

month as the size of their intervals because they could not find sufficiently many posts to conduct statistical analysis with any shorter time frame, and they studied language change across an entire population without regard to the language production of individual members of that population. Moreover, while language change and learning both undoubtedly occur over time, and we are used to thinking about steady learning processes through the lens of regularly-occurring classes in offline settings, there is no reason to believe that learning through social media occurs at a steady rate. A month-long gap between a user's post and their next post does not represent one month of learning time. It may represent a month during which they read the forum religiously and wrote many drafts, but it is just as likely to represent a month during which the user didn't think about social media at all. A real time framing does not account for this variation.

In contrast, an ordinal framing of time considers language change through the prototypical "post-feedback-post" cycle:

- User makes a post.
- They receive feedback, primarily in the form of replies, upvotes, and downvotes.
- User incorporates the feedback (consciously or not) and makes their next post.

The ordinal framing carries its own danger: namely that a user who makes many posts in very quick succession likely is not fully engaged in incorporating feedback into each post. They simply have not had enough time to do so. Notice also that a "lurker" (defined by Preece et al as "someone who has never posted in the community to which he/she belongs"[49]) might be a long-time observer of their community incorporating feedback through watching reactions to other posters, but would not be credited with receiving this feedback under the ordinal framing.

Both framings approximate the process of language change. Since I am directly interested in analyzing changes in the *language production* of individual novices as a

potential learning process, I opt to use the ordinal framing of time in my work. If I assume that a relatively consistent quantity of learning activity is required to master the norms of a community, I would then like to be able to approximate this with an amount of time a novice has spent within the community learning. A real time framing offers very high variance: Novices who have spent a month within the community may have truly spent anywhere from a month down to just a minute mentally engaged with that community. The "post-feedback-post" cycle of the ordinal framing, on the other hand, relates well to the concept of a novice interacting with a tutor. A novice who has made 10 posts has engaged mentally with the community 10 times and received feedback 10 times, even if we cannot directly determine to what degree the novice has incorporated that feedback into their writing. Both framings offer an approximation. However, I believe that of the two, the ordinal framing offers a closer approximation to the hidden true learning process of a novice.

1.5 Chapter overviews

The remaining sections of this document are arranged as follows: In Chapter 2, I outline relevant literature from the Computational Sociolinguistics community and surrounding relevant disciplines. I continue in Chapter 3 with an overview of my primary Reddit dataset and its relevant characteristics. In Chapter 4, I present the methods and results of my classification studies. In Chapter 5, I present the methods and results of my computational analysis of individuals' language production trajectories. I conclude in Chapter 6 with a synthesis of the overall results so far along with my thoughts on promising future directions for my analysis.

Chapter 2 reviews work primarily from Computational Sociolinguistics, a term given to an interdisciplinary body of work on the borders between computer science,

sociology, and linguistics. Its chief unifying feature is the treatment of linguistic artifacts (e.g. social media text records) through computational methods toward a greater understanding of language as an artifact of society. Much existing work in this emerging field looks at how elements of our identities (e.g. demographics) inform the way we communicate online. I situate my notions of novicehood and veterancy within a community and language as an artifact of learning in this context.

Chapter 3 provides context on the Reddit dataset that I use in my studies. I discuss its terminology, as well as characteristics of its subforums that impact my methods and their application.

Chapter 4 covers my classification studies, addressing RQs 1-3. I demonstrate the variation of language use between Subreddit communities and within individual communities between novice and veteran members.

Chapter 5 contains several studies examining individuals' trajectories of language use, and addresses RQs 4-6. These studies are exploratory in nature, and I construct a theory of independent appropriation and winnowing learning phases that appears to explain my results.

Chapter 6 concludes with a synthesis of my results and observations, and proposes several future avenues for investigation.

1.6 Definitions

In this section, I present for reference definitions that I apply throughout my work. I subdivide the definitions into general definitions, terms that bear specific meaning in the context of Reddit and StackOverflow, and terms referring to categories of community members.

1.6.1 General definitions

Definition (ROC-AUC)**.** Given a sample set and a test with variable sensitivity, I plot the true positive rate (recall) against the false positive rate (1 - specificity). This is the ROC curve, and ROC-AUC is defined as the area under the curve.

Definition (Bag of words model)**.** A post consists of an unordered set of tokens, with a token present in the set indicating that one or more copies of the token could be found in the full text of the post.

Definition (True Positive Rate (Recall))**.** The fraction of true positives divided by the sum of true positives and false negatives. This is sometimes also referred to as the Sensitivity. For clarity, I use "Recall" to refer to the True Positive Rate and "Sensitivity" to refer to the variable discriminatory setting of my classifier.

Definition (False Positive Rate)**.** The fraction of false positives divided by the sum of true negatives and false positives. This is equivalent to 1− Specificity.

Definition (Vocabulary)**.** A probability distribution which, given a specific Reddit user (or set of users), Subreddit, and token, returns the probability that a new post from the user(s) in the given Subreddit will contain the given token. The size of a vocabulary is the number of such tokens with non-zero probabilities.

1.6.2 Reddit and StackOverflow definitions

Definition (Age)**.** The *ordinal* position of a post within its author's corpus. i.e. a poster's age 1 post is their first post, their age 2 post is their second post, and so on. Also may refer to the poster themself, e.g. "at age 20, a poster is now considered a veteran".

Definition (Community member)**.** In Reddit, a user, indicated by their username, who produces language within a particular given Subreddit. The language produced may be either in the form of posts or comments, and includes "language" production consisting of links to images.

In StackOverflow, the same, but the user produces language within the StackOverflow Q&A forum either by asking questions, answering questions, or commenting.

Definition (Post)**.** An entry within a social media forum that starts one or more *threads*. Most top-level views, e.g. the front page of a Subreddit, only show the titles and a limited preview of posts. A **comment** is a response to a post or to another comment, generally only viewable through viewing the specific post at the root of the comment's thread. In StackOverflow, an **answer** is a specific type of comment that directly attempts to answer the question posed in the post and tends to look different from other comments. The act of *posting* can refer to creating posts or comments, and throughout my studies I refer to a **post** in this more general way as a catch-all for text content produced within a community.

Definition (Poster/User)**.** A person within a community. While both Reddit and StackOverflow content can be freely read without formal membership, posting requires logging into an account with a unique *username*. Within both Reddit and StackOverflow metadata, a post's *author* field(s) refer to the user who created the post. Note that a single poster may be active within multiple communities.

Definition (Subreddit)**.** A forum within the Reddit website. Each Subreddit follows the naming scheme "r/name", corresponding to the URL that a reader would visit to view the latest top posts within the community.

Definition (Timestamp)**.** An instant in real time, measured in seconds after January 1, 1970. Also known as "Unix Time". Reddit and StackOverflow metadata includes a Timestamp.

1.6.3 Community member categories

Notice that an individual poster will be a novice and a veteran at different points in their tenure, and will create novice and veteran posts to match.

Definition (Veteran)**.** A community member who has produced a significant number of posts within the community. A post is considered to be a veteran post if it was made by a veteran community member *and they had achieved veterancy by or before making the post*. A veteran might be considered by virtue of experience to be an authority on the community.

Definition (Novice)**.** A community member who has not yet produced enough posts to be considered a veteran, but does eventually become a veteran. A post is a novice post if it was made by a novice community member. A novice is (possibly unintentionally) undergoing a process of joining the community and may be learning its linguistic norms.

Definition (Tourist)**.** A poster who never achieves veterancy (within my dataset) is a **tourist**. A post they create is called a tourist post.

Chapter 2

Natural Language Processing and Online Communities

In this chapter, I outline prior work that informs my studies. The main body of related research comes from the computational sociolinguistics community, which consists of computer scientists, sociologists, and linguists united through our interest in computational and statistical investigation of elements of and effects of language online.

2.1 Who am I? Identity through language in online spaces

Computational sociolinguistics includes a significant body of work which blurs the lines between online personas and offline identities. In their survey of the field, Nguyen et al describe this subset of work as analysis of online language production to "automatically infer social variables from text" [39]. This generally involves predicting demographic information, but may also include other elements of offline

identities such as mapping relationship networks[19]. This kind of study offers a statistical technique which can bypass a common problem with online data: internet users don't often self-report demographic data, and when they do they sometimes lie. When discussing less clear-cut demographics such as measures of social status or experience, users may not know or may accidentally mis-identify. User profiling studies tend to focus on predicting the types of demographic information that we see used in academic and non-academic analysis everywhere. Gender[1, 8, 12, 15, 22, 48, 56, 57], age[1, 9, 40], geographic location[20], social status[5, 16, 23, 47], ethnicity, and political affiliation all appear within the corpus.

I extend this body of literature by considering language production through the lens of veterancy within an online social media community, with a focus on novices' conversational language patterns. Recall that I define a veteran in an online community loosely as someone whom a member of the community might consider by virtue of experience to be an authority on the community. Although the term is new, the concept of veterancy is not a totally novel demographic in this field. Several prior studies consider membership duration and posting frequency. Nguyen and Rosé examine features of language production and their ability to predict membership duration in a breast cancer forum, and draw the same connection that I make between the language use patterns of long-duration members and community norms[41]. Fields et al consider length and frequency of membership along with gender in their analysis of the programming patterns in the Scratch online community, with an eye toward the use of specific programming patterns among "Newbie", "Young", "One-year", and "Oldie" accounts[21]. Unlike most demographics considered in the user profiling literature, veterancy is a property of a person in conjunction with a specific online community. Veterancy is similar to expertise, examined in the Scratch online community by Huang and Peppler[30], but whereas an expert is presumed to be objectively

better at some external task, veterancy implies a subjective degree of understanding and power in relation to the community itself.

Prior work includes both papers which conduct demographic prediction as an end in itself, and papers for which it is a tool to assist with other analysis. For example, Prabhakaran et al[48] examine how gender impacted displays of power through language, Dadvar et al[15] use automated detection of gender to select the appropriate gender-specific cyberbullying classifier, and Im et al[31] uses NLP techniques generally to identify Russian Twitter trolls. In contrast, work by Sarawgi et al[57], Gianfortoni et al[22], and Sap et al[56] focuses on improving the technical task of gender prediction trained on one platform and tested on another. In some cases, as with analyses of gender by Burger et al[8] and of age by Nguyen et al[40], computational analysis is specifically contrasted against manual analysis to demonstrate that the new tools are better at predicting the demographic in question.

I follow both arguments, first verifying that my tools do in fact distinguish veteran vocabularies from those of different types of non-veteran community members, and then using those tools to make narrative sense of the differences I observe.

Work in this area acknowledges that demographics are often intertwined with each other and with other elements of conversation. Linguistic features which can be used to distinguish gender and age, for example, overlap[1]. Features that appear to indicate a community member's gender may in fact indicate the topic of discussion[57]. Researchers estimating the relative power of speakers through measuring linguistic coordination may find that they are actually measuring in part the gender of the speakers[16]. It is well understood in Sociology research that perceptions of authority (who has it, how much) are influenced by demographics and particularly influenced by race and gender[58].

In addition to linguistic markers of demographics, other social dynamics such as social roles can be a confounding factor. Zhang et al show how different elements of linguistic variation can predict future anti-social behavior in an online space[62]. They find that analysis of posts in Wikipedia talk page discussions, in bag of words form, yielded 56.7% accuracy at predicting conversational failure. This rate, significantly better than chance, suggests that some detectable portion of Reddit posts may be part of attempts to troll rather than attempts to sincerely engage within a community and learn its norms.

Although I am wary of predicting demographics and then conducting veterancy analysis on top of that, I discuss the potential for overlap between community recognition of authority, its linguistic markers, and member demographics in my results. I address the effect of topics of discussion directly, using the presence of independent Subreddit communities that substantially share topics to demonstrate the "topic effect" on my data and its limits. Regarding trolling and related anti-social behavior, I limit most of my studies to community members with sustained posting behavior (i.e. they achieve veterancy). This is unlikely to weed out all trolling behavior, however given the high level of overall noise within my data I accept some degree of insincere interactions by novices and veterans as part of that noise.

2.2 Who are we? Community language dynamics in online spaces

A second corpus of work within computational sociolinguistics considers the dynamics of language within communities. This includes the dynamics of individuals interacting with their community and dynamics of the community without regard to individuals.

It is comparatively much smaller than the user profiling corpus. I hope through my work to expand it.

Work in this area extends offline community analysis. The corpus draws significant inspiration from Lave and Wenger's communities of practice[35]. Although the lack of a shared practice or project in the traditional sense removes the formal concept from consideration, Nguyen et al note in their survey of computational sociolinguistics that "core" and "peripheral" community membership play significant roles when examining online community dynamics[39]. Legitimate peripheral participation in a community of practice may likewise be analogous to well-known social media behaviors such as lurking.

Earlier work on online communities, such as that by Cassell and Tversky[10], looks at non-public communities in educational settings. More recent work has moved to more free-flowing community settings such as the online forums I work with. These communities' language production is examined through the lens of jargon and slang by Nguyen and Rosé[41], as well as by Danescu-Niculescu-Mizil et al[17] and by Hemphill and Otterbacher with a specific focus on gender[27].

There is even less extant work on changes in community language production over time, with only three papers identified. Danescu-Niculescu-Mizil et al[17] break their data into month-long chunks for analysis, allowing them to consider not only how users adapted their language from month to month, but also how the characteristics of the community's language over all change. They consider conceptual distances between individual language use and community norms, but while they account for how norms may change over time, they do not address how different users' posts may variably contribute to what members of the community consider to be their language norms. That is, they measure cross-entropy between the bag of words form of a user's post and the distribution of bag of words forms of all users' posts for that

month. Kershaw et al[33] examine the creation and spread of new vocabulary within Twitter and Reddit, also using one month time periods. These very large groupings of posts by time provide more data for statistical analysis—Kershaw points out that they initially attempted one week time periods and found they had insufficient data density—but make it impossible to conduct analysis on the trajectory of language for individual posters. Hua et al consider editing behavior in Wikipedia talk page conversations, although their focus is on changes to existing text, rather than the evolution of text production[29].

Unlike prior studies, I examine individual Reddit users' evolutions in language production over time, with a deliberate focus on the natural social media cycle of making a post, receiving replies (feedback), and then making a new post. I intentionally abstract away "real" time to look at the dynamics of individuals interacting with their communities in spaces designed for asynchronous conversation.

Chapter 3

Understanding Reddit Data

In this chapter, I provide context on the social media website Reddit, from which I drew the data for my study, and how these settings influenced my study design. The website remains active as of June 2021, and can be found at www.reddit.com. My Reddit data comes from a dataset of English-language posts from the site's founding in 2007 through May 31, 2015. More recent versions of this dataset can be acquired from the Reddit community r/datasets. I have chosen not to upgrade to a more recent dataset version as the 2015 dataset already includes sufficient quantities of data for my study, and I do not expect any effect on my research questions from adding 2016-2020 data.

All text within my dataset is publicly available. While some Reddit users are "verified", linking their usernames to offline identities, I avoid analysis of forums within my data that cover sensitive topics and avoid presenting results that could be tied to individual real humans.

Although it is difficult to rigorously measure social media use due to the difficulty of measuring reading activities and the difficulty of getting reliable answers in user surveys, Table 3.1 shows one 2020 pop-scientific survey of active user counts, which

Network	Active users (millions)
Facebook	2,701
YouTube	2,000
WhatsApp	2,000
Messenger	1,300
WeChat	1,203
Instagram	1,082
TikTok	800
QQ	691
Sina Weibo	550
Qzone	517
Reddit	430
Kuaishou	400
Snapchat	397
Pinterest	367
Twitter	326

Table 3.1: The top fifteen social media websites by active users in 2020[18]. Shaded rows are primarily non-English in content.

offers a rough approximation of which social media sites are the most significant. Reddit sits high on the list of English-language social media websites, and is the largest such platform with clearly delineable sub-community boundaries. Within the United States, "nearly a quarter of U.S. young adults" self-reported using Reddit in 2019[32].

3.1 Organization of Reddit Data

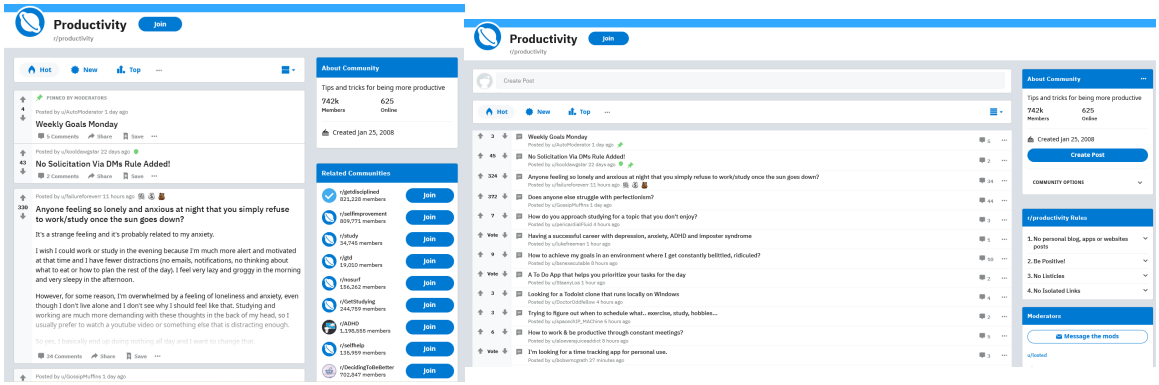
I use Reddit as my primary dataset due to its unusual internal structure, which allows me to analyze community membership much more cleanly than is possible in other social media. Many social media sites are organized around this notion of a *friendship* or its parasocial counterpart, the *follow*. In these communities, tracking community membership itself is a serious challenge, and behavioral dynamics involved in joining

and leaving communities are completely obscured. Reddit, by contrast, supports a very simple way to think about the boundaries of communities: the Subreddit.

3.1.1 The Subreddit

Reddit consists entirely of various user-created discussion forums called Subreddits. Within those Subreddits, users read posts, write posts, and respond to the posts of others. Responses can take the form of upvotes and downvotes that are visible in aggregate, or the form of comments which are visible individually—and may be voted on in turn. Reddit users can subscribe to different Subreddits, and each Subreddit is essentially permanent. Each user sees an individualized front page when they visit the website. While they may skip directly to a page showing recent posts for a specific community, there is some evidence to suggest that viewing the personalized front page may be the dominant way in which users begin each interaction with their Subreddit communities[24]. This presents a possible threat to the validity of my analysis, since it calls into question the degree to which Subreddits are experienced as separate communities by their members. However, Buntain and Golbeck find that it is rare (3%) for a Reddit user to be a "frequent participant" in more than one Subreddit[7]. Thus, the personalized front page of a Reddit user active in one Subreddit likely closely resembles the community-specific front page of that Subreddit. Community-specific front pages can be accessed directly by link within the Reddit website, or directly by URL (e.g. www.reddit.com/r/productivity as shown in Figure 3.1).

Because Subreddits are user-created, with a very low barrier to creation, I conduct analysis on specific Subreddits, chosen to highlight factors that may affect learning patterns while remaining computationally tractable and theoretically interesting. There is no particular organization or pattern to Subreddit communities or their



(a) (b)

Figure 3.1: The r/productivity Subreddit front page as seen in the (a) standard and (b) compact views, screenshots taken May 25, 2021.

names. Some are catch-all topics of discussion, such as r/politics with its 5.7 million subscribed members as of January 2020. Large, high profile Subreddits such as this one offer at most a diffuse internal community, with community norms dominated by the norms of affinity groups that operate as distinct communities within the space. Other communities have a very specific purpose, such as r/SubredditSimulator with 413,000 subscribers and the odd distinction of being a community populated only by robots. A Subreddit such as this with strict norms limiting member speech naturally also makes for a poor setting for my analysis.

Both of these communities are quite large by Reddit standards. The majority of Subreddits have only one or two posts. While it is useful that Reddit does not provide tools for deleting Subreddits, and so these small Subreddits persist, I focus on Subreddits of intermediate size (1,000-300,000 posts) that have been active for significant periods of time. I introduce my three primary Subreddit foci in Section 3.1.5 below.

3.1.2 Community membership models: subscribers vs readers vs posters vs responders

In order to discuss the language production patterns of members of social media communities, it is necessary not only to have a notion of what a community is but also a theory for what constitutes a community member. Reddit allows users to subscribe to Subreddit communities, which appears at first glance to be what we're looking for. Unfortunately, the reality is not so simple. While a Subreddit itself is permanent, a subscription is not. Community membership can change as rapidly as a Reddit user wants it to, and no one records which communities a user joins or leaves. Moreover, it is not at all clear that subscription accurately captures community membership. Reddit users do not need to subscribe to a community to either read its content, or add content of their own. A Subreddit's subscribers thus may not accurately reflect the users that actually form its community of practice.

Setting aside the idea of modeling community membership through self-identified subscriptions, we can also model membership through concrete actions. In doing so, I follow in the footsteps of the largest body of computational sociolinguistics work, where individuals' characteristics are inferred from visible online activity. A Reddit user engages in four distinct types of activity within a Subreddit community: They read, they vote, they write, and they comment.

The first and second types of action are unfortunately invisible. Reddit users do not even need to log in in order to read in the website, which helps lower the barriers for new users discovering content, but also makes it impossible for a researcher to track who is reading what. Similarly, while the total vote count for each post and comment is generally visible, these votes cannot be traced back to the individuals who cast the votes using my data.

Writing posts and commenting, in contrast, are highly visible activities. For every post or comment, my dataset includes a wealth of information. (See Appendix A.) Critically, the available information includes the author of the statement and a timestamp. I can identify all posts and comments made by a particular user within a particular Subreddit, and I can place the posts and comments in the order that they were made. As a result, in my studies when I refer to a member of a Subreddit I specifically refer to their membership in terms of post and comment creation. Recall my definition of a community member:

Definition (Community member)**.** In Reddit, a user, indicated by their username, who produces language within a particular given Subreddit. The language produced may be either in the form of posts or comments, and includes "language" production consisting of links to images.

It is a conservative definition, in that not all true community members can be detected. A lurker who reads every post but never comments cannot be seen this way. Nor can we identify a more active user who contributes votes, thus shaping through Reddit's core algorithms what content will be shown on the front page of that Subreddit, until they begin to produce content of their own.

3.1.3 Comments and posts

I focus on visible language production in the form of posts and comments. They are conceptually similar activities, both involving language production and the opportunity for feedback from the community. The most significant immediate differences between them are that a comment lacks a title and must be made in response to an existing post. A comment is therefore not a candidate to be shown on any user's front page, but by this clearly reduced visibility may be less taxing for a user to create

than a post. Examples of posts and comments can be found in Section 3.1.5. For my analysis, I viewed this difference as not significant enough to justify creating separate age measurements for posts and comments or otherwise separating the two types of language production, and I refer to both together as "posts" in my studies.

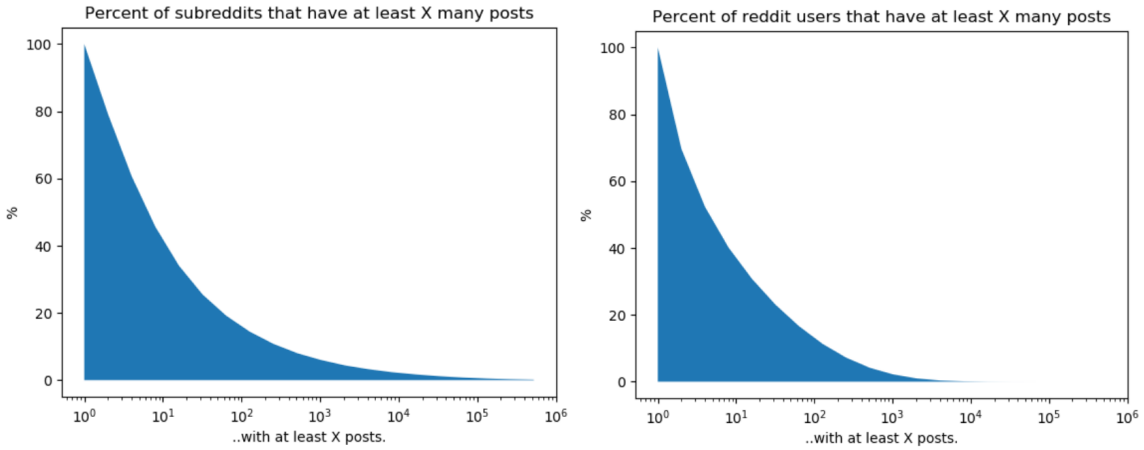
3.1.4 Geography of Reddit

The first guideline for thinking about Reddit as a dataset is to realize that *everything* is an exponential decay function.

Subreddit sizes

There is a size sweet spot for my analysis techniques. Too small, and I don't have enough data to work with to see any real pattern. Too large, and I run into technical difficulties with pre-processing which prevent me from completing the analysis. Recall that my dataset covers 2008-2015, and covers English-language Subreddits only, at least in theory. Manually skimming through the dataset suggests that a few non-English Subreddits may have slipped through. There are 224,621 Subreddits in total within my dataset. Mapping the Subreddits in terms of post and comment count led to the following results: The largest Subreddit of all was the r/AskReddit community with 154,501,316 posts and comments. As shown in Figure 3.2, the plurality, but not the majority, of Subreddits contained only a single post.

When considering the posting habits of Reddit users, note that my data is technically about the posting habits of Reddit *usernames*. A single human being might have multiple Reddit monikers, and some Reddit users may be accounts shared by multiple humans, corporate accounts that get passed from social media coordinator to social media coordinator, or even bot-controlled accounts that post automatically.



(a) (b)

Figure 3.2: Population statistics for all Subreddits (left) and all Reddit users (right). Note the log scales on the X axis.

Ignoring all of these complications, the distribution of post counts across Reddit users in Figure 3.2 shows a remarkably similar pattern to the one I observe regarding Subreddit sizes. Once again, the plurality of Reddit users only make a single post, while the top user AutoModerator boasts 3,677,676 posts.

Robot users

As the name suggests, AutoModerator is a bot-controlled user. How much are non-human users warping the language production of the Reddit dataset? At first glance, there is reason for some concern. Of the top ten Reddit users (Table 3.2), all but the 10th are robots (and the 10th is a deleted account that I can't verify either way).

Using the names of the top posters as a guide, I created a first-pass filter on usernames that divides Reddit users into "suspected bots" and "suspected humans" based on the presence of "auto", "bot", "log", "transcribe", or "poster" in their username. Out of 13,212,801 Reddit users (in my dataset), 13,135,241 registered as "suspected human" and only 77,560 came up as "suspected bot". Plotting the

Username	Post Count
AutoModerator	3,677,676
conspirobot	575,535
ModerationLog	547,604
autowikibot	402,076
PoliticBot	388,395
imgurtranscriber	360,244
dogetipbot	358,093
qkme_transcriber	301,844
TweetPoster	293,290
Late_Night_Grumbler	279,007

Table 3.2: The top ten Reddit users by post count.

Username	Post Count
AutoModerator	3,677,676
conspirobot	575,535
ModerationLog	547,604
autowikibot	402,076
PoliticBot	388,395
imgurtranscriber	360,244
dogetipbot	358,093
qkme_transcriber	301,844
TweetPoster	293,290
hit_bot	125,636

Table 3.3: The top ten suspected bot Reddit users by post count.

internal distributions of each population with respect to post count generated the surprising result that the distributions are virtually identical.

A closer look at the top ten of each category (table 3.3 and table 3.4) shows some false negatives, such as the bot MTGCardFetcher. Others, such as morbiusgreen and -rix, do appear to be human posters upon brief inspection of their posts.

The low volume of suspected bots relative to suspected humans and the identical distributions of postcounts suggests that my filter is not very precise. If human posters are much more common than bot posters, even a relatively small rate of false positives among suspected bots would result in the false positives dominating and

Username	Post Count
Late_Night_Grumbler	279,007
morbiusgreen	185,501
-rix	172,958
pixis-4950	171,898
Franciscouzo	170,706
UnluckyLuke	162,505
Lots42	154,504
PornOverlord	136,747
MTGCardFetcher	125,794
matts2	123,819

Table 3.4: The top ten suspected human Reddit users by post count.

	r/ArtFundamentals	r/productivity	r/vim
Total posts	3,308	30,501	68,601
Median post length	60	25	26
Mean post length	106.7	46.8	43.4
Post length st.dev.	135.8	69.8	61.9
Median post age	10	2	14
Mean post age	296.2	8.1	81.0
Post age st.dev.	429.5	17.8	215.7

Table 3.5: General post statistics for the r/ArtFundamentals, r/productivity, and r/vim Subreddits. Post length statistics are measured in tokens.

producing the results I see. On the other hand, Norlander estimates that roughly a quarter of Reddit posts are produced by bots [42]. They note, however, that almost all bots they detected produced either politically charged posts or low-effort posts. The first category is likely to produce an extremely unevenly distributed population of bot posts. Accordingly, I opt to treat robot users as regular users. As long as I restrict my analysis to low-profile Subreddits without a political focus, they should contribute a vanishingly small amount of noise to my analysis and not skew my results in a significant way.

3.1.5 Specific Subreddit foci

While I hope to shed light on patterns that apply across social media, even into offline informal learning settings, no analysis can be complete without consideration of the quirks of my particular sample space. Within the Reddit milieux, I conduct some degree of analysis on dozens of different Subreddits. However, there are three in particular through which I conduct the bulk of my studies.

I choose the first two foci, r/productivity and r/vim, as each is a clear affinity space built around something external to the Subreddit. They are thus among the Subreddits that can best make the case for having a "practice" in the sense of a Community of Practice. Each Subreddit functions as a forum where members of this community and newcomers may discuss, and there is in both cases clear direction from the community as to the main topic of conversation yet few restrictions on how that discussion may proceed. (Each Subreddit has a Rules sidebar, and the current rules for these Subreddits do not impact conversational language use. I observe no indications that they previously had more restrictive rules, but it is technically possible.) The lack of an adjacent external forum means that my data covers more-or-less the entire picture of the community's discourse. (Contrastingly, a Subreddit with such an adjacent external forum might see a significant amount of users learning in one forum and then speaking in the other, muddying my results.) Both r/productivity and r/vim are medium-sized Subreddits with a long history of activity within the time frame of my dataset, these factors making my computational approach feasible. In contrast to each-other, the r/productivity community uses a low degree of obvious community-specific jargon relative to the Reddit language background, while r/vim uses a high degree of jargon. Since relative jargon-heavy-ness is not a well-defined concept, I apply a computational measure for it in Section 5.3. This measure also

finds that r/productivity is a low-jargon community, and r/vim is a high-jargon one. In addition to being a high-jargon community, r/vim's focus on computer software may mean that its user demographics skew toward white men. I do not know enough about the r/productivity community to guess whether the same issue may be present. Since Reddit users are pseudonymous, I cannot confirm whether this is the case for either Subreddit, but it would be valuable to include in future studies communities focused on diverse topics known to contain significant populations from different backgrounds.

My third focus Subreddit, r/ArtFundamentals, is chosen because it is an intentional yet informal learning community. Like the first two, it is built around an external activity. Unlike them, the external activity is something that we can easily imagine fitting into a school curriculum—practicing art skills. It is much smaller than either r/productivity or r/vim, which is not ideal, but it is large enough for my classifier to function properly. One possibility for future work would be to use this community to help me identify other intentional learning communities within Reddit that have larger membership rolls, thus allowing me to better examine the effects of an informal community's intention toward learning on novices' language production. r/ArtFundamentals is not selected for contrasting jargon-heaviness, and rates moderately jargon-heavy according to my methods in Section 5.3.

Table 3.5 shows overall statistics for these Subreddits. Figures 3.3-3.7 show sample posts and comments drawn from the front page of each community. Note that Reddit's rating algorithm incorporates both user votes and recency, so posts on the front page in 2021 are not actually contained within my dataset but are rather representative of the type of content that can be found within the community. A reader looking at a specific post can view more comments by scrolling down or expanding individual comment threads. Not all comments are immediately visible. For each Subreddit,

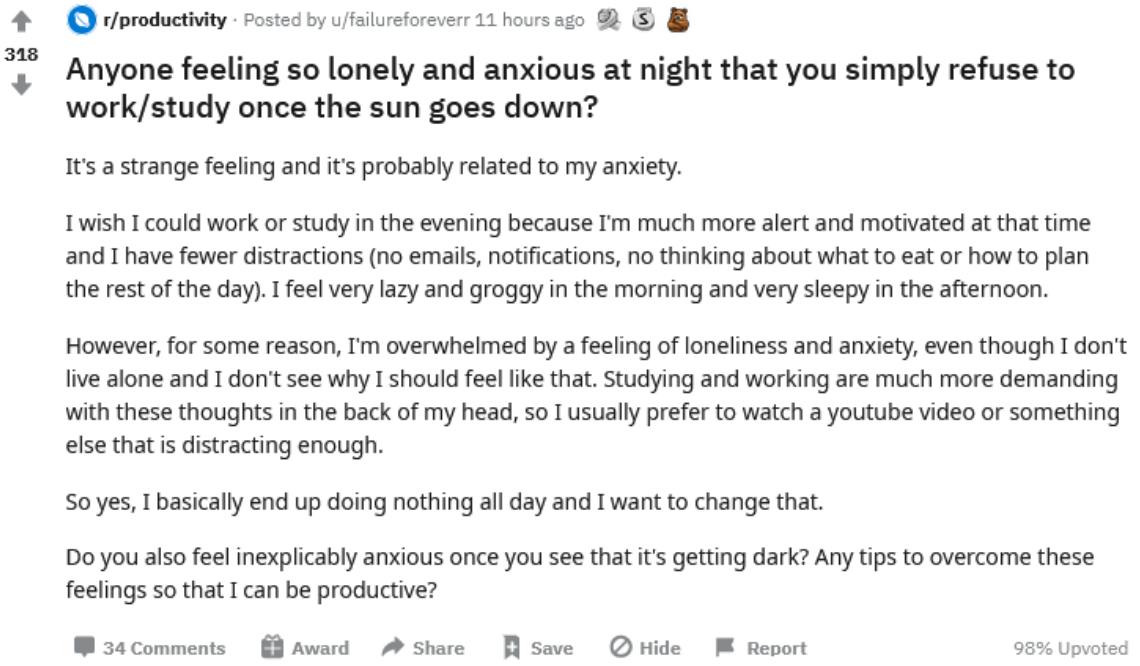


Figure 3.3: The top-rated post in the r/productivity Subreddit front page on May 25, 2021.

I also provide their self-described "About Community" statement and a brief text description of the community.

r/productivity

Tips and tricks for being more productive[54]

The **r/productivity** community consists of 31 thousand posts in my dataset and bills itself as a place to share productivity tips. It is best characterized by its lack of organized structure, and by its lack of obvious jargon. Informal survey of a sample of posts from the community revealed no terms that were recognizable as community-specific to a human outsider. This does not exclude the possibility of more subtle jargon or non-vocabulary dialect patterns.

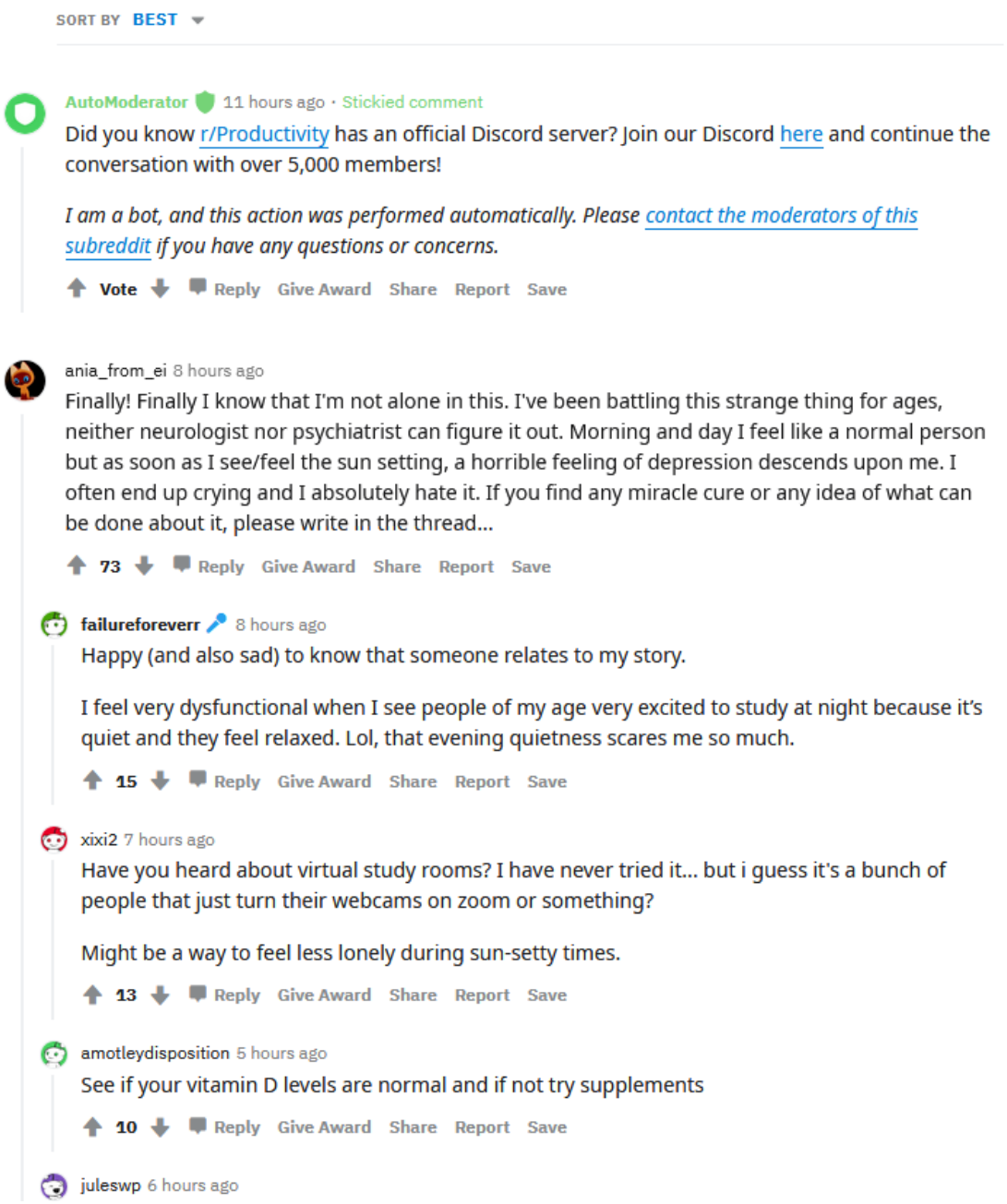


Figure 3.4: The top comments on the post in Figure 3.3.

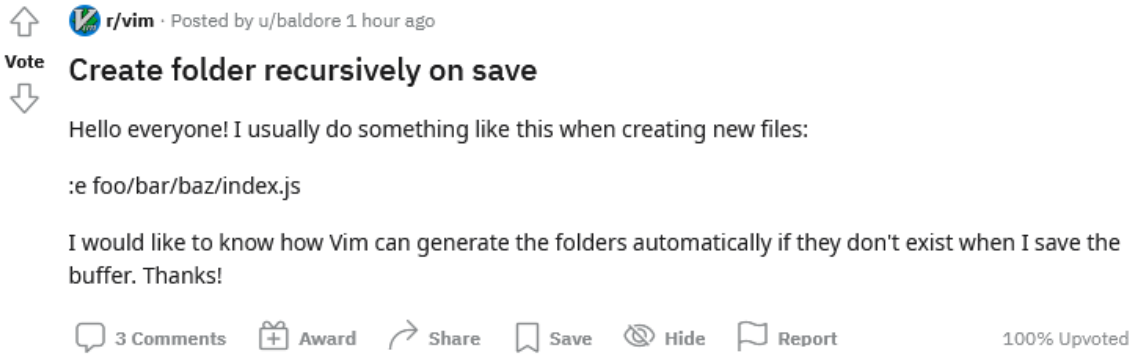


Figure 3.5: The top-rated post in the r/vim Subreddit front page on June 20, 2021.

r/vim

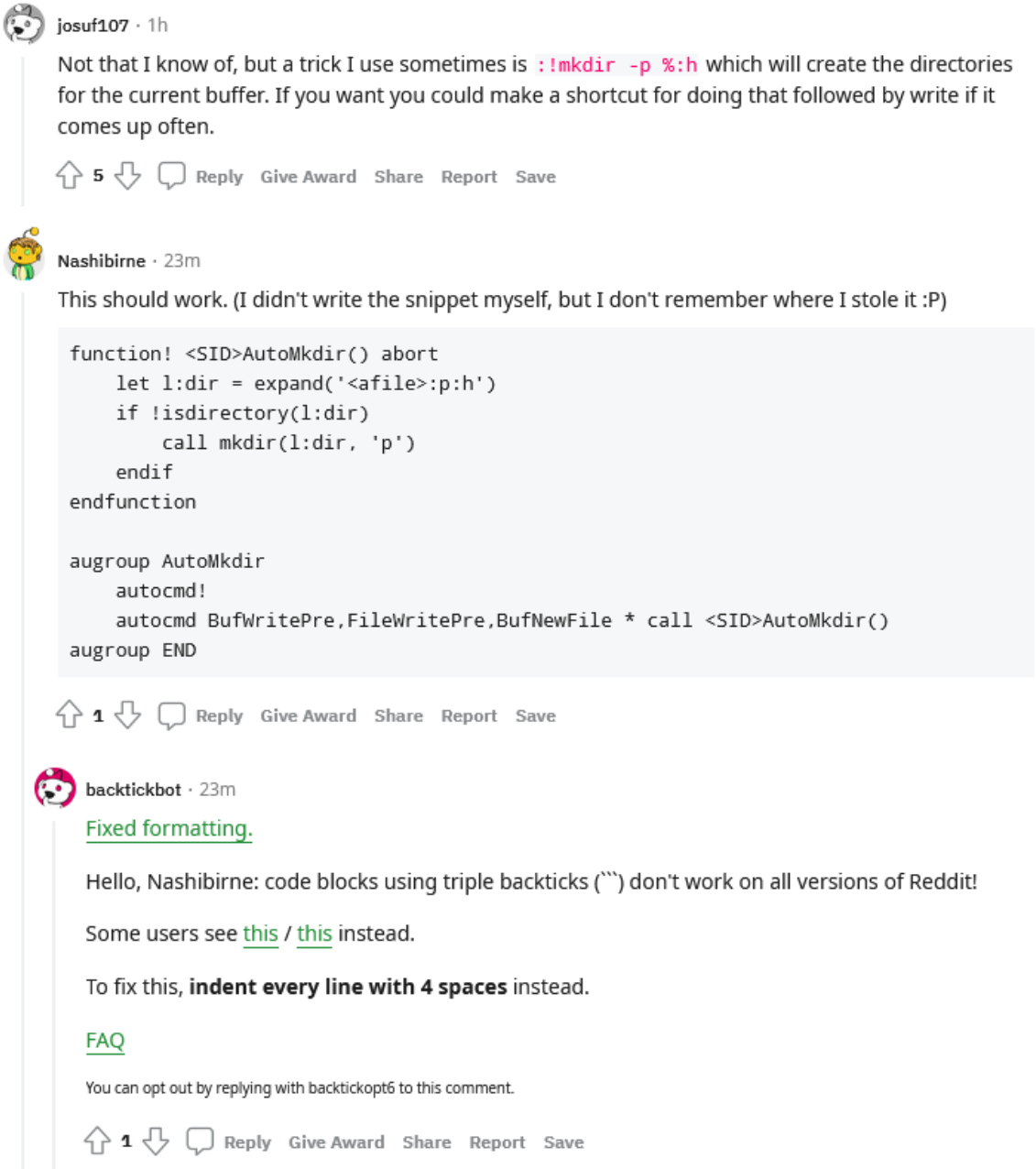
> *For Vim enthusiasts and anyone interested in Vim.*[55]

In comparison, the 69 thousand post **r/vim** community displays a very high amount of technical vocabulary and Vim-specific jargon. It is organized as a community of interest around the text editor Vim.

r/ArtFundamentals

> *Everyone keeps telling you that you need to practice your fundamentals. What the hell does that mean, and how do you do it? This Subreddit's all about concrete exercises that you can do to improve your fundamentals. We'll give you homework and we'll tell you where you're going right and wrong. Check out https://drawabox.com for more info.*[53]

The **r/ArtFundamentals** community, at 3.3 thousand posts, is close to the minimum size at which my analysis can function. I focus on it due to its unusual quality as an informal but intentional learning community. The founder and moderator (u/Uncomfortable) functions consistently as a mentor, but a variety of other users can be seen offering advice and feedback within the community's threads.



josuf107 · 1h

Not that I know of, but a trick I use sometimes is `:!mkdir -p %:h` which will create the directories for the current buffer. If you want you could make a shortcut for doing that followed by write if it comes up often.

5 Reply Give Award Share Report Save

Nashibirne · 23m

This should work. (I didn't write the snippet myself, but I don't remember where I stole it :P)

```
function! <SID>AutoMkdir() abort
    let l:dir = expand('<afile>:p:h')
    if !isdirectory(l:dir)
        call mkdir(l:dir, 'p')
    endif
endfunction

augroup AutoMkdir
    autocmd!
    autocmd BufWritePre,FileWritePre,BufNewFile * call <SID>AutoMkdir()
augroup END
```

1 Reply Give Award Share Report Save

backtickbot · 23m

Fixed formatting.

Hello, Nashibirne: code blocks using triple backticks (```) don't work on all versions of Reddit!

Some users see this / this instead.

To fix this, **indent every line with 4 spaces** instead.

FAQ

You can opt out by replying with backtickopt6 to this comment.

1 Reply Give Award Share Report Save

Figure 3.6: The top comments on the post in Figure 3.5.

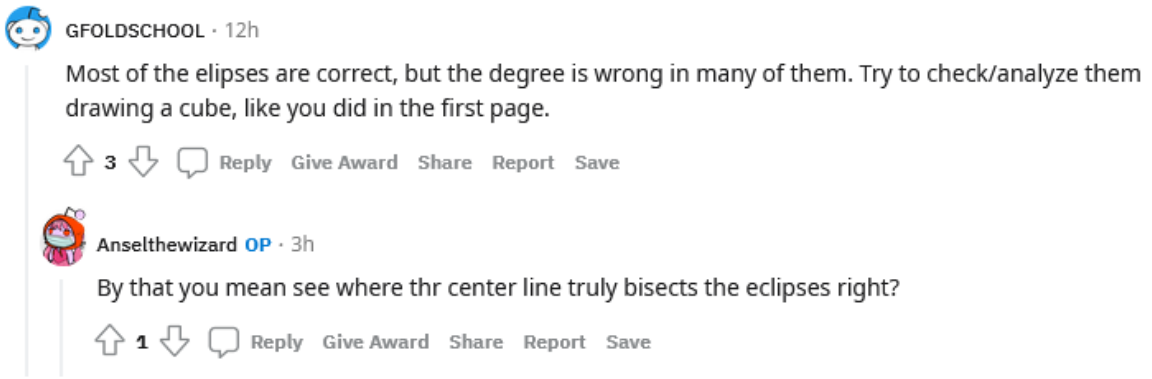


Figure 3.7: Comments on the top post in in the r/ArtFundamentals Subreddit on June 20, 2021. The post itself contains no text, but rather a series of art exercise pictures and is titled “250 Cylinder challenge- critique welcomed”.

Chapter 4

Classifying Language of Posts, Members, and Communities

In this study, I set out to establish 3 characteristics of language production in Reddit:

> **Research question 1:** Is vocabulary used within different Subreddits sufficiently distinct for a simple classifier to differentiate?
>
> **Research question 2:** Do detectable differences in vocabulary arise from differences in informal language use, or from the dominant topic of discussion?
>
> **Research question 3:** Do detectable differences in vocabulary appear between novice and veteran posters within a single Subreddit?

These three questions serve as a necessary justification for my overall investigation of learning through the lens of language use within social media. The first and second address concerns that meaningful language use variation may simply not be present within sub-forums of a single social media website. In order to study the dynamics of linguistic variation, the linguistic variation must first exist! The third research

question addresses a similar concern specifically targeted at my focus on linguistic variation as a lens for studying learning. In order to analyze how learners change their language use over time, there must be detectable differences between the language use of novices and "experts". I replace experts with veterans for purposes of my analysis because in my informal communities it is not clear what would constitute an "expert" on the community and its norms, but the notion of a veteran community member remains coherent.

I address these research questions using a neural net classifier, described in detail in section 4.2 below. While my neural net classifier cannot directly supply a theoretically justified framework for language differences, nor even a human-comprehensible one, it is the most sensitive tool I have for detecting differences that exist. If any of these three hypotheses of detectable differences are unsupported using a neural net classifier method, then less sensitive methods I apply are highly unlikely to provide useful results.

These research questions might be alternately phrased beginning with *"Under what circumstances. . . "* There are a wide variety of classifiers available, and an entire vibrant subfield within computer science is dedicated to improving their performance. In this study, I intentionally limit myself to a simple feed-forward neural net, as I find this tool is sufficient. I apply several useful filters to the raw data, but my filter choice is deliberately restricted to only those that appear necessary. In Chapter 6, I discuss what the selection of necessary filters might say about the nature of language use in social media communities.

I consider *language production* to be written content produced within a Subreddit, including both posts and comments. (For the remainder of this section I will again refer to both interchangeably as "posts".) I ignore image content for technical reasons; since all of my data comes from years-old Reddit posts, the data are archived and

the images used are often no longer available. Similarly, I ignore the content of links as these may be unavailable or may have changed. I do retain the text of links; an unusually high rate of links is detectable, as is an unusually high rate of links to particular websites. Following my focus on conversational, informal language production, I also ignore the often-formulaic posts' titles, including only their body text.

Within language production, a user's *vocabulary* is a distribution of the frequency with which they use different words. I simplify this analysis by ignoring capitalization, numbers, repetition, and most punctuation. I tokenize each post using a black box word tokenizer from the Natural Language ToolKit (NLTK)[4], which conducts some expert inference to help break each block of text into meaningful word-like tokens. For example, "can't" is subdivided into two separate tokens "can" and "n't". I apply NLTK's default word tokenizer—while tokenization is a live topic of research, there are standard solutions available which are sufficient for my purposes.

Definition (Vocabulary)**.** A probability distribution which, given a specific Reddit user (or set of users), Subreddit, and token, returns the probability that a new post from the user(s) in the given Subreddit will contain the given token. The size of a vocabulary is the number of such tokens with non-zero probabilities.

A tokenized post still includes numbers, punctuation, token order, and repeated tokens. I strip numbers and punctuation by simply filtering away all non-alphabetic token elements. Empty tokens that result from this filtering, I discard. Both numbers and punctuation can contain elements of meaning, but they are content-poor compared to alphabetic text tokens. Through removing them, I make it easier for my classifier to meaningfully separate different informal languages. In the case of numbers, there is a very high amount of token variation compared to the amount

of meaning variation present[38]. The differences between the tokens “156”, “157”, and “159” (for example) may not carry the same semantic meaning as the differences between the tokens “cap”, “car”, and “cat”. Punctuation has both less variation and more meaning, but its meaning tends to be related to syntax, and when considering vocabulary I cannot make use of markers for syntax. Similarly, I cannot make good use of information I derive from the order of words, and so I abstract this away as well. I use a *bag of words* model, which considers each post as a mathematical set of the tokens present.

Definition (Bag of words model)**.** A post consists of an unordered set of tokens, with a token present in the set indicating that one or more copies of the token could be found in the full text of the post.

Although the term ‘bag of words’ can refer to either sets or multisets, my model abstracts away repetition; each subsequent use of a token within a post contains significantly less useful information than the first use. Figure 4.1 shows a sample of my data handling. Using my bag of words model, I can take a user’s past posts as a sample to directly generate an estimate of their vocabulary. The probability that a new post by Reddit user U will contain the token T can be calculated to be:

$$Pr[U,T] = \frac{\text{\# of posts by U which contain T}}{\text{total \# of posts by U}} \tag{4.1}$$

This formula foreshadows my Token Frequency Distribution Dictionaries, used directly in Chapter 5.

```
Try keeping a journal, at the start and/or end of the day, just write down all
your thoughts, It gets it all out of your head and in front of you so you can
see what it is that clouding your thoughts.

['try', 'keeping', 'a', 'journal', ',', 'at', 'the', 'start', 'and/or', 'end',
'of', 'the', 'day', ',', 'just', 'write', 'down', 'all', 'your', 'thoughts',
',', 'it', 'gets', 'it', 'all', 'out', 'of', 'your', 'head', 'and', 'in',
'front', 'of', 'you', 'so', 'you', 'can', 'see', 'what', 'it', 'is', 'that',
'clouding', 'your', 'thoughts', '.']

['front', 'clouding', 'down', 'so', 'out', '.', 'at', 'what', 'and/or', 'is',
'it', 'journal', 'your', 'and', 'in', 'write', 'can', 'you', 'thoughts',
'the', 'try', 'that', ',', 'day', 'start', 'of', 'see', 'keeping', 'gets',
'a', 'head', 'all', 'end', 'just']

['front', 'clouding', 'down', 'so', 'out', 'at', 'what', 'andor', 'is', 'it',
'journal', 'your', 'and', 'in', 'write', 'can', 'you', 'thoughts', 'the',
'try', 'that', 'day', 'start', 'of', 'see', 'keeping', 'gets', 'a', 'head',
'all', 'end', 'just']
```

Figure 4.1: Tokenization of a sample post from r/vim.

4.1 Why vocabulary?

Vocabulary is one of many relevant characteristics of language production, alongside idiom, grammar, syntax, and context-specific conversational dynamics. I focus on vocabulary as one of these characteristics. Its relative straight-forwardness makes it an easy target for initial machine learning based language analysis. Much prior work analyzing qualities of language production, such as Fields et al's analysis of programming in Scratch[21], relies on expert inference, pre-selecting known significant language characteristics. In contrast, papers looking specifically at vocabulary, such as Kershaw et al's examination of coinage[33], have been able to eschew expert inference. While ultimately it is valuable to expand computational analysis into all aspects of language production, this history marks vocabulary as a natural starting point.

In addition to 1-grams (single tokens, i.e. vocabulary), I considered using 2- or 3-grams (ordered pairs or triplets of tokens) as an approximation of idiom. Unfortunately, experiments with 3-grams ran into severe technical difficulties. The sparsity of individual 3-grams increased the noise level in resulting data dramatically, to the point where analysis was both computationally prohibitive and overwhelmed by noise. While a single Subreddit might have a vocabulary of 10,000 unique tokens, with many repeated appearances of the same tokens throughout a sample of posts, the same Subreddit might contain 1,000,000 unique 3-grams with a much smaller proportion of repeats. (For example, the r/productivity Subreddit contains 36,218 unique 1-grams and 879,575 unique 3-grams.)

In several places throughout my studies, I apply a filter to consider only *topic-relevant* vocabulary. Topic-relevant tokens appear in a given Subreddit with at least 4 times greater frequency than they appear in Reddit as a whole. Tokens that appear more frequently within the Subreddit are likely, although not guaranteed, to be relevant to the primary "topic" of discussion within the Subreddit, and thus more likely to be meaningful to users within that community. I use this filter in the same way I strip punctuation, to hide the many tokens which provide less meaning without applying specific expert knowledge of a given Subreddit community. This additional step has the ability to reduce noise levels, but is incompatible with the first two research questions of this chapter, both of which revolve around comparison of vocabularies in settings with different topics. It also further reduces vocabulary sizes. I find vocabulary sizes on the order of 2,000 unique topic-relevant tokens when filtering those with less than 4 times greater frequency above the Reddit baseline.

```
model = torch.nn.Sequential(
        torch.nn.Linear(d.n_words,d.n_hidden),
        torch.nn.Linear(d.n_hidden,d.n_categories),
        torch.nn.LogSoftmax(dim=1)).to(device)
```

Figure 4.2: Neural net model used throughout these experiments. Note that 'device' is CUDA, used to boost performance through parallelism, and does not affect the model's logic. n_hidden, the height of the hidden layer, is set to 128. n_words and n_categories are the dictionary size and number of sources, respectively, and are derived from the data.

4.2 Classification

In these experiments, I use a simple classifier implemented using the PyTorch deep learning library[45]. The code formalizing the model itself can be seen in Figure 4.2. For consistency, I use this same classifier model in each experiment. As a simple "feed-forward" classifier, it eschews the feedback loops that characterize many current classification approaches such as LSTM, which reduces its relative discriminatory power. However, it is sufficient for my analysis and relatively simple to implement. Additionally, its behavior is relatively straightforward to explain and justify to audiences not versed in Machine Learning.

The classifier's initial inputs are posts in bag of words form as defined above. Again, each vocabulary word is tagged as either present in or absent from the post. The input layer for a particular training or testing run consists of a node for each possible word–collectively, the *dictionary*–and is populated with 1s for the words present in the post, with all remaining possible words in the dictionary being populated with 0s. The total dictionary varies from experiment to experiment and is calculated during preprocessing along with the tokenization step. The dictionary size is recorded in Figure 4.2 as n_words.

The classifier's outputs are categories. For my experiments, these are the source of each post. In a binary classification experiment to determine whether posts from

r/vim and r/productivity can be distinguished from each other, for example, the output layer consists of two nodes, one each for r/vim and r/productivity. My model implements general n-ary classification, with the number of categories determined by the experiment and recorded in the model as n_categories. Within this report, I conduct binary classification experiments and 3-ary classification experiments. The latter occur when distinguishing novices, veterans, and tourists simultaneously.

Between the input layer and the output layer I place a single hidden layer. The width of this hidden layer is recorded as n_hidden in the model, and is set to 128 throughout my experiments. The connections between the hidden layer and the input and output layers are randomized prior to each experiment.

Completing the neural net are two linear transitions, one from the input layer to the hidden layer and the other from the hidden layer to the output layer. The weights on all of these edges are randomized prior to each experiment. Weights are updated through the training process after each batch of 10 training posts. In order to support the neural net selecting a category on every input, I apply a LogSoftmax function to the output layer for each guess. Thus, the classifier will report the most likely category for a given input rather than reporting several with lower certainty.

Since many Subreddits include posts that are mostly or entirely picture-based, and the text of such posts or comments may be as brief as a single "lol", perfect classification is not an achievable goal. The classifier may assign all such posts to one source when in reality a non-trivial portion of them may be from another source. I have avoided analyzing Subreddits that are entirely devoted to image-sharing in order to mitigate this concern, but some image-heavy, text-light posts exist within the communities I examine. This generates an asymmetric set of error rates, where guesses of source A may see large numbers of false positives, while guesses of source B see few false positives but many false negatives. One way to visualize these error

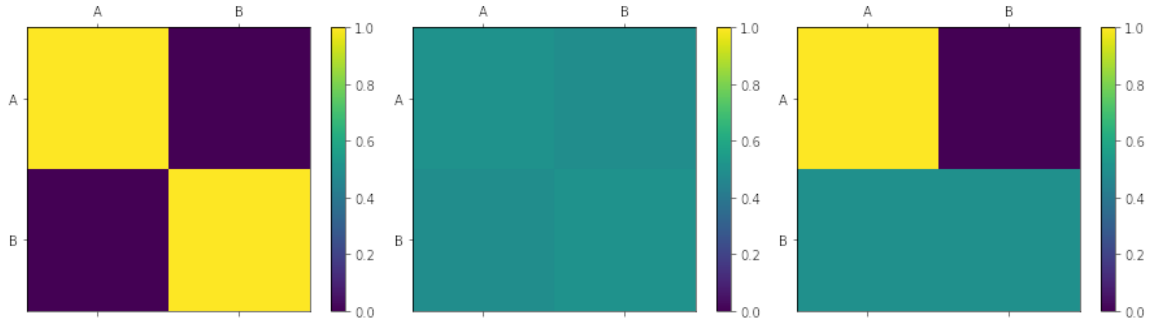


Figure 4.3: Three confusion matrices for a binary classifier showing (a) perfect classification, (b) failure to classify with posts from each source guessed correctly 51% of the time, and (c) asymmetric error where each source contains half "ambiguous" posts and all ambiguous posts are guessed to be from source A.

rates is through a confusion matrix as in Figure 4.3. Each row in a confusion matrix represents posts from the same source, while each column represents posts that the classifier guessed to be from the same source.

Values closer to 1 along the main diagonal (top left to bottom right) indicate that the classifier is guessing better. Note that values across each row must sum to 1. This is not a restriction on columns; if the classifier guesses that everything is from source A, then column A will be all 1s and the other columns 0s. Figure 4.3 shows three prototypical confusion matrices. In (a), a perfect classification correctly guesses every time resulting in 1s on the main diagonal. In (b), the classifier guesses nearly at random. An untrained classifier, or one trained on features orthogonal to the features of interest, may display this pattern. (See Appendix B for examples.) A trained classifier that has failed to learn the actual differences between two sources might display this pattern, but is more likely to highlight a single column as explained below. In (c), we see a common asymmetric error pattern. As the row associated with source A is guessed accurately and the row associated with source B is guessed uniformly at random, the classifier appears to have learned to recognize source A and not source B. In reality, in this example, there are three rough groups of items which

we would call "definitely A", "definitely B", and "ambiguous", with each source consisting of half definite items and half ambiguous items. The classifier correctly identifies all of the definite items, but guesses source A for all ambiguous items. When a high percentage of items are ambiguous, this can look like an entirely yellow column. This is a common classifier failure mode, since the classifier is designed to amplify small differences in order to learn. (Again, see Appendix B for examples.) This last error pattern overlaps significantly with a failure mode that occurs when features within the neural net receive a weight of zero. These weights are then "stuck" at zero and cannot be further updated, which can lead to the classifier refusing to classify any inputs as belonging to one or more sources. If this occurs, those columns will appear solid purple. Updating weights in batches after multiple training runs usually prevents this particular failure mode.

In the next section, I describe some of the math behind verifying that my classifier is separating truly different languages, as opposed to inventing differences within the same informal language.

4.3 Verifying the diversity of language within Reddit

I apply the simple classifier described in the previous section to posts sourced from multiple unrelated Subreddits in order to verify the existence of language variation between them. For binary classification, I calculate the "receiver operating characteristic area under the curve" (ROC-AUC) of the two input sets. This provides a measure of true positives (correct guesses by the classifier) in comparison to false positives (incorrect guesses).

Definition (ROC-AUC)**.** Given a sample set and a test with variable sensitivity, I plot the true positive rate (recall) against the false positive rate (1 - specificity). This is the ROC curve, and ROC-AUC is defined as the area under the curve.

Definition (True Positive Rate (Recall))**.** The fraction of true positives divided by the sum of true positives and false negatives. This is sometimes also referred to as the Sensitivity. For clarity, I use "Recall" to refer to the True Positive Rate and "Sensitivity" to refer to the variable discriminatory setting of my classifier.

Definition (False Positive Rate)**.** The fraction of false positives divided by the sum of true negatives and false positives. This is equivalent to $1-$ Specificity.

ROC-AUC score is a common benchmark for binary classification. If we consider a binary classification with a source A of interest and a source B distractor, classification errors can be broken down into false positives (items from source B where we guess A) and false negatives (items from source A where we guess B). We can imagine a classifier with a variable sensitivity rating, such that dialing up the sensitivity will result in more A guesses and dialing it down will result in more B guesses. Thus, increasing the sensitivity will result in better recall (fewer false negatives) but worse specificity (more false positives). Conversely, lowering the sensitivity results in worse recall but better specificity. Since both types of errors are relevant to us, the ROC-AUC score which incorporates both is a good measure of classification success. Because the classifier's guesses are mediated through the machine learning process, there is no guarantee that recall and specificity are formally monotone with respect to sensitivity, but given the non-adversarial setting we may claim that their aggregate behavior should approximate what we would see if they were monotone.

In many scenarios, ROC-AUC is calculated by testing many sensitivity ratings of the same classification exercise. For my experiments, the classifier has an implicit

sensitivity rating, but not one that I can directly set. Nor can I put, even after the fact, a number on the sensitivity rating that an experiment used. The training process generates guesses with some degree of confidence, and determines for itself what degree of confidence is sufficient to make a particular guess.

This is arguably a weakness of my experimental procedure. However, as I explain below, I am able to identify three points on the ROC curve even without the ability to directly measure the sensitivity of my tool, which combined with my claim of approximate monotonicity allows me to calculate a reasonable approximation of the "true" ROC-AUC. Classifiers similar to the one I employ can sometimes report their degree of confidence in each guess, and we can imagine dialing up or down the sensitivity of such a classifier by modifying the required degree of confidence for the classifier to make a particular guess. This would allow for multiple ROC measurements, but would not actually address the underlying issue. The sensitivity of a classifier is bound up in the way weights are updated during the learning process. My classifier and others like it are designed to amplify small differences, to intentionally overstate their confidence as part of the machine learning process. This significantly warps any resulting confidence measurements, to the degree that reporting them would be more misleading than helpful.

While I cannot directly measure the sensitivity rating of the classifier, it is straightforward to measure its true positive and false positive rates in the output, providing a single data point along the ROC curve as shown in Figure 4.4. Likewise, I can anchor either end of the curve with the fact that a classifier with sensitivity threshold 1 would naturally produce 0 positive guesses and a classifier with sensitivity threshold 0 would produce only positive guesses. I can thus approximate the ROC-AUC as the

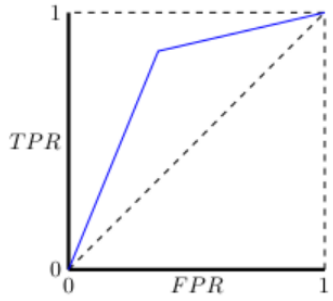


Figure 4.4: Measuring the degree to which a tool can achieve high True Positive Rate and low False Positive Rate simultaneously, the ROC-AUC score is the area under the blue curve.

area of a quadrilateral whose area is simply

$$\text{ROC-AUC} = \frac{TPR - FPR + 1}{2} \tag{4.2}$$

which is a formula readily available after training the classifier.

Note that since I cannot directly control the sensitivity of the classifier, I cannot control *which* point on the ROC curve I get. A poor choice of sensitivity rating could lead to a difference between the true ROC-AUC and my estimate. If we assume that both the TPR and FPR are monotone with respect to sensitivity, then they are also monotone with respect to each other. However, while we may assume monotonicity there is no theoretical basis to assume that the ROC curve is smooth or even integrable. With a data point having FPR a and TPR $f(a)$, the true ROC-AUC is bounded,

$$f(a) - af(a) \leq \text{ROC-AUC} \leq af(a) + 1 - a \tag{4.3}$$

resulting in the error of my approximation being at worst $1/2$ as the classifier finds TPR or FPR close to either 0 or 1.

Calculated values deviating from 0.5 indicate more success separating languages from the two different sources. Given the high degree of variability inherent to language production generally, I take a ROC-AUC at or above 0.6 to indicate with strong confidence the separability of the languages. A ROC-AUC between 0.55 and 0.6 indicates a weak degree of separability which might be realizable with a more sophisticated classifier. Pairs of informal languages in this latter range are distinct languages, but the differences between them are more subtle or more specific. A human expert attuned to which word choices particularly matter may be able to cleanly classify between them if the differences occur in specific language use, or may not even be able to notice that the languages are distinct if the differences are subtle and diffuse. For comparison, note that the data behind the three confusion matrices in Figure 4.3 produce the following ROC-AUC scores: 1 (a), 0.51 (b), and 0.75 (c).

Running my classifier pairwise on my primary three communities of interest (r/vim, r/productivity, and r/ArtFundamentals) produced the following ROC-AUC scores:

- r/ArtFundamentals vs r/productivity: 0.91
- r/ArtFundamentals vs r/vim: 0.91
- r/productivity vs r/vim: 0.82

For reference, I also ran a repeated classification trial of r/productivity against itself, with 10 runs, producing a mean ROC-AUC score of 0.50 and standard deviation of 0.01. This suggests that a ROC-AUC score threshold of 0.55 for weak separability is a conservative estimate, and likewise the strong separability score threshold of 0.6 is also conservative.

I can thus answer my first research question: Yes, the vocabulary used within different Subreddits is sufficiently distinct for a simple classifier to differentiate.

r/3amjokes	*/r/3amjokes - for all the stupid humor of sleep deprivation. "So bad, its good" Have you been up for longer than a normal human being can operate? Good. Have you just laughed at a joke that wouldn't be funny otherwise?*
r/dadjokes	*Welcome to r/dadjokes - a homely place for the best and worst of jokes that make you laugh and cringe in equal measure. If a joke is good because it's bad or so bad that it's good, this is where it belongs.*
r/puns	*For the instances of puns in daily life.*

Table 4.1: Self descriptions for each joke community, according to their "About Community" blurbs.[50, 51, 52]

4.4 Addressing differences in topics of conversation

A reader might reasonably ask at this point whether my classifier truly separates the languages spoken by members of these different Subreddits, or whether it simply classifies by topic of discussion. While there may not be a shared project as in a community of practice, each Subreddit community shares a specific interest that defines the community and discussion within the boundaries of the community naturally focuses on that topic. For example, given the Subreddits r/ArtFundamentals and r/productivity, a post which mentions "art" is quite likely to have come from the former. Members of r/productivity may talk about art, but they're less likely to, and in particular less likely to talk about art within the r/productivity community.

To address this "topic effect", I analyze several Subreddits with very similar topics. r/3amjokes, r/dadjokes, and r/puns, as their names suggest, all revolve around the sharing of (and groaning at) really bad jokes. Moreover, although there are certainly many kinds of humor, these three communities dedicate themselves to quite similar ideas of humor as shown in Table 4.1. As such, we should expect very little topic effect to appear in their language production.

Running my classifier in each of these communities, pairwise, results in the following ROC-AUC scores:

- r/3amjokes vs r/dadjokes: 0.57
- r/3amjokes vs r/puns: 0.63
- r/dadjokes vs r/puns: 0.61

While these scores are noticeably lower than those between communities without a shared focus, they remain significant. Thus I can answer my second research question: the "topic effect" is real, but it also cannot account for all of the classifier's ability to differentiate informal languages.

4.5 Veterancy

Recall that a *veteran* is defined as:

Definition (Veteran)**.** A community member who has produced a significant number of posts within the community. A post is considered to be a veteran post if it was made by a veteran community member *and they had achieved veterancy by or before making the post*. A veteran might be considered by virtue of experience to be an authority on the community.

This definition supposes a link between quantity of prior posts and some degree of informal authority. In the language of Communities of Practice, we might say that a veteran is equivalent to a core member of the community, someone who a member of the community would likely perceive as being a knowledgeable authority on the community itself.

That there is some correlation between post quantity and domain authority (if not actual expertise) seems natural, but what is the nature of the link? If there is

causation, is it that posting more causes one to develop the domain authority, or that users who are already domain authorities tend to post more? I hypothesize based on prior experience as a member of various forums that even if a newcomer is "objectively" an authority on the norms and practices of an online community, other posters are unlikely in the extreme to grant them that status. In Chapter 5, I conduct an analysis on whether a user's initial posts can be used to predict their longevity, which would also have been a sign of this direction of causality. That experiment finds no detectable link, further suggesting that domain authority is not an innate quality that causes posting. Note that neither analysis disputes that, once a community member has achieved veterancy, their status may compel them to *continue* to post more. That initial posting quantity to some degree causes veterancy is impossible to fully prove, but I provide evidence for this model through my analysis of how novices' language use changes as they post more and become veterans. Speculatively, I might directly address the question of causation through a future experiment where I identify conversational markers of community members deferring to or otherwise acknowledging the veterancy of their fellows. I could then compare these instances to the veterancy levels of the speakers as estimated by their language use to tell to what degree the interlocutors were reacting to the learned writing patterns of the speakers.

Setting aside questions of causation, there is also a question of just how great a quantity of posts needs to be in order for veterancy to be achieved. Consider a hypothetical user U. If they have made 40 posts within Subreddit r/S, are they a veteran? To address this question, each of my analyses applies a "veterancy threshold", a selected number of posts at which I consider a user to be predominantly a veteran. If I set a veterancy threshold of 20 posts, then poster U's 20th and later posts all occurred while they were a veteran member of the community. However, their first 19 posts occurred before they achieved veteran status, and I identify these as "*novice*"

posts. Notice that U is both a novice and a veteran at different times, although any one of their posts can be unambiguously labeled as either a novice post or a veteran post and not both.

It turns out that posters who never reach veterancy are a significant portion of any Subreddit's contributors, if not necessarily a significant portion of the community. I call this class of posters "tourists". Since Reddit is an ongoing platform and all of the Subreddit communities I examine are active today, it is quite likely that at least some posters identified as tourists in my 2007-2015 data are now veterans, and the posts I identified as tourist posts should have been classified as novice posts. However, for purposes of my analysis separating novice and tourist language production remains sound. Users who did not manage to achieve veterancy during the time period I study are statistically more like a prototypical tourist than novices who do achieve veterancy somewhere during the time period. In the r/productivity Subreddit, for example, tourists whose next post at their current average posting rate would occur after the end of my dataset constitute just 7% of recorded tourists in that community. Of this 7%, the average interval between posts is just over 10 months. In r/vim, the fraction is slightly higher at 9%, also with an average post interval of 10 months. The arguable exception are tourists who begin posting toward the end of my data collection (i.e. in 2015), but this group is vanishingly small relative to the size of my data set and unlikely to significantly raise the overall level of noise present within the data. Of tourists in the r/productivity Subreddit and the r/vim Subreddit each, only 0.8% both started posting in 2015 and would make their next post after the end of my dataset. A deeper analysis might introduce a real-world time estimate for how long a user must go between appearances in a Subreddit to reset their tenure and be best described once again as a tourist, but I leave this for a future study.

4.5.1 Veterancy threshold determination

Where not otherwise discussed, I set a default veterancy threshold of 10 posts. Given the general exponential decay pattern of post frequency (which holds for individual Subreddits just as it does for Reddit as a whole), I want to set a veterancy threshold that lands slightly into the long tail of the post count distribution and so should exclude most true novice and tourist language production. I find that 10 posts meets this criterion for the Subreddits that I examine.

Setting an appropriate veterancy threshold is a statistical task based on the theoretical notion that veteran identity roughly follows an "s" curve. Initially (zero posts), a new member would be assigned novice status by virtually every member of the community. Eventually (post infinity), virtually every member would accord them veteran status. Individual community members might disagree on precisely when the transition occurs, with individual variation driven by individual quirks far out of scope for us to characterize, but there is some smooth transition from one identity to the other. Ideally, a veterancy threshold would be placed roughly at the midpoint of this curve.

It's worth noting that different individuals will adapt to community norms at different speeds, and different individuals will be accorded veteran status at different speeds. Aside from variable learning aspects, there is a high likelihood that biases held by community members will be applied to this process, and simply asking members who is a veteran and who is not will lead to their subconscious or conscious biases skewing the results to the point of uselessness. I raise this concern not because it is practical to ask in the Subreddits of interest "who is a veteran member of this community?" (It isn't.) However, machine learning could be used to generate a preliminary estimate of community language norms and then veteran identity could be

determined by proximity to those norms, leading to a distinct and grounded calculation of a veterancy threshold for each individual poster in the community. I leave this analysis to a future study, and instead apply my expert knowledge to place the threshold.

4.6 Analysis of novicehood and veterancy within Subreddits

Recall that in this chapter I seek to demonstrate that the vocabularies of veterans, novices, and tourists are meaningfully different within any given Subreddit community. I address this question by applying my classifier to a Subreddit, training it on a subset of posts labeled veteran, novice, or tourist, and then testing it against a different subset from the same Subreddit.

I find that my classifier behaves differently in different Subreddit communities. The separability of language elements within these communities may be linked to characteristics of the communities themselves. To illustrate these classification differences, I apply several variations on my classification analysis to my three primary communities of interest: r/ArtFundamentals, r/productivity, and r/vim. I further consider how their distinct properties may affect my results.

When initially presented with tourist, novice, and veteran groups from the same Subreddit for 3-ary classification, the classifier is generally unable to differentiate between novices and veterans. The sole exception I find is in r/ArtFundamentals, where tourists and novices are not distinguishable from each other, but veterans are distinguishable with a true positive rate >0.8 and low false positive rates even in 3-ary classification. In Figure 4.5 below, I report the confusion of my classifier for each

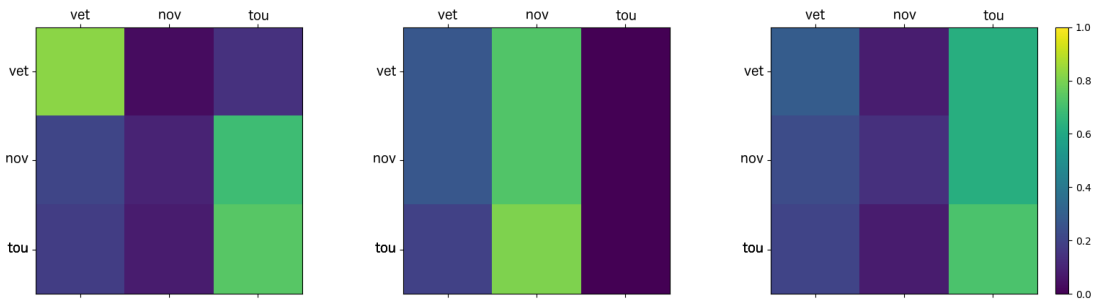


Figure 4.5: Confusion matrices for veteran posts, novice posts, and tourist posts in r/ArtFundamentals (left), r/productivity (center), and r/vim (right) with no preprocessing of data beyond the bag of words model and removing non-alphabetic tokens.

Subreddit. Each row indicates the true source of a post, and the column indicates the guess that the classifier made. A high score indicates that a high fraction of posts from that source group were given that guess. Thus, we can see in r/productivity that the majority of all posts are guessed to be novice posts, indicating a failure of the classifier to learn the difference between categories. A similar pattern is displayed in r/vim, where most posts are guessed to be tourist posts. Repeated trials in this setting produce similar results to those shown in Figure 4.5, although which column the classifier highlights varies from trial to trial. Rerunning the categories in pairs produces ROC-AUCs of 0.59 for novices vs tourists, 0.53 for novices vs veterans, and 0.47 for tourists vs veterans in r/productivity; in r/vim we find ROC-AUC scores of 0.55 for novices vs tourists, 0.52 for novices vs veterans, and 0.55 for tourists vs veterans. In both cases, novice and veteran language production is not distinguishable at even the "weak" threshold I set, and while tourists are (barely) distinguishable in these particular runs, this effect is not consistent.

Noise is a major factor influencing the separability of different types of posts. To reduce the noise, I run several variations of pre-processing. Note that the classifier itself is unchanged; I instead alter the posts themselves.

4.6.1 Topic-relevant vocabulary

One classic approach to reducing noise in text data is to remove a stoplist of the most common words in the language. These words are presumed to not contain meaningful information, as they tend to be universal terms appearing everywhere and most of them serve a syntactic rather than semantic purpose in a sentence. A more significant approach is to look only at words that appear more frequently within a Subreddit relative to Reddit as a whole. Limiting analysis to only these "topic-relevant" words may help a classifier sort signal from noise by immediately excluding the vast majority of terms while not excluding any local jargon that novices might be in the process of learning. While the topic effect mentioned previously is a natural concern when doing this type of filtering, tourists, novices, and veterans within the same Subreddit share the same topic and the same set of topic-relevant vocabulary.

In order to filter for topic-relevant vocabulary within a Subreddit, I compare the frequency of vocabulary words within that Subreddit to their frequency in a random sample of Reddit as a whole. I consider a vocabulary word to be topic-relevant if it occurs at least 4 times as often within the Subreddit as elsewhere. I also restrict my analysis to vocabulary words that appear at least 4 times within the Subreddit. This removes the many globally unique 1-grams which are not relevant to my analysis even though they technically occur more often within my Subreddit of interest.

I apply my filtered bag of words model as input to my classifier. Each post, already a bag of words, is reduced to a much smaller bag of topic relevant words. I then train and test my classifier. Figures 4.6, 4.7, and 4.8 show the results of applying this filtered analysis to r/ArtFundamentals, r/productivity, and r/vim.

In the case of r/ArtFundamentals, not much changes with the introduction of a filter for topic-relevant vocabulary. Veteran posts remain distinguishable from novice

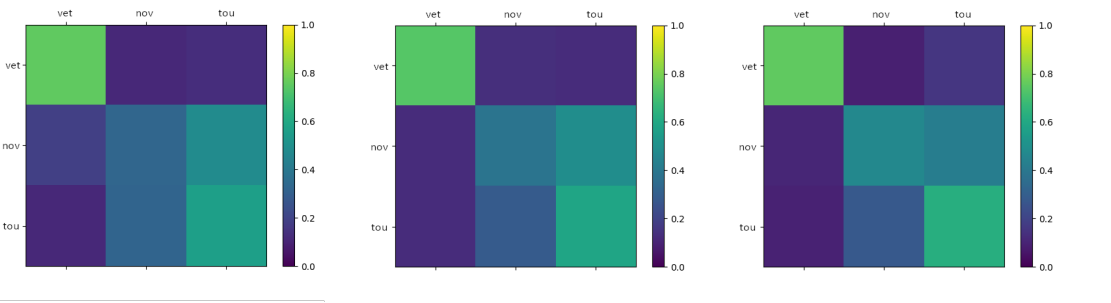


Figure 4.6: Confusion matrices for veteran posts, novice posts, and tourist posts in three successive training runs on r/ArtFundamentals with topic filtering.

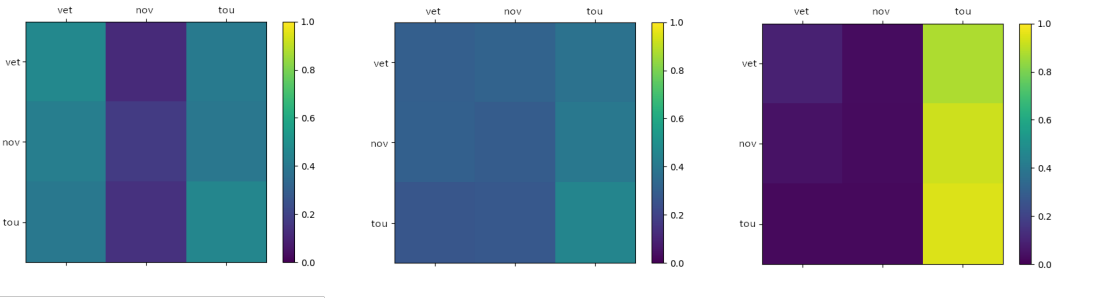


Figure 4.7: Confusion matrices for veteran posts, novice posts, and tourist posts in three successive training runs on r/productivity with topic filtering.

and tourist posts with ROC-AUC scores between 0.78 and 0.83, while novice and tourist posts cannot be readily distinguished from each other. Speculatively, this may relate to the community's focus on informal teaching and learning, with some veterans deliberately taking on a role as teachers and as a result producing different language, while both novices and tourists take on the student role.

Three successive analysis runs on the filtered version of r/productivity produced no ROC-AUC scores above the weak separability threshold. The precise failure of the classifier to differentiate novice, veteran, and tourist language production in this setting and the r/vim setting varies from analysis run to analysis run.

Three runs on the filtered version of r/vim produced highly variable results. ROC-AUC scores veterans vs novices were consistently below the weak separability thresh-

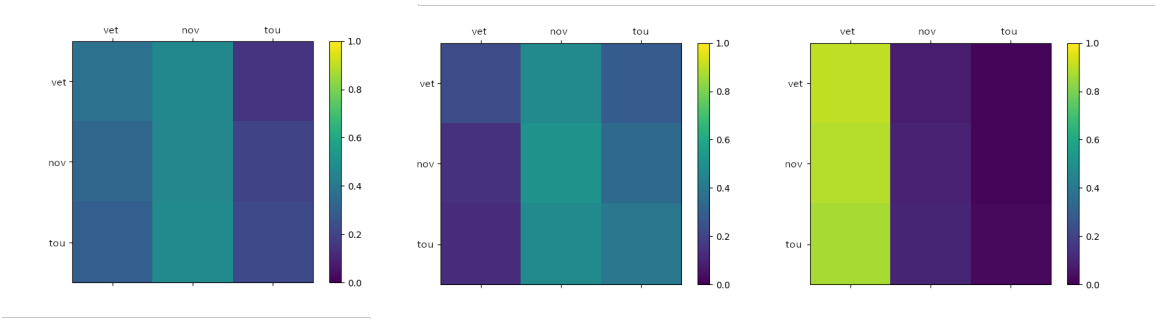


Figure 4.8: Confusion matrices for veteran posts, novice posts, and tourist posts in three successive training runs on r/vim with topic filtering.

old. For veterans vs tourists the scores are, left to right: 0.61, 0.47, and 0.94. For novices vs tourists the scores are, again left to right: 0.63, 0.58, and 0.54. In spite of the characteristic failure pattern displayed in the third run where nearly all posts are guessed to be veteran posts, it is possible that filtering for topic-relevant vocabulary uncovers the distinguishability of tourist posts in the r/vim community. The simple classifier, however, appears to be at best marginally effective at differentiating even these posts from novice and veteran posts.

Overall, topic filtering does not appear to provide a significant improvement over the baseline classifier performance. It does achieve slightly higher ROC-AUC scores, suggesting a slightly better signal-to-noise ratio, but not sufficiently improved to reliably distinguish community member categories from each other. As with any classification failure, these results may indicate either that the classifier was unable to overcome noise present in the data to "see" the underlying pattern, or that there is no underlying pattern for the classifier to observe. A reader should at this point maintain a healthy skepticism about the idea of conversational language differences between tourists, novices, and veterans. The only reliable result *so far* has been to differentiate veterans in an intentional, if informal, learning community. These members may have explicitly taken on "teacher" roles, which might plausibly have effects on their

language production that the inconsistent mentorship roles of community veterans might simply not have.

I ground each successive modification of the classifier or its inputs (posts) in a direct theoretical justification in order to avoid filtering away important elements of my data. In this case, the high degree of variance in the ROC-AUC scores between training runs suggests that data sparsity may be an issue for the topic-filtered approach. Prior to topic filtering, the median post length (in tokens) for r/ArtFundamentals, r/productivity, and r/vim are respectively: 60, 25, and 26 tokens. Topic filtering reduces these lengths further, and may reduce them below levels at which classification is viable. This could account for some of the classification difficulties, as extremely short posts are often nigh-impossible to accurately classify. One common solution to this issue is to use longer samples. While I do not have the ability to simply select longer sample posts, I can easily combine or augment posts to simulate longer utterances from community members.

4.6.2 Sliding window post augmentation

The sliding window is an approach where rather than take each post individually, I identify a target quantity of text and augment each post to make sure that it includes at least that amount of text. Various implementations of this approach all address the same concern. Statistical analysis such as classifiers do not work properly when individual pieces of input data are too small, so data that are linked together are analyzed together in chunks that are more computationally tractable. In Reddit post data, where some posts even in text-heavy Subreddits may still consist of a single link or image, using a sliding window allows me to avoid trying to classify such posts on

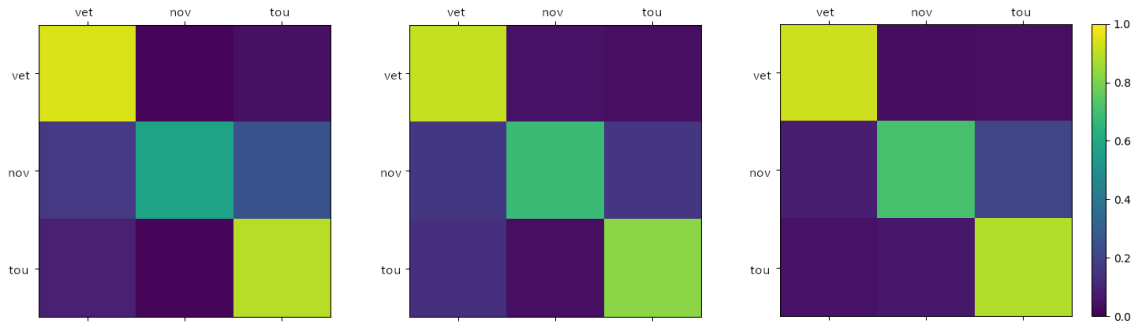


Figure 4.9: Confusion matrices for veteran posts, novice posts, and tourist posts in three successive training runs on r/ArtFundamentals with sliding window analysis.

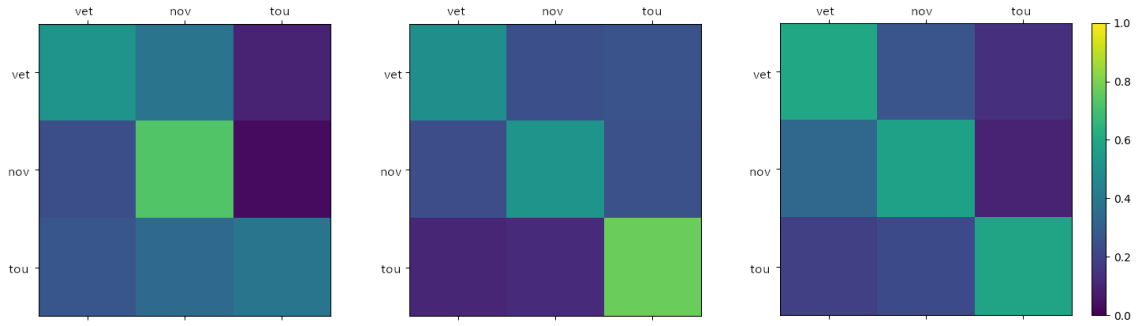


Figure 4.10: Confusion matrices for veteran posts, novice posts, and tourist posts in three successive training runs on r/productivity with sliding window analysis.

their own, instead classifying them in combination with subsequent text-containing posts made by the same author.

The size of the target window is a tunable variable which I set to 100 tokens. This is close to the average post length for r/ArtFundamentals, and roughly twice the average post length for r/productivity and r/vim. (See Table 3.5.) I apply the post augmentation step after tokenization but prior to building the bag of words model, so an augmented post under the sliding window approach may only contain 70 or 80 distinct tokens by the time it reaches the classifier. This is still enough to provide a robust set of features for the classifier to work on for every post in the collection, thus reducing noise.

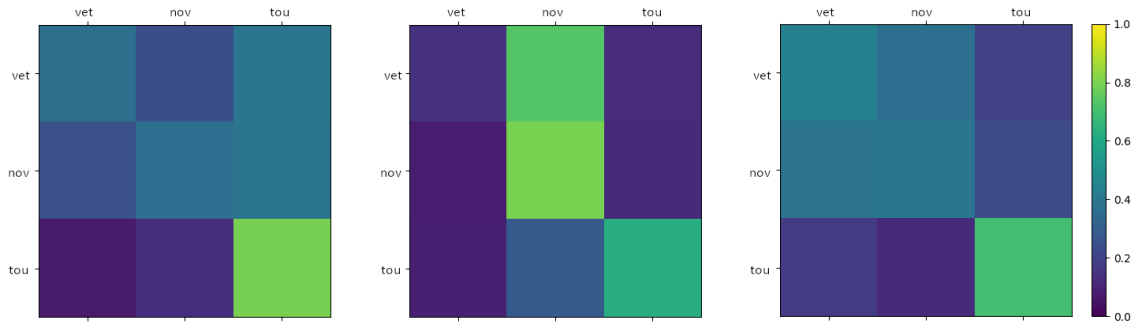


Figure 4.11: Confusion matrices for veteran posts, novice posts, and tourist posts in three successive training runs on r/vim with sliding window analysis.

In my first variation of sliding window augmentation, I augment each post from user U with the chronologically following post(s) from the same user until the window size target is reached. The posts seen by the classifier include a separate entry for every post, but there may be significant overlap between them if user U made short posts. Classifying my three communities of interest under this variation produced the confusion matrices shown in Figures 4.9-4.11. The classifier is much more successful at separating veteran, novice, and tourist language in all three communities under this modification.

In r/ArtFundamentals, the classifier almost completely learns to recognize both veteran and tourist language production, suggesting the existence of clear markers for each that tend to appear with relatively high frequency. That novice language production displays some overlap is not surprising, since I expect some overlap between novice identity and both tourist and veteran identities due to individual variation within any learning process. It is quite likely that individuals right at the boundary between novice and veteran, for example, use language that sometimes aligns with either side of that artificial boundary.

Posts in the r/productivity community display less perfect separability than those in r/ArtFundamentals, but the classifier does successfully learn all three categories

	r/ArtFundamentals	r/productivity	r/vim
veteran vs novice	0.87	0.62	0.56
veteran vs novice	0.86	0.56	0.55
veteran vs tourist	0.93	0.70	0.61*
veteran vs tourist	0.93	0.62*	0.66
novice vs tourist	0.74	0.72	0.64*
novice vs tourist	0.69***	0.56**	0.59*

Table 4.2: Aggregate ROC-AUC scores (averaged over 3 runs each) for veteran posts, novice posts, and tourist posts from r/ArtFundamentals, r/productivity, and r/vim using post-by-post sliding window augmentation (white background) and continuous text string sliding window augmentation (shaded background). Calculations which included runs with <0.5 ROC-AUC score had that score converted to a score equally distant above 0.5 before averaging (marked * per inverted score).

simultaneously, indicating that that small post size was a significant barrier to successful classification.

In r/vim, separability is markedly improved over previous iterations, but the classifier still cannot reliably separate veteran language production from that of novices. Tourists, however, are now clearly identifiable. Speculatively, this may be related to the high-jargon technology-oriented nature of the r/vim community, where novices can easily look up vocabulary words relevant to the community, and the primary challenge is learning to properly restrict their usage to the appropriate times.

ROC-AUC scores from pairwise analysis are shown in Table 4.2. The resulting scores are dramatically higher and well above the threshold for strong separability in both the r/ArtFundamentals and r/productivity communities, suggesting again that short post length is a major issue for classification of posts from these communities. The scores for r/vim also improved, although not as much, with strong separability of tourist posts and weak separability of veterans from novices. This success in r/vim relative to the more ambiguous confusion matrix is a result of the relative ease of binary classification relative to 3-ary classification.

My second variation of sliding window augmentation addresses a concern with the prior augmentation: augmentation of short posts may lead to the vocabulary present in those short posts being over-represented within the classifier's input set. In this variation, I avoid assigning the same tokens to multiple post inputs by pre-processing a user's corpus as a continuous string of text. I take 100-token slices of that text, labeled with the age of the post where the slice begins, and classify these slices. The resulting input set can include multiple posts from a given author of the same age (if an author wrote a post longer than 100 words), and can "skip" ages (if an author wrote several short posts in a row). Table 4.2 shows the pairwise ROC-AUC scores for this variation as well. This approach turns out to be less effective than post-by-post augmentation, although it also dramatically improves over the no-augmentation condition. In particular, we see a high frequency of ROC-AUC inversions. Deviation from 0.5 indicates successful separation of the two language sources regardless of the direction of the deviation, but it is unclear why this approach would result in the classifier often learning to separate sources *backwards*. Simple overtraining would result in ROC-AUC scores very close to 0.5, whereas the median inverted score in this set is 0.40.

In either variation, it is clear that at least within some Subreddit communities, the positive hypothesis for Research Question 3 is robustly supported: Detectable differences in vocabulary do appear between novice and veteran posters in the Subreddit, and also between those posters and tourist posters. The degree of difference varies from community to community, likely as a result of the specific circumstances of that community.

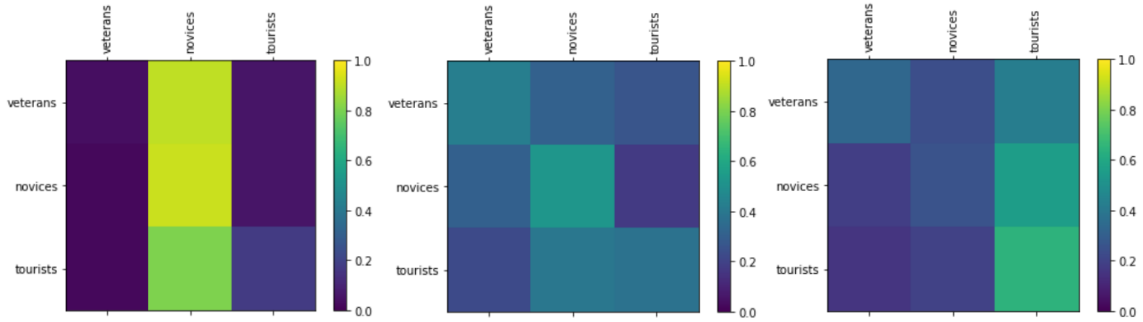


Figure 4.12: Confusion matrices for veteran posts, novice posts, and tourist posts in three successive runs on posts tagged "HTML" in StackOverflow.

4.7 StackOverflow

While the primary focus of my classification experiments is Reddit, I repeat several experiments within the context of the community formed by the StackOverflow "HTML" tag. The techniques I apply within the context of Reddit should transfer to social media text data in other online communities. The informal community bounded by StackOverflow's tagging mechanism serves as an example of such a community, as many users will only interact with tags relevant to their expertise or interests. Similar to Reddit's personalized front pages, StackOverflow allows a user to watch one or more tags, thus filtering the content they see to only include questions with those tags. Initial classification of veterans, novices, and tourists for posts using the "HTML" tag displays a high degree of uncertainty, as seen in Figure 4.12. A closer inspection of ROC-AUC scores, however, reveals that even unfiltered and absent sliding window augmentation, tourists are distinguishable. Table 4.3 shows ROC-AUC scores for three successive runs.

	HTML tag
veteran vs novice	0.55
veteran vs tourist	0.55**
novice vs tourist	0.75*

Table 4.3: Aggregate ROC-AUC scores (averaged over 3 runs each) for veteran posts, novice posts, and tourist posts from StackOverflow posts tagged "HTML". Calculations which included runs with <0.5 ROC-AUC score had that score converted to a score equally distant above 0.5 before averaging (marked * per inverted score).

4.8 Thoughts and future directions

My vocabulary-based classification strongly suggests that there are real differences between the language produced by novices, veterans, and tourists in specific Subreddits, as well as between separate Subreddit communities. I use my classifier largely to help me demonstrate that this difference exists in the face of a highly noisy environment where alternative analyses might be prone to false positives. That the lens of vocabulary is sufficient for a classifier to detect differences suggests there may be hope for analysis of subtler elements of language production as well. Moreover, this confirms that vocabulary is an integral part of the language patterns that novices must learn when picking up informal languages.

While I subdivide posts into tourist, novice, and veteran categories, each of these is a heterogeneous group made up of individuals who have additional salient identities. The features my classifier identifies as signifying veterancy, for example, represent features that best statistically predict veterancy across the entire training sample. Thus, features that predict a subset of veterans may not be recognized. For example, veterancy, and in particular the recognition of veterancy within a community, may not be demographics-neutral. While members of different racial demographics in the US may use social media at comparable rates, this does not imply that they produce comparable amounts of posts, and particularly does not imply that they

receive comparable amounts or types of feedback on their posts. Analysis of a social media setting where demographics are known could shed some light on the interactions between racial identity (for example) and veterancy within that community, and perhaps provide some guidance for analysis like mine that occurs in settings where users are anonymous.

Data sparsity is a clear threat to validity of future studies, as it was a challenge for my analysis. For most of the communities I considered, specific pre-processing to address data sparsity was required in order to see positive results. Additionally, my initial investigations of idiom through 3-gram tokens were largely unsuccessful as a result of these data sparsity challenges. Based on my experiences, it seems appropriate to incorporate handling data sparsity into the core design of future research in this area. This could include measurements of sparsity, such as recording the frequency of tokens or other artifacts that only appear once within the dataset or recording the number of features included in each datapoint. One example of the latter is my measurement of the numbers of tokens appearing in each post. In addition to measuring sparsity, future work can conduct experiment design so as to limit its effects. This might include placing a greater focus on analyzing data fram larger communities, or applying methods similar to my sliding window post augmentation to combat it.

Classifiers specialize in the prediction of individual new or unknown data points. One future direction for this type of analysis could be as an assistant to personalized learning tools. Such a tool could include conversational data (as well as evaluative questions) to help target useful educational material to the student. This type of approach would likely need a more sophisticated breakdown of the novice category than is present in my study, which would be tailored to the specific learning setting. Outside the domain of education, this analysis could be used, with extreme caution,

to detect social media users evading community bans. A new user producing veteran text patterns within close proximity to the banning of a veteran user could be that same person returning with a new account. It may be more difficult for the ban-evader to fool a generalized novice-vs-veteran analysis than analyses which focus on language patterns specific to that user's prior posts.

Chapter 5

Individuals' Trajectories of Language Production

In this chapter, I zoom in to examine changes in individuals' language production over time, focusing again on word choice (vocabulary). For a given Subreddit community, I consider the following research questions:

> **Research question 4:** As a novice becomes an experienced member of the community, does their generated vocabulary size for posts within the community grow?
>
> **Research question 5:** Does the generated vocabulary size of a user's initial posts within a community predict their longevity?
>
> **Research question 6:** How do novices' word choices in posts change over time with respect to a prototypical distribution of veterans' word choices?

Having demonstrated in Chapter 4 that tourists, novices, and veterans within a Subreddit display different patterns of language production through their word choices, it now makes sense to ask how an individual poster changes their vocabulary

over time. I begin with an analysis of vocabulary size, with the initial hypothesis being the common-sense idea that as a novice learns to speak the local language, the *size* of their vocabulary should grow in parallel with their increased knowledge and transition from a peripheral participant to a core community member. RQ5 addresses a specific concern that learning may not occur and the patterns I see might in fact be a symptom of a process where new-comers who already fit in with the community norms stay (becoming novices and subsequently veterans) while new-comers who don't fit in leave (becoming tourists). If this were the case, we would expect initial proximity to community norms to predict how much a user posts, including prediction by initial vocabulary size. Of course, while vocabulary size provides a straightforward measurement of individuals over time, vocabulary is as more about *which* words than *how many*. I measure the statistical distance between novices' posts and a distribution that approximates a prototypical veteran post. Hypothetically, a novice's posts should initially land quite far away from the prototype veteran post and as the novice learns community norms their posts should begin to approach, although I expect some distance due to the random and noisy nature of my data even when comparing the posts of actual veterans to the prototype. As My results in this chapter complicate the picture for each of these questions. While vocabulary acquisition does appear to occur, I find that there are also periods where the learning process a novice undergoes with respect to their vocabulary would be better described as "winnowing". I find no indications that initial vocabulary size predicts longevity, but it remains to be shown whether other elements of a new-comer's initial identity might prove predictive. And while in some Subreddits novices' posts do indeed converge on the prototypical veteran post, in other Subreddits they diverged from it instead.

Larger scale classification tasks can sometimes mask the presence of multiple conflicting patterns like these. By looking at the behavior of individual Reddit users, I

can unveil patterns that would otherwise remain hidden under a cloud of background noise. Variation between people is not the only source of noise I contend with, and so I use a variety of statistical methods to strengthen the signal. These methods, described in detail for each experiment, generally fall into three categories:

- Apply a filter to focus on relevant vocabulary over common low-content words
- Apply windowing techniques to reduce variance from post length
- Apply user aggregation where it won't interfere with analysis

This chapter consists of two experiments. The first looks at individual trajectories in the abstract setting of vocabulary size. The second looks at generating a prototypical "veteran" word choice distribution called a Token Frequency Distribution Dictionary (TFDD), and examines individual trajectories toward or away from these TFDDs.

5.1 Vocabulary size

To find evidence of learning processes through language production, I look for markers of fluency. Within an informal language community, fluency consists of understanding and mastery of the community and its language norms. One common marker of fluency among foreign language learners is a student's vocabulary. As Read puts it:

> "Scholars who carry out research on vocabulary size are not claiming that learners can meet their language needs simply be increasing the number of words they know. Obviously reading comprehension involves grammatical competence, an understanding of how texts are organised, background knowledge of the subject matter and other abilities in addition to vocabulary knowledge. Rather, the point is that adequate knowledge of words is a prerequisite for effective language use. Learners whose vocabulary is below a certain threshold level struggle to decode

> the basic elements of a text, to the extent that they find it hard to develop any higher level understanding of the content."[43]

Since novices in a social media community come in already speaking the formal language (English, for example), one marker of fluency in an informal language community is then the student's community-specific vocabulary. In other words, I seek evidence of their understanding of and proper use of jargon.

A novice's true vocabulary consists of words they understand and can produce, and thus consists of two elements: their received vocabulary and their generated vocabulary. Goulden et al refer to this distinction as between receptive and productive knowledge[25], and Read describes it as the difference "between being able to recognise a word when you hear or see it and being able to use it in your own speech or writing"[43]. A novice's received vocabulary consists of the words that they can accurately define and recall, such that when primed by the words' use in correct context by other community members, the novice can respond appropriately. A novice's generated vocabulary consists of the words that they can generate on demand when given an appropriate context. Most standardized tests exclusively measure received vocabulary, while when students in a language class complain that they are unable to find the words to say the ideas that they want to convey, they are complaining about limitations of their generated vocabulary.

In my online communities, unlike in a controlled setting such as a classroom, I do not have the ability to survey community members for their received vocabularies. Nor do I have access to experts who can generate true definitions of jargon against which to compare novice usage. In fact, received vocabulary is largely inaccessible in my setting. In contrast, generated vocabulary is plainly available. My dataset includes every word produced within the community by each member and significant

metadata around the context of each produced word. While I cannot ask whether a novice truly understood each word they used, I can construct an approximation by considering whether they used each word at all. The breadth of a community member's generated vocabulary can serve as a marker for their ability to fluently speak the informal language of their community. It is worth noting that, while detected change in vocabulary sizes may indicate a learning process, not all changes in vocabulary size result from learning processes and not all learning processes will cause changes in vocabulary size.

Given generated vocabulary size as a marker of fluency, I ask several questions about its nature and function as a marker. We may intuitively assume that as a novice learns to speak the informal language of their community, their generated vocabulary within the community will grow larger. Does this actually occur? If so, we would expect to see that vocabulary sizes for veteran posts would be on average larger than those for novice posts. We may also intuitively assume that changes to the vocabulary sizes of sub-populations within the community are driven by changes within those sub-populations. But are changes to individuals driving changes in the data (suggesting a learning process), or are changes driven by which individuals are included within those sub-populations (suggesting a process of weeding out poor fits)? If learning is not occurring and a filtering process is instead dominant, we would expect that initial elements of language production would predict longevity, where longevity is defined as the total number of posts an individual produces within the community.

5.1.1 Experiment design

I consider user language production of text posts and comments within the community, ignoring picture content, link targets, votes, and post titles. I apply NLTK's

default word tokenizer and remove non-alphabetic tokens such as punctuation and numeric tokens. I retain duplicates, however, for this analysis.

Since many posts are individually too short for a reasonable analysis of generated vocabulary size on their own, I combine multiple sequential posts from the same user together such that each unit of analysis contains at minimum a threshold number of 1-gram (word) tokens. I thus generate a vocabulary size estimate for each "age" using the same sliding window approach I use in classification experiments. To calculate the total tokens for a given age, I begin with the user's post at that age and append subsequent posts until the total tokens in the extended post exceed my threshold window size of 1000 tokens. The resulting extended posts are thus on average slightly larger than 1000 tokens, with outliers theoretically unbounded but in practice up to about 2000 tokens. Two alternative approaches would be to either cut off text mid-post to hit exactly 1000 tokens for each data point, or to scale over-long extended post results down so that a 2000-token extended post which scored 400 unique tokens would be assigned a normalized score of 200. The former alternative approach would skew results by prioritizing text produced earlier in a post, which I avoid in the approach I take. The latter suffers the more serious flaw that a post which is twice as long does not statistically include twice as many unique tokens. Applying this normalization would require calculating a normalization curve which is likely different within each Subreddit community, and likely unique to each individual. In practice, while I observe outliers of up to 2000 tokens, the average length of a post within each of my three primary communities (r/ArtFundamentals, r/productivity, and r/vim) is between 50 and 150 tokens. I therefore ignore the variations in vocabulary size that result from variations in augmented post length.

5.1.2 Results

Figure 5.1 shows individual trajectories for members of each of my primary communities of interest. Sample community members were selected according to longevity, defined as the maximum age at which I can generate an extended post containing at least 1000 total tokens. The top row consists of the highest longevity member of each community, with rows 2 and 3 showing randomly selected members whose longevity was in the middle (row 2) or low but enough for reach veterancy (row 3). While each pattern looks superficially distinct and noise levels are quite high, several distinct patterns can be seen repeated throughout these results. In order to aid visual inspection of these plots, I apply LOWESS (locally weighted scatterplot smoothing)[13] moving regression curves to each subfigure. LOWESS is selected due to its function-agnostic nature, meaning I do not need to assume anything about the data in order to produce the regression, and its strength in dense scatterplots. Thus, the LOWESS curve is most useful where dense datapoints would otherwise render a subfigure unreadable. This strength does come with a corresponding weakness where data is sparse, but the regression is provided at these points as well for completeness even though it does not necessarily match potential patterns visible through visual inspection.

Over the long term, all three of the most prolific posters produced a wide range of individual data points, suggesting that even a 1000 word window may be too narrow for blunt measurement of generated vocabulary size. While the LOWESS moving averages of all three remain within a range of 400-460 unique vocabulary tokens per 1000, there does not appear to be much consistency in long-term trajectories.

During posters' formative early posts, however, I identify two common patterns that seem to appear for many posters. The first pattern consists of an initial rise followed by a drop. This pattern can be seen most clearly in the posts of *TheEncour-*

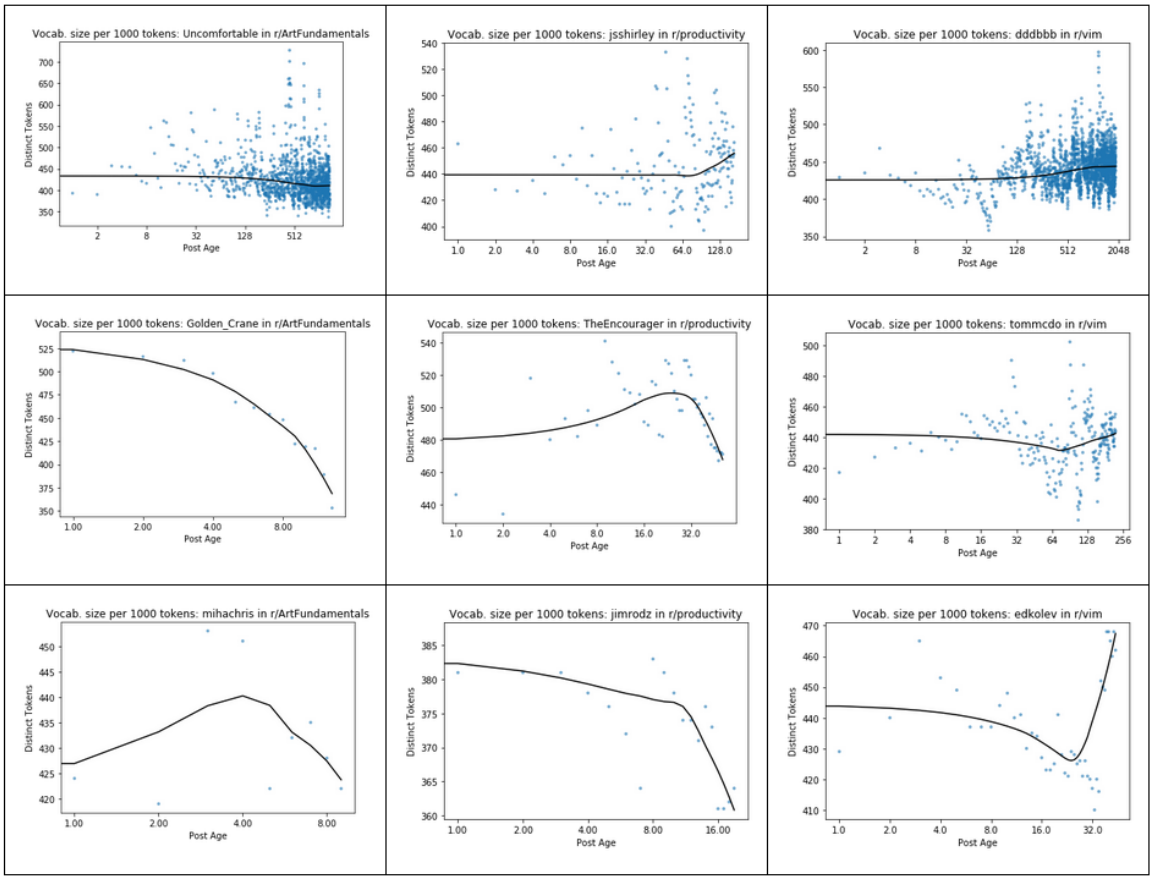


Figure 5.1: Nine examples showing individual trajectories of generated vocabulary sizes against post age in r/ArtFundamentals (left), r/productivity (center), and r/vim (right). Age is displayed using a log scale to highlight the formative initial posting period.

ager in the center of Figure 5.1, but is more subtly present in *tommcdo* and *edkolev* on the right and bottom right, as well as perhaps *mihachris* on the bottom left. At a point somewhat after we would expect the user in question to achieve veterancy, vocabulary size rebounds. A second pattern, seen in the posts of *dddbbb*, *Golden_Crane*, and *jimrodz* in my sample, displays what appears to be the second half of the first pattern. Posters' vocabulary sizes begin at a local peak, decay, and then after sufficiently many posts rebound. I speculate that these patterns may in fact mask the same phenomenon: Novices first undergo a phase of appropriation during which they begin to use local jargon. This increases their vocabulary size. Subsequently, as veteran community members react to the novices' often incorrect use of topic-relevant terms, the novices respond by reducing incorrect usage and perhaps reducing overall usage of jargon. This results in a rise and fall "arc" of vocabulary size. I further speculate that novices may overcorrect at both extremes, which would explain the rebound in vocabulary size seen around the 30-60 post range in several parts of my sample. One interesting aspect of this model is that the same arc should appear with novices' use of individual words as they learn, overuse, and reduce incorrect usage of each word in turn.

While individual trajectories display considerable variation within each Subreddit, when we treat the community overall as a single combined individual and track its generated vocabulary, we see a steady rate of roughly 420 unique tokens per 1000 total tokens within all three communities as in Figure 5.2. Individual variation appears to average out and does not indicate a single broader pattern within any of these communities. It is, however, possible that two or several conflicting patterns exist that largely balance each other. If the arc and half-arc patterns I describe above are present in roughly equal quantities, for example, they could cancel each other out in the aggregate analysis.

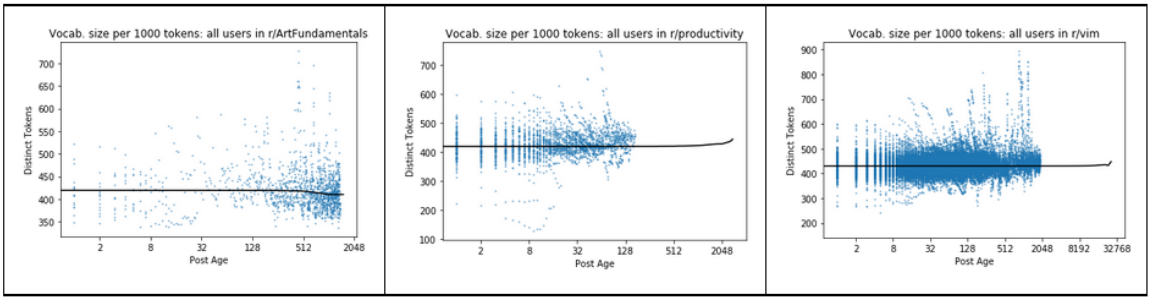


Figure 5.2: Generated vocabulary sizes for all users across all ages in r/ArtFundamentals (left), r/productivity (center), and r/vim (right) plotted together.

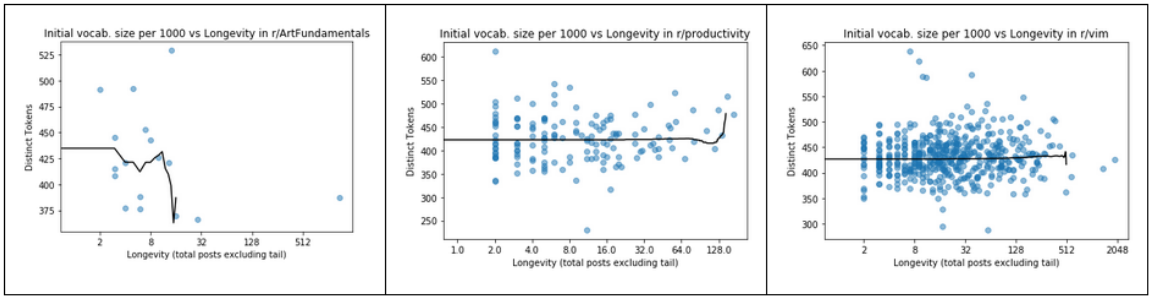


Figure 5.3: Individuals' age 1 generated vocabulary sizes compared against their longevity (maximum age of post with at least 1000 remaining tokens). Measuring longevity rather than maximum age excludes the post "tail" with insufficient remaining text for analysis.

While evidence points to a complicated answer to my first research question, plotting the first generated vocabulary data point from each community member against their longevity in Figure 5.3 shows that there is no correlation between the two. This is consistent with the hypothesis of an "arc" of vocabulary use, where novices begin by learning to imitate jargon and then subsequently learn to limit their incorrect usage before eventually bouncing back to equilibrium, given that individuals may begin their posts already at the peak of the arc. Additionally, while only a half or a third of posters within a given community may produce enough posts to be considered "veterans" in the community, only 10-20% of each community produced enough text to fill a single 1000-token window. I speculate that there may be a possibility to predict longevity with a similar approach, but it would require developing methods to calculate generated vocabulary sizes using less total text.

I might like to use these results to rule out the hypothesis that low initial generated vocabulary size predicts low longevity, but these data do not support so strong a claim. I see no evidence of a weeding out process, but it remains possible that tourists to a community tend to be weeded out before they are able to generate 1000 tokens of text material. Further experiments would be required to determine whether this is the case.

5.2 Token frequency distributions

In this study, I consider the trajectories of individuals' language production with respect to the full distributions of words used by veterans within their community (see RQ6). This framing supposes that the community holds a set of shared language norms that veteran members largely observe, and which novices must learn. This is an intuitive hypothesis, shared by prior work in the field[17, 41].

As in my previous studies, I consider language production to consist of text posts and comments produced within the community, ignoring picture content, link targets, votes, and post titles. I continue also to focus on user vocabulary, using the same bag of words model and tokenization process. Here, I explicitly develop prototypical vocabulary distributions for veterans of each community analyzed and measure distance of individual posts from those distributions. Danescu-Niculescu-Mizil et al use a similar approach to measure distance between posts and a prototypical distribution for the entire community, with cross-entropy as the distance measurement[17]. Nguyen and Rosé use KL-Divergence as their distance measurement to compare users' posts against their posts from the first week after forum registration, and against their posts after their first year post-registration[41]. I call such a distribution a Token Frequency Distribution Dictionary (TFDD) and refer to the mathematical proximity of a post to a given distribution as its *score*. The score that I use (formally defined in Equation 5.1 below) is similar to cross-entropy; the primary difference being that cross-entropy takes the log of individual elements, while I do not. In this setting, any of these three divergence measurements should perform similarly by allowing us to compare linguistic proximity and differentiate between text production that is like or unlike a prototypical distribution. Since word choices are not independent variables, however, no distance measurement can provide more than this relative comparison. For example, when both halves of a two-word phrase appear in a post, we effectively give double credit in the calculation. This is not an issue when trying to identify typical and atypical posts, since the granularity is large and all posts are scored according to the same rubric, but it would interfere with attempts to draw information-theoretical conclusions from the raw scores.

5.2.1 TFDDs

Each TFDD is defined on a set of posts and consists of a mapping from tokens (words, using the NLTK default word tokenizer as a black box) to probabilities, indicating the chance that a uniformly randomly selected post in bag of words form from the set will contain that particular token.

Recall Figure 4.1 shows a sample of my data handling. First, I apply NLTK's default word tokenizer. I then remove duplicates, and finally remove non-alphabetic tokens and parts of tokens.

Consider a hypothetical Subreddit community r/A. Implicitly, there exists some *ground truth* distribution which predicts the vocabulary used by members of r/A within their Subreddit. Every possible token has some true probability of occurring within a randomly selected r/A post. Note that these probabilities, while accurate, are not independent. It is highly likely that some tokens appear together with high frequency, and that others appear together only rarely. In the example post above, the semantically related "start" and "end" tokens appear together, as do the syntactically related "out" and "of" tokens. Nevertheless, this ground truth distribution exists. If we have access to the ground truth distribution, we can use it to estimate the likelihood that a given post comes from r/A. We cannot measure the probability with perfect accuracy, since the distribution only gives us the probabilities for each token assuming they are independent, and we know they aren't actually independent. I do not have direct access to the ground truth distribution either, but a TFDD built from a sample of posts from r/A can approximate the ground truth distribution. Thus, given a post and a TFDD, I can construct a measure of the distance between the post and the TFDD, approximating the probability that someone posting in r/A could have produced the given post.

Mathematically, I define a *word score* measuring the similarity of a post to a reference TFDD D in terms of 1-gram (word) tokens. Each token in the bag of words model of the post has an associated probability in D. Since all of the probabilities are numbers between 0 and 1, a natural way to measure the probability of the post as a whole is to combine the probabilities of the individual words. However, this results in short posts having an inflated word score, and long posts having a deflated word score. I normalize by post length, with the resulting formula, which provides the arithmetic mean of the TFDD scores of the word-tokens in the post.

$$\text{wordScore}(\text{post}, D) = \frac{1}{\#\text{ words in post}} \sum_{\text{word} \in \text{post}} D(\text{word}) \tag{5.1}$$

My wordScore statistic is also similar to the "cross product sum" used in cryptography at the individual character level[3, 34].

5.2.2 Experiment design

For this study, I construct one TFDD representing the language model of veteran posters with respect to word choice. I include all posts with age meeting or exceeding my veterancy threshold. For example, using my default veterancy threshold of 10 posts, only the final three posts from a user with 12 posts total would have been included.

I consider each poster who eventually becomes a veteran, and examine their sequence of posts. I calculate the wordScore of each post against the veteran TFDD, including both their novice and veteran posts for comparison. I exclude from consideration tourist posters who do not eventually make the veterancy threshold number of posts within the Subreddit. Table 5.1 shows the number of included and excluded posters for each of my primary Subreddits of interest. Given the small sample size

	r/ArtFundamentals	**r/productivity**	**r/vim**
Total posts	3,308	30,501	68,601
Tourist posters	447	10597	6169
Veteran/novice posters	39	466	1381
Veteran x2 posters	10	139	655

Table 5.1: Population statistics for the r/ArtFundamentals, r/productivity, and r/vim Subreddits. Counts displayed are calculated assuming a veterancy threshold of 10 posts. "Veteran x2" posters have at least 20 posts and are also included in the "Veteran/novice" line. e.g. There are 1381 total veteran posters in r/vim, of whom 655 have produced at least 20 posts.

(N=39) for r/ArtFundamentals, I opt not to conduct this experiment on that community.

Measuring similarity to the veteran distribution across all words in the corpus leads to a large amount of noise, which I address in a second experiment by filtering to look specifically at token frequency similarity among words that are topic-relevant, as defined in the previous study. The relative frequencies of words as compared to their prevalence in the background Reddit data follows an exponential decay pattern in all Subreddits I examined, with inflection points variable but usually between factors of 10 and 50. As in my classification experiments, I select a factor of 4 threshold for relevance.

5.2.3 Results

In my first TFDD experiment, I generate veteran poster TFDDs for two Subreddits: r/productivity and r/vim. Recall that the former community focuses on trading productivity tips, and tends to contain no human-identifiable jargon, while the latter hosts discussions about the text editor Vim and is characterized by a high and human-noticeable degree of jargon. I plot the posts of each novice poster within each community, comparing the divergence of their first, second, third, etc post from the

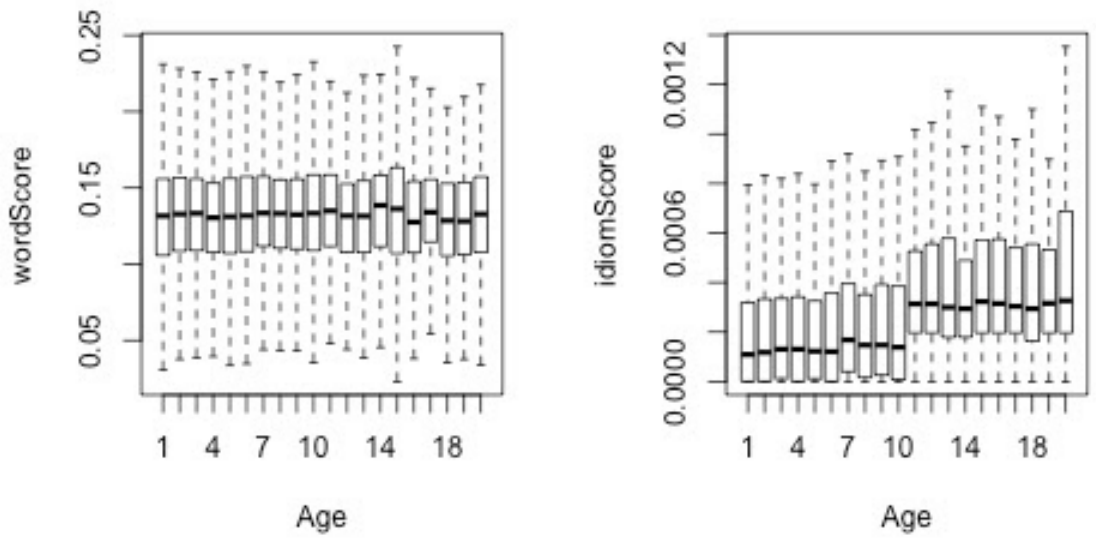


Figure 5.4: An example of wordScore (left) and idiomScore (right) values by age across all posters active in the r/productivity Subreddit, showing median score and first and third quartiles for ages 1-20.

veteran distribution. I set a veterancy threshold at 10 posts, but include in the figures below the wordScore values for posts above the veterancy threshold for comparison purposes. My initial hypothesis was that novices would adapt their language production through feedback and wordScore values would rise as they gained experience, indicating that novices used language closer to the prototypical veteran distribution, on average, as they learned to operate within the community better.

Despite theoretical reasons that I might have expected to see change with this model, my word choice language model is unable to detect learning. Experiments with different veterancy thresholds for particular Subreddits also reveal no changes in the wordScores. Notably, the average wordScore values remain consistent across a range of ages that includes the very ages which were used to generate veteranWordScore, revealing that my wordScore analysis cannot distinguish novice posts from veteran posts.

For completeness, I also run this same analysis with ordered triplets of tokens. These triplets approximate idiom, as unusually common phrases should appear as distinct probability spikes in a TFDD. My idiomScore values, derived using identical methods except for as noted using 3-grams instead of 1-grams, do differ by a factor of 2-3 between novice and veteran posts. However, this difference is explained by the sparsity of my 3-grams data. This data is littered with "unique" 3-grams, each essentially equivalent. However, since the veteran posts are used to generate the veteran TFDD that I compare against, unique 3-grams from these posts increase that post's idiomScore while unique 3-grams in novice posts decrease it. This is a statistical artifact, and does not imply a meaningful change in an individual's posting habits. Within each category (novice, veteran), idiomScore also remains constant with respect to age, and so I am unable to detect learning through this approximation of idiom. A typical pattern comes from the r/productivity Subreddit (see Figure 5.4), where the average and median wordScore for each age remain close to 0.1 across all ages measured, while the idiomScore values have means and medians close to 0.0002 for ages below the veterancy threshold, abruptly rising to close to 0.0004 for ages over the veterancy threshold.

In my second TFDD experiment, I restrict input tokens to look at only topic-relevant vocabulary, tokens which occur with at least 4 times greater frequency within the Subreddit as within Reddit as a whole and which occur at least 4 times total within the given Subreddit. The latter restriction is imposed to filter out "unique" tokens that happened to occur within the Subreddit and nowhere else, but are otherwise meaningless.

I detect two distinct patterns in TFDD similarity with age, shown in Figure 5.5. In r/productivity, we see clear convergence toward the veteran word distribution as poster age increases. However, in r/vim we see the opposite pattern, where increased

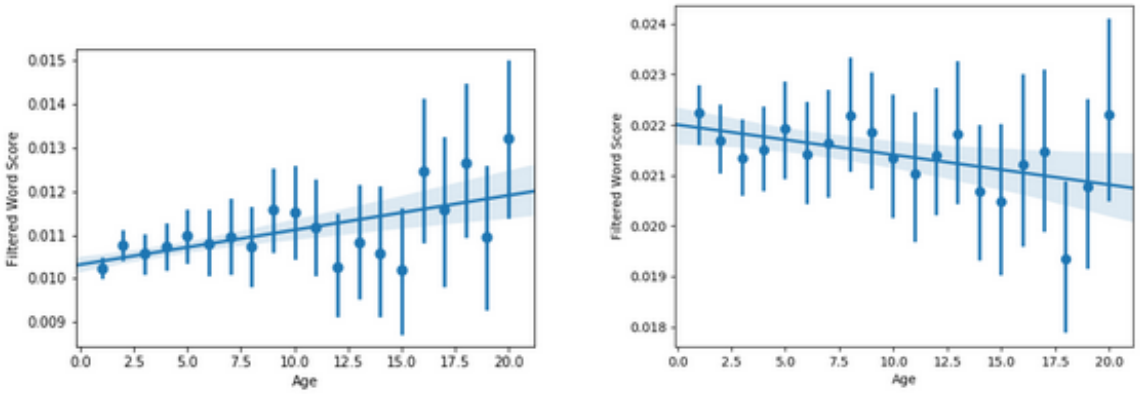


Figure 5.5: WordScores for topic-relevant vocabulary to the veteran TFDD across post age in r/productivity (left) and r/vim (right).

age of post led to a divergence from the veteran word distribution. The effect of age on word frequency similarity is significant with $p < 0.01$. The increased variance with age is likely explained by the decreased sample size as age increases. While tourists are excluded from this analysis and so every poster in the sample has a post at each age up through the veterancy threshold, only 30% of veteran r/productivity posters and 47% of veteran r/vim posters produced a full 20 posts. (See Table 5.1.)

5.3 Comparing Subreddit and Reddit background distributions

My results from the TFDD experiments suggest that prevalence of community-specific jargon within a Subreddit's discourse may have a significant impact on how people learn to speak the informal language of that community. Recall that I selected r/productivity and r/vim as examples of low- and high-jargon communities. However, that selection was based on human inspection and intuition. When I look at r/productivity, I do not see words that I recognize as jargon. I see words that are largely familiar

and common. When I look at r/vim, I see words that I do recognize as jargon, and which I would not expect to see elsewhere. In this section, I present an approach to quantifying the jargon-heaviness of different Subreddit communities.

My method is informed by prior work from Vilhena et al on mapping academic fields and the "cultural holes" (their term) that make it more difficult for readers from adjacent fields to consume material within the given field[59]. They propose a communication efficiency measure

$$E_{ij} = \frac{H(X_i)}{Q(p_i||p_j)} = \frac{-\sum_{x \in X} p_i(x) \log_2 p_i(x)}{-\sum_{x \in X} p_i(x) \log_2 p_j(x)} \tag{5.2}$$

to quantify how inaccessible field i is to readers familiar with field j. In this formula, $H(X_i)$ is the Shannon entropy of field i, defined over the distribution X_i of relevant communicative tokens. Vilhena et al focus on scholarly phrases, but note that the formula does not depend on which tokens are relevant. West and Portenoy apply the same analysis scheme to comparing usage of mathematical jargon between academic papers[61]. $Q(p_i||p_j)$ is the cross-entropy of the i and j distributions, defined as the sum of the Shannon entropy of i plus the Kullback-Leibler Divergence (KL-Divergence) of i from j. In the result, bounded in $(0, 1]$, lower scores indicate a greater degree of *in*accessibility. Intuitively, at $E_{ij} = 0$ a reader used to field j would have to look up every scholarly term they encountered in a paper from field i. We can also use $1 - E_{ij}$ as a relative distance measure.

Unfortunately, using E_{ij} to measure jargon heaviness normalizes away the internal complexity of a Subreddit's informal language. In my setting, rather than measuring accessibility of one field i against the backdrop of all other fields j, I am interested in measuring the divergence of various Subreddits from a single background sample representing Reddit as a whole. I could replace the single background distribution with

a collection of individual alternative Subreddit distributions, but a simpler solution is to use the part of E_{ij} prior to the normalization I want to avoid. This leads me to KL-Divergence, which has been used in related work by Nguyen and Rosé, where it filled the same role as my sum of WordScores[41].

Using KL-Divergence, I can approximate the divergence of an individual community's language norms from those of another community, such as the Reddit background. However, this approximation comes with several caveats. First, as I explain at in Section 5.2, the interdependence of tokens within the language norms of a community means that any statistical measure I generate must either factor in token interdependence or accept that we get a rougher approximation. I opt for the more straightforward calculation, with the understanding that large scale differences between should remain unaffected but smaller ones may be obscured. Second, absent a more concrete definition of jargon, the magnitude of any distance measurement between community language distributions is meaningful only as relative magnitude. In other words, I can indicate which Subreddits diverge *more* from the Reddit baseline and which diverge *less*. For this analysis, I apply a slightly modified version of Kullback-Leibler Divergence that I refer to as Pointwise Kullback-Leibler Divergence (PKLD).

KL-Divergence in a discrete distribution space is formally defined as the extra information required to encode a specific population P using a coding scheme designed for a background population Q.

$$D_{KL}(P, Q) = \sum_{x \in X} P(x) \log \left(\frac{P(x)}{Q(x)} \right) \tag{5.3}$$

D_{KL} is an asymmetric measure, and you may note that the sum is conducted over events x in the discrete probability space X. In my setting, an event x is a possible

post in bag of words form, and the probability space X encompasses all possible posts in bag of words form. This appears to be an excellent measure of the surprisingness of posts in P (a Subreddit) relative to the background distribution Q (a random sample of Reddit posts). Unfortunately, these distributions are extremely sparse. Except for extremely short posts, it is unlikely that a particular set of tokens will occur together as the bag of words form of a post more than once in the entire dataset. An event which occurs once within P and zero times within the background sample Q does not tell us much about the difference between P and Q, and D_{KL} is dominated by this type of event. Note also that $P \not\subset Q$. Either including or excluding P from the sample of background Subreddits would result in a valid measure; I opt to exclude P in all cases.

Rather than summing over events (possible posts), I measure Pointwise KL-Divergence by summing over tokens in the dictionary of tokens which appear within P. PKLD is a novel measurement. The fact that $\sum_{x \in X} P(x) = 1$ means KL-Divergence is bounded in $(0, 1]$. Since multiple tokens appear within the same post, PKLD could be higher in settings with longer average posts and higher average tokens per post. To account for this, I scale by the average tokens per post.

In order to be able to practically run the analysis, my Reddit background distribution is calculated using a random sample of Subreddits, each with at least 1000 posts and no more than 300,000 posts. The results shown below use a background sample of 160 Subreddits, of which the largest single contributor is r/Warhammer40k constituting 10.5% of the background sample posts. The background sample has an average post length of 27.5 tokens. Since even this sample may be vulnerable to jargon of the background Subreddits, I also skip tokens that appear more frequently in Q than in P. This is reflected in the formula by taking the maximum of 0 and $\log(P(t)/Q(t))$, which zeros out terms where $P(t) < Q(t)$. Conceptually, these are

tokens that relate to topics not discussed within P, or are jargon belonging to one or more Subreddits in Q.

The resulting formula, with token t in the dictionary σ and recalling that both $P(t)$ and $Q(t)$ are scaled by the average length of a post in P and Q respectively, is:

$$PKLD(P,Q) = \sum_{t \in \sigma} P(t) \max(0, \log\left(\frac{P(t)}{Q(t)}\right)) \tag{5.4}$$

I calculate the PKLD scores for r/ArtFundamentals, r/productivity, and r/vim. These results, shown in Table 5.2, corroborate my intuition that r/productivity is a low-jargon community while r/vim is characterized by higher amounts of jargon. r/ArtFundamentals also scores highly in PKLD. I also show the top 10 tokens contributing to the PKLD of each Subreddit. Unsurprisingly, the name of the Subreddit appears at the top of this list for two of the three. Top contributing tokens include terms unique to the local jargon (e.g. "vimrc" or "gtd") as well as terms that are particularly frequently used within the community but likely well understood outside of it (e.g. "lesson" or "task").

In Figure 5.6, I compare PKLD scores with Subreddit size for each of the Subreddits used for the background sample Q in my PKLD calculations. I exclude each Subreddit from the background while calculating its own PKLD. This comparison reveals a slight negative correlation between Subreddit size and PKLD score, with a Pearson correlation coefficient of -0.176. Overall, we can see that r/vim lands on the high end of PKLD scores for its size, while r/productivity lands on the low end for its size. While r/ArtFundamentals scores very close to r/vim, the former is perhaps only slightly above average in jargon heaviness for its size.

The PKLD approach to estimating jargon heaviness of a social media community remains experimental, especially since the underlying concept of jargon heaviness is

	ArtFundamentals		productivity		vim	
Avg. post length		66.47		33.71		31.83
PKLD		0.9007		0.4680		0.8972
Top 10	lesson	0.029	productivity	0.0074	vim	0.063
	ellipses	0.020	tasks	0.0067	vimrc	0.015
	drawing	0.019	evernote	0.0056	plugin	0.011
	lines	0.013	app	0.0055	use	0.0089
	boxes	0.013	productive	0.0046	plugins	0.0088
	draw	0.012	task	0.0043	lt	0.0078
	homework	0.0095	work	0.0038	emacs	0.0071
	ellipse	0.0095	calendar	0.0033	file	0.0060
	pen	0.0094	gtd	0.0033	gt	0.0059
	organic	0.0093	day	0.0033	gvim	0.0052

Table 5.2: Jargon-heaviness ratings of r/ArtFundamentals, r/productivity, and r/vim. Also included are the individual ratings of the top 10 most topic-relevant words from each community.

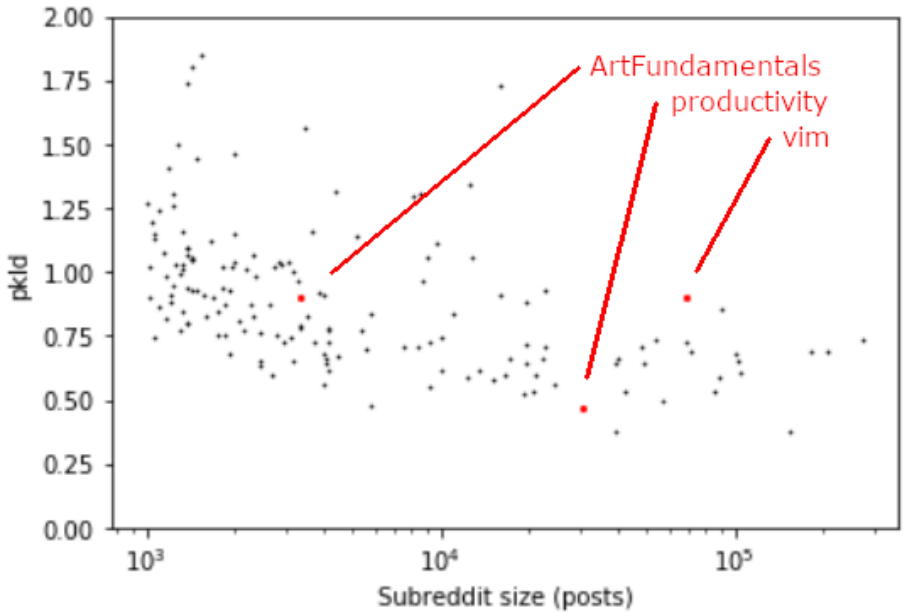


Figure 5.6: Comparison of Subreddit size and calculated Pointwise Kullback-Leibler Divergence scores across all 160 Subreddits used in the random sample. Three Subreddits of interest (not included in the background sample) are shown in red. This figure excludes 7 outlier Subreddits with PKLD scores between 2 and 8.

not formally defined. These results match those that human observers may reach intuitively following the concept of jargon heaviness directly, which suggests that PKLD is at least relatively close to the target concept. Further research into jargon-use in social media communities would need to develop a framework for thinking about what constitutes jargon and its use, which would in turn be used to refine the PKLD measurement to best fit the desired framework. For this study, the framework of interest is to provide a statistical analysis which includes all tokens and can provide non-anecdotal evidence for variable rates of jargon use in different Subreddit communities. PKLD fulfills this role, and additionally highlights that communities may naturally tend to reduce their jargon-heaviness as they grow. Whether this is a result of the effects of community growth, a tendency for low degrees of jargon to cause growth, or simple correlation remains to be seen.

5.4 Thoughts and future directions

In r/productivity, I find the expected pattern of novices increasing their use of local jargon over time. This relates to a learning process where novices appropriate vocabulary as they encounter it through interactions within the community. Note that the appropriation, while present, is not very strong. With a veterancy threshold at 10 posts, a novice only increases their topic-relevant vocabulary use by about 10% before reaching the veterancy threshold. Above that threshold, appropriation may continue, but the poster's own vocabulary use is also included in the veteran TFDD and so we cannot tell the difference between a veteran user appropriating additional local jargon and that same veteran user becoming more fixed in their own personal idiosyncratic language use. Given Danescu-Niculescu-Mizil et al's findings that "old" members tend to become more linguistically conservative and ultimately get left be-

hind by the community,[17] it seems at least plausible that veteran users becoming fixed in their specific personal language patterns is a factor.

In r/vim, the opposite pattern appears. Filtered word scores are much higher across the board, providing additional evidence of a community with greater reliance on specialized vocabulary. However, the highest rates of relevant word use appear among novices, and filtered word scores decrease with poster age. I speculate that this pattern might be evidence of novice overuse of topic-relevant vocabulary. In settings where novices know jargon is important, they may know what that jargon is but not how to use it appropriately. Even the use of two relevant terms where one would serve increases wordScore values, and of course my analysis has no way to indicate whether a term was used appropriately or not. As in the previous case, the adaption is relatively subtle, with a 10% drop in topic-relevant vocabulary use between a novice's first post and the veterancy threshold of 10 posts.

Combining my TFDD analysis, my vocabulary size analysis, and Danescu-Niculescu-Mizil et al's "life stage" analysis, a coherent picture of social media user language learning begins to emerge: Initially, a user undergoes a period primarily characterized (in terms of vocabulary) by the acquisition of new jargon. The user may begin this linguistic appropriation prior to contributing new posts to the community, and may in fact be likely to conduct this phase while "lurking" if the jargon is obvious. I speculate that when a user knows they do not know the right words, they may as a result be more reluctant to contribute, and the pattern we see in r/vim exemplifies this reluctance with new posters waiting until they can produce a higher quantity of local jargon than veterans do before daring to post.

After the period of appropriation, a user subsequently undergoes a period of linguistic winnowing. As I find through vocabulary size, the rate of jargon use decreases. At the same time, Danescu-Niculescu-Mizil et al see the distance from the dominant

language model of that month decreases. Together, this suggests that the novice is decreasing their over-use of local jargon and approaching the community norms.

At some point thereafter, the user achieves veterancy within the community. While we can all continue to learn throughout our lives, it appears that linguistic change decreases with veterancy. Vocabulary size recovers from its low point at the end of the winnowing phase, suggesting that the user may no longer be as responsive to feedback from other community members, or equivalently may feel more comfortable expressing themself within the community through idiosyncratic word choice. Danescu-Niculescu-Mizil et al suggest the rate of change of the user's language also drops, suggesting the same. My TFDD analysis relates user word choice to the word choices of prototypical veterans including the veteran being analyzed, but it seems probable that a future analysis which compared veteran word choices to those of *other* veteran posters might be able to confirm the hypothesis that veteran community members' language diverges from the community norms with increased age.

Altogether, these three phases paint a picture of novices learning the language of the community through distinct phases of appropriation and winnowing, and veterans speaking that language but notably not doing much learning of their own. Qualitative analysis of individual learners could shed additional light on markers for each phase and transition points between them. This overarching pattern is one hypothetical learning pattern, and there is no guarantee that community members will experience every (or indeed any) part of it, nor that users will engage with each piece for the same duration. It remains a pattern, however, that we can now specifically look for within quantitative and qualitative data in order to help make sense of online language use.

Chapter 6

Conclusions and Future Work

In these studies, I set out to map features of informal learning in Reddit and StackOverflow communities. My results form initial steps toward that understanding, but much work remains to be done.

Through the lens of vocabulary, we can see novices adapting their language use patterns over time. A feed-forward neural net classifier can distinguish novices from veterans, demonstrating the presence of language changes. Restriction to topic-relevant vocabulary surprisingly does not seem to improve classification, while text sample size is a major factor and so windowing to avoid short, confounding posts is quite successful.

In contrast, topic-relevance is critical to revealing trends in novices' frequencies of use of different words, as seen in my TFDD experiments. We see novices in a low-jargon setting steadily increasing their use of topic-relevant vocabulary, while novices in a higher-jargon setting appear to reduce their use of jargon. This is indicative of the twin challenges of appropriating specialized vocabulary and learning to restrict its usage correctly. In a setting with subtle jargon, the former challenge may dominate.

It is interesting to compare these results, derived using ordinal time, against prior results obtained using real time and life stage. Members of a medical support community displayed convergence toward a distribution of 2nd year members' posts[41]. Interestingly, this mirrors my low-jargon-setting results, despite the reasonable expectation that such a community might be a relatively high-jargon setting. Participants in two beer-rating communities saw a precipitous drop in distance from community language norms in the first third of their active time in the community, followed by a slower increase over their remaining tenure[17]. The life stage framing unfortunately obscures the initial stages that are of most interest to me, but the behavior of veterans late in their life cycle provides useful context.

In terms of the two patterns that I see in my results, because the domains I study are informal spaces where language norms exist but may not be actively studied, it is possible that veterans and experts in the low-jargon r/productivity Subreddit are unable to provide feedback which directly prompts novices with the correct terminology. Compare this scenario to a formal English classroom, where a teacher may present students with a list of vocabulary words for the week. Veteran members of a high-jargon setting such as the r/vim Subreddit or the r/ArtFundamentals Subreddit may hold a more conscious awareness of the specialized vocabulary they use, and so be able to prompt novices to use that vocabulary as well. In this latter case, the challenge of correct usage becomes dominant, resulting in an over time decrease in usage. This is reflected in the second TFDD experiment, where novice posts over-used jargon and were initially *too* like a prototypical veteran post before settling down, as well as in individuals' trajectories of vocabulary size, where the initial appropriation phase is followed by a phase where novice vocabulary size shrank.

Just as different dominant patterns seem to hold prominence in different Subreddits, it is quite likely that different conflicting or overlapping patterns can be found

within a single community. We can see this dynamic at work when comparing individuals' trajectories of vocabulary size, where both initial rises and initial drops can be seen among the highlighted examples, and the overall averages in Figure 5.2 which display no changes at all. Unsupervised clustering within veteran or within novice posts might provide some hints about significant sub-group patterns, but I suspect future breakthroughs will require analysis of non-anonymous social media settings. I believe the most significant variation to be found will be rooted in how posters' experience existing in online spaces as a result of their offline characteristics. For example, someone who has experienced social media as a source of harassment—unfortunately common among members of minoritized communities—may change their language production differently from someone who has not. While members of the communities I study are anonymous to me, they may know (or hypothesize) about each other, and linguistic markers such as pronoun use could lead to other members of their communities reacting to them differently.

One potential avenue for progress while continuing to look at anonymous social media would be to systematically identify microaggressions using Breitfeller et al's approach[6] in a Subreddit and look at the speaker and the recipient of the microaggressions. This would allow me to investigate one type of difference in experience with social media. It would be particularly meaningful if microaggressions were most commonly directed at tourists in the community, which could suggest that tourist vs novice identity is shaped in part by how community members react to a newcomer.

In future experiments I also hope to delve deeper into the interplay between novices and their ersatz mentors. I have identified distinct expansion and contraction phases, although it is likely these overlap more than my aggregate data can show, but what are the processes through which new vocabulary words are introduced to a novice, and can I track decreases in word use to feedback directed at them?

Expanding outward from vocabulary, I plan to return to examining idiom and ultimately word embeddings. Through these more sophisticated aspects of language, I hope to find elements of novice language change that are less shrouded in noise, and possibly more clearly tied to known elements of learning processes.

I finally plan to broaden my domain of study to examine dynamics of language change and learning in other social media settings. I have largely focused so far on extremely informal and uncontroversial communities. While I plan to continue to avoid hot-button issue communities, there should be a new wealth of online text data from semi-formal and formal instruction settings that sprung up in response to the COVID-19 pandemic. Identifying conversational exchanges within these new communities may allow me to apply my techniques directly to questions around learning *informal* language norms of *formal* learning communities.

In the end, this dissertation presents a novel analysis that is broadly interdisciplinary but also tractably narrow in scope. Within computational sociolinguistic analysis of informal language learning, there are myriad future directions my research can take. I have laid down several building blocks on which I hope to rest a future body of work, and I am excited to begin the next phase of my research.

Bibliography

[1] Shlomo Argamon, Moshe Koppel, James W Pennebaker, and Jonathan Schler. 2007. Mining the blogosphere: Age, gender and the varieties of self-expression. *First Monday* 12, 9 (2007). `https://doi.org/10.5210/fm.v12i9.2003`

[2] Mikhail M Bakhtin. 1986. The Bildungsroman and its Significance in the History of Realism. *Speech genres and other late essays* 10 (1986), 21.

[3] Friedrich Ludwig Bauer. 2002. *Decrypted secrets: methods and maxims of cryptology.* Springer Science & Business Media.

[4] Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit.* O'Reilly Media, Inc.

[5] Philip Bramsen, Martha Escobar-Molano, Ami Patel, and Rafael Alonso. 2011. Extracting social power relationships from natural language. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1.* Association for Computational Linguistics, 773–782.

[6] Luke Breitfeller, Emily Ahn, David Jurgens, and Yulia Tsvetkov. 2019. Finding microaggressions in the wild: A case for locating elusive phenomena in social media posts. 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).* 1664–1674. `http://dx.doi.org/10.18653/v1/D19-1176`

[7] Cody Buntain and Jennifer Golbeck. 2014. Identifying social roles in reddit using network structure. In *Proceedings of the 23rd international conference on world wide web.* 615–620.

[8] John D Burger, John Henderson, George Kim, and Guido Zarrella. 2011. Discriminating gender on Twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing.* Association for Computational Linguistics, 1301–1309.

[9] John D Burger and John C Henderson. 2006. An Exploration of Observable Features Related to Blogger Age.. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*. Menlo Park, CA, 15–20.

[10] Justine Cassell and Dona Tversky. 2005. The language of online intercultural community formation. *Journal of Computer-Mediated Communication* 10, 2 (2005), JCMC1027. https://doi.org/10.1111/j.1083-6101.2005.tb00239.x

[11] Courtney Cazden, Bill Cope, Norman Fairclough, Jim Gee, Mary Kalantzis, Gunther Kress, Allan Luke, Carmen Luke, Sarah Michaels, and Martin Nakata. 1996. A pedagogy of multiliteracies: Designing social futures. *Harvard educational review* 66, 1 (1996), 60–92.

[12] Morgane Ciot, Morgan Sonderegger, and Derek Ruths. 2013. Gender inference of Twitter users in non-English contexts. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. 1136–1145.

[13] William S Cleveland. 1979. Robust locally weighted regression and smoothing scatterplots. *J. Amer. Statist. Assoc.* 74, 368 (1979), 829–836.

[14] Michael Cole, Vera John-Steiner, Sylvia Scribner, and Ellen Souberman. 1978. Mind in society. *Mind in society the development of higher psychological processes. Cambridge, MA: Harvard University Press* (1978).

[15] Maral Dadvar, FMG de Jong, Roeland Ordelman, and Dolf Trieschnigg. 2012. Improved cyberbullying detection using gender information. In *Proceedings of the Twelfth Dutch-Belgian Information Retrieval Workshop (DIR 2012)*. University of Ghent.

[16] Cristian Danescu-Niculescu-Mizil, Lillian Lee, Bo Pang, and Jon Kleinberg. 2012. Echoes of power: Language effects and power differences in social interaction. In *Proceedings of the 21st International Conference on World Wide Web*. ACM, 699–708. https://doi.org/10.1145/2187836.2187931

[17] Cristian Danescu-Niculescu-Mizil, Robert West, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. No country for old members: User lifecycle and linguistic change in online communities. In *Proceedings of the 22nd International Conference on World Wide Web*. ACM, 307–318. https://doi.org/10.1145/2488388.2488416

[18] Brian Dean. 2021. Social Network Usage & Growth Statistics: How Many People Use Social Media in 2021? https://backlinko.com/social-media-users#most-popular-social-media-platforms. Accessed: 2021-05-24.

[19] Christopher P Diehl, Galileo Namata, and Lise Getoor. 2007. Relationship identification for social network discovery. In *AAAI*, Vol. 22. 546–552.

[20] Jacob Eisenstein, Brendan O'Connor, Noah A Smith, and Eric P Xing. 2010. A latent variable model for geographic lexical variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 1277–1287.

[21] Deborah A Fields, Yasmin B Kafai, and Michael T Giang. 2017. Youth computational participation in the wild: Understanding experience and equity in participating and programming in the online scratch community. *ACM Transactions on Computing Education (TOCE)* 17, 3 (2017), 1–22. `https://doi.org/10.1145/3123815`

[22] Philip Gianfortoni, David Adamson, and Carolyn P Rosé. 2011. Modeling of stylistic variation in social media with stretchy patterns. In *Proceedings of the First Workshop on Algorithms and Resources for Modelling of Dialects and Language Varieties*. Association for Computational Linguistics, 49–59.

[23] Eric Gilbert. 2012. Phrases that signal workplace hierarchy. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*. ACM, 1037–1046. `https://doi.org/10.1145/2145204.2145359`

[24] Maria Glenski, Corey Pennycuff, and Tim Weninger. 2017. Consumers and curators: Browsing and voting patterns on reddit. *IEEE Transactions on Computational Social Systems* 4, 4 (2017), 196–206.

[25] Robin Goulden, Paul Nation, and John Read. 1990. How large can a receptive vocabulary be? *Applied linguistics* 11, 4 (1990), 341–363. `https://doi.org/10.1093/applin/11.4.341`

[26] Jonathan Grix. 2018. *The foundations of research*. Macmillan International Higher Education.

[27] Libby Hemphill and Jahna Otterbacher. 2012. Learning the lingo?: gender, prestige and linguistic adaptation in review communities. In *Proceedings of the ACM 2012 Conference on Computer-Supported Cooperative Work*. ACM, 305–314. `https://doi.org/10.1145/2145204.2145254`

[28] Andrew Gary Darwin Holmes. 2020. Researcher Positionality–A Consideration of Its Influence and Place in Qualitative Research–A New Researcher Guide. *Shanlax International Journal of Education* 8, 4 (2020), 1–10.

[29] Yiqing Hua, Cristian Danescu-Niculescu-Mizil, Dario Taraborelli, Nithum Thain, Jeffery Sorensen, and Lucas Dixon. 2018. Wikiconv: A corpus of the complete

conversational history of a large online collaborative community. *arXiv preprint arXiv:1810.13181* (2018).

[30] Joey Huang and Kylie Peppler. 2019. Studying Computational Thinking Practices Through Collaborative Design Activities with Scratch. (2019). https://doi.dx.org/10.22318/cscl2019.933

[31] Jane Im, Eshwar Chandrasekharan, Jackson Sargent, Paige Lighthammer, Taylor Denby, Ankit Bhargava, Libby Hemphill, David Jurgens, and Eric Gilbert. 2019. Still out there: Modeling and Identifying Russian Troll Accounts on Twitter. *arXiv preprint arXiv:1901.11162* (2019).

[32] L Jacques, E Carpenter, T Valley, B Alvarez, and J Higgins. 2021. Medication or surgical abortion? An exploratory study of patient decision-making on a popular social media platform. *American journal of obstetrics and gynecology* (2021). https://doi.org/10.1016/j.ajog.2021.05.011

[33] Daniel Kershaw, Matthew Rowe, and Patrick Stacey. 2016. Towards modelling language innovation acceptance in online social networks. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*. ACM, 553–562. https://doi.org/10.1145/2835776.2835784

[34] Solomon Kullback. 1976. *Statistical methods in cryptanalysis*. Vol. 4. Aegean Park Press.

[35] Jean Lave, Etienne Wenger, et al. 1991. *Situated learning: Legitimate peripheral participation*. Cambridge University Press.

[36] Adriana M Manago, L Monique Ward, Kristi M Lemm, Lauren Reed, and Rita Seabrook. 2015. Facebook involvement, objectified body consciousness, body shame, and sexual assertiveness in college women and men. *Sex roles* 72, 1 (2015), 1–14.

[37] H Richard Milner IV. 2007. Race, culture, and researcher positionality: Working through dangers seen, unseen, and unforeseen. *Educational researcher* 36, 7 (2007), 388–400.

[38] Aakanksha Naik, Abhilasha Ravichander, Carolyn Rose, and Eduard Hovy. 2019. Exploring numeracy in word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 3374–3380.

[39] Dong Nguyen, A Seza Doğruöz, Carolyn P Rosé, and Franciska de Jong. 2016. Computational sociolinguistics: A survey. *Computational Linguistics* 42, 3 (2016), 537–593. https://doi.org/10.1162/COLI_a_00258

[40] Dong Nguyen, Rilana Gravel, Dolf Trieschnigg, and Theo Meder. 2013. "How old do you think I am?" A study of language and age in Twitter. In *Seventh International AAAI Conference on Weblogs and Social Media*.

[41] Dong Nguyen and Carolyn P Rosé. 2011. Language use as a reflection of socialization in online communities. In *Proceedings of the Workshop on Languages in Social Media*. Association for Computational Linguistics, 76–85.

[42] Brian Norlander. 2018. Reddit Bot Classifier. `https://briannorlander.com/img/Reddit_Bot_Classifier.pdf`. Accessed: 2021-03-23.

[43] Felicity O'Dell, John Read, Michael McCarthy, et al. 2000. *Assessing vocabulary*. Cambridge University Press.

[44] Seymour Papert. 1972. Teaching children to be mathematicians versus teaching about mathematics. *International journal of mathematical education in science and technology* 3, 3 (1972), 249–262.

[45] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.). Curran Associates, Inc., 8024–8035. `http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf`

[46] Andrew Perrin. 2015. Social media usage. *Pew research center* 125 (2015), 52–68.

[47] Vinodkumar Prabhakaran, Owen Rambow, and Mona Diab. 2012. Who's (really) the boss? Perception of situational power in written interactions. *Proceedings of COLING 2012* (2012), 2259–2274.

[48] Vinodkumar Prabhakaran, Emily E Reid, and Owen Rambow. 2014. Gender and power: How gender and gender environment affect manifestations of power. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 1965–1976.

[49] Jenny Preece, Blair Nonnecke, and Dorine Andrews. 2004. The top five reasons for lurking: improving community experiences for everyone. *Computers in human behavior* 20, 2 (2004), 201–223.

[50] Reddit. 2020. *"About Community" text from the r/3amjokes sidebar.* https://www.reddit.com/r/3amjokes/

[51] Reddit. 2020. *"About Community" text from the r/dadjokes sidebar.* https://www.reddit.com/r/dadjokes/

[52] Reddit. 2020. *"About Community" text from the r/puns sidebar.* https://www.reddit.com/r/puns/

[53] Reddit. 2021. *"About Community" text from the r/ArtFundamentals sidebar.* https://www.reddit.com/r/ArtFundamentals/

[54] Reddit. 2021. *"About Community" text from the r/productivity sidebar.* Retrieved March 23, 2021 from https://www.reddit.com/r/productivity/

[55] Reddit. 2021. *"About Community" text from the r/vim sidebar.* https://www.reddit.com/r/vim/

[56] Maarten Sap, Gregory Park, Johannes Eichstaedt, Margaret Kern, David Stillwell, Michal Kosinski, Lyle Ungar, and Hansen Andrew Schwartz. 2014. Developing age and gender predictive lexica over social media. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP).* 1146–1151.

[57] Ruchita Sarawgi, Kailash Gajulapalli, and Yejin Choi. 2011. Gender attribution: tracing stylometric evidence beyond topic and genre. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning.* Association for Computational Linguistics, 78–86.

[58] Ryan A Smith. 2002. Race, gender, and authority in the workplace: Theory and research. *Annual Review of Sociology* 28, 1 (2002), 509–542. https://doi.org/10.1146/annurev.soc.28.110601.141048

[59] Daril A Vilhena, Jacob G Foster, Martin Rosvall, Jevin D West, James Evans, and Carl T Bergstrom. 2014. Finding cultural holes: How structure and culture diverge in networks of scholarly communication. *Sociological Science* 1 (2014), 221.

[60] Etienne Wenger. 1999. *Communities of practice: Learning, meaning, and identity.* Cambridge University Press.

[61] Jevin West and Jason Portenoy. 2016. Delineating Fields Using Mathematical Jargon. In *Proceedings of the Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL).* 63–71.

[62] Justine Zhang, Jonathan P Chang, Cristian Danescu-Niculescu-Mizil, Lucas Dixon, Yiqing Hua, Nithum Thain, and Dario Taraborelli. 2018. Conversations gone awry: Detecting early signs of conversational failure. *arXiv preprint arXiv:1805.05345* (2018).

Appendices

Appendix A

Reddit Dataset Raw Format

Column	Used	Description	Sample post
created_utc	x	Timestamp when the post was created.	1417392000
Subreddit_id		Unique id for the Subreddit where the post was made.	t5_2r9po
link_id			t3_2nvxfv
id		Unique id for the post itself.	cmhems8
author	x	Name of the account which made the post.	eror11
score_hidden		Reddit hides the upvote/downvote score of some posts.	false
body	x	Full text of the post body.	Is he the joey?

edited		True iff the text of the post was edited after creation.	false
archived		Reddit posts are archived after a period of time. Media embedded in archived posts is no longer accessible, and comments can no longer be added.	false
score		Derived measure of the "score" of a post. Involved in sorting posts when viewing.	3
name			t1_cmhems8
retrieved_on		Timestamp when the curator of my dataset retrieved this post.	1425748675
author_flair_ css_class			null
ups		Number of upvotes on the post.	3
controversiality		Derived measure of how "controversial" a post is. May affect some ways of sorting posts when viewing.	0
Subreddit	x	Name of the Subreddit where the post was made.	Guildwars2
author_flair_text			null
parent_id		Unique id for the immediate post/comment that a comment is responding to. Matches the *name* of that post.	t1_cmhe1zw
distinguished			null
gilded			0
downs		Number of downvotes on the post.	0

Table A.1: Raw data format for my Reddit dataset. Only columns marked as Used were used for my analysis. Explanations are provided for clarity but are my interpretation and are omitted for columns where I am not confident in my understanding of their use.

Appendix B

Classifier Training Errors Examples

B.1 Undertraining

In Figure B.1, I conduct 0 training runs (rather than the usual 10,000) on the r/productivity Subreddit. The classifier operates purely off of the initial random weights in the neural net. Since these weights have absolutely no relation to the data, there is no hope of any correct classification. Instead, we see two of the basic failure modes of a classifier. In the left example, the initial weights land such that the classifier guesses near equally between novice and tourist. In the middle and right examples, the initial weights cause all posts to be guessed as one category. Since the posts in this example have not been stripped of stopwords like "the" or "an", these two examples might be the result of one or several of these words being given a significant initial weight.

B.2 Training on a homogeneous language

In Figure B.2, I conduct the usual 10,000 training runs on the r/vim Subreddit. However prior to training I apply veteran, novice, and tourist tags to posts uniformly

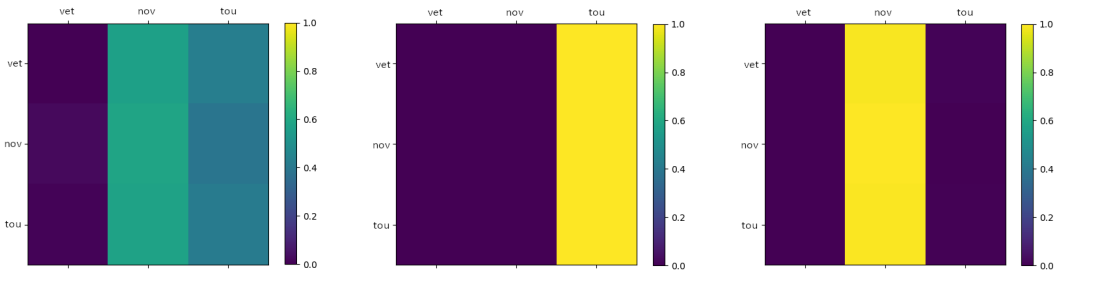


Figure B.1: Novices, veterans, and tourists in the r/productivity Subreddit, as wildly guessed at by a completely untrained classifier.

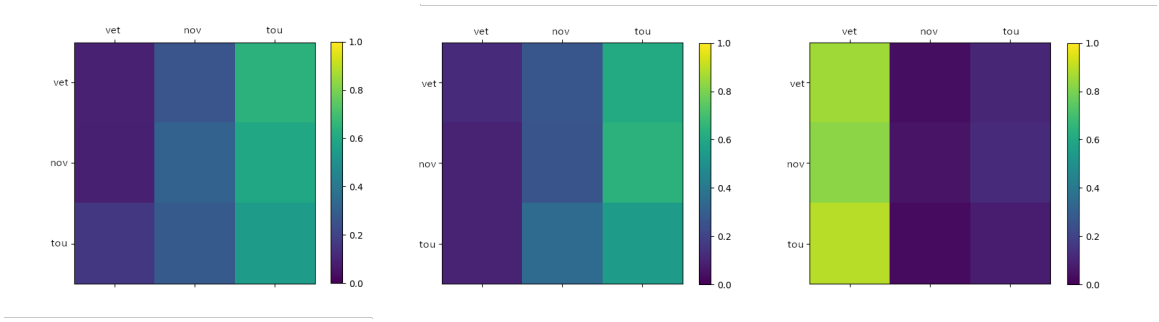


Figure B.2: Novices, veterans, and tourists in the r/productivity Subreddit, as wildly guessed at by a completely untrained classifier.

at random. Since the tags bear no relation to the underlying text, there is nothing for the classifier to correctly learn from. It may learn characteristics specific to the training set, based on the particular random choices made, but these characteristics will not appear in the test set. We see the same two basic classifier failure modes as in Section B.1. One critical difference is that in this case the classifier is separating text into distinct categories, albeit spurious ones. This makes the "vertical bar" error mode somewhat less likely, with neural net features that flag all inputs as coming from a particular source likely smoothed away. Note, however, that the third example in this mode still displays this error mode. The probability of its occurrence is reduced but not eliminated.

Chapter 7

List of Abbreviations

CS	Computer Sciences
FPR	False Positive Rate
KL-Divergence	Kullback-Leibler Divergence (See Section 5.3.)
LS	Learning Sciences
LOWESS	LOcally WEighted Scatterplot Smoothing
ML	Machine Learning, a subfield within Computer Sciences
MOOC	Massive Open Online Course
NLP	Natural Language Processing, a subfield within Computer Sciences
PKLD	Pointwise Kullback-Leibler Divergence (See Section 5.3.)
ROC-AUC	Receiver Operating Characteristic – Area Under the Curve (See Section 4.3.)
RQ	Research Question
TFDD	Token Frequency Distribution Dictionary (See Section 5.2.)
TPR	True Positive Rate