Generalizable Identity Classifiers from Self-Reporting Statements on Reddit

Aditya Kashyap

A THESIS

In

Data Science

Presented to the Faculties of the University of Pennsylvania in Partial
Fulfillment of the Requirements for the Degree of Master of Science in Engineering

2018

Prof. Chris Callison-Burch

Prof. Lyle Ungar

Abstract

This thesis presents a descriptive analysis of self-identity on Reddit. Following previous work on mental health disorders that use patterns like "I was just diagnosed with depression", we analyze self-identification on Reddit more broadly. We show which group memberships people tend to assert, and analyze which forums people disproportionately self-identify or self-distance in. To show the type of socio-linguistic studies that can be performed, we analyze linguistic traits of different groups. In addition, we define a heuristic that should be followed while building binary identity classification models on using Reddit.

Contents

List of Figures

Chapter 1

Introduction

In sociology and social psychology, an in-group is a social group that a person psychologically identifies as being a member of. An out-group is a group with which an individual does not identify. Social Identity and Self Categorization Theories [49, 52] established the foundations for theoretical work on group identity several decades ago. These theories, despite differences, agree that the psychological categorization of in-groups and out-groups lead to phenomena such as in-group favoritism, out-group derogation [49], and group polarization [40].

Predicting social identities or in-groups of users in social media is useful in research across several domains like sociology, demography and public health, where human behaviour is studied. For example, studies that have used social media have focused on assessing population attitudes towards health related issues like tobacco use [42, 33] and vaccines [44, 18]. Other research has focused on studying issues like the onset of postpartum depression [17], suicide risk [29], sleep disorders [39] and distribution of fitness levels across geography [27] among others. Some of the recent studies have focused on analyzing cultural violence and peace [38], investigating illegal wildlife trade [24], identifying substance use risk [31] and detection of violent extremists. [30] Using social media platforms for research offers several advantages:

- Researchers can actively track real-time updates on users attitudes and behaviours as they

emerge.

- As posting on social-media is voluntary, users may report opinions with greater fidelity than they would otherwise do with interviews or surveys.
- Analyzing social media is low cost and can be automated, unlike surveys and interviews

These pros however, have a major con associated with them; Social Media data may not always represent the population that researchers are interested in analyzing. Identifying and quantifying bias in the experimental results is extremely challenging. This is due to a lack of identity indicators like age, gender, race among others. Therefore, creating a method to accurately and reliably detect the identity of users would expand the use of social media as a research tool for social and behavioural sciences and public health.

Most of the past research on predicting information about social media users has focused on building models that predict basic identities like gender [19, 32, 9, 14], age [58, 47, 12] and ethnicity [54, 10]. In this work, we focus on building models from self-reporting statements on Reddit that can predict a richer variety of identities in addition to these basic identities, like profession (artist, athlete,..), interests (cat person, dog person,..), hobbies (cyclist, gamer, runner,..), origin (american citizen, aussie, asian,..), religion (agnostic, atheist, catholic, hindu,..), sexual orientation (lesbian, gay, straight,..) and others (addict, alcoholic, ginger, gun owner, introvert,..).

Reddit is a popular social media platform with over 250 million users worldwide, consisting of a large number of discussion forums (called subreddits) focusing on different topics. Users do not state their name on their profiles, but use anonymous nicknames instead. There is evidence showing that having name anonymity increases self-disclosure in social media [37], making Reddit a promising platform to explore self-identification of in-groups and identities.

Our work assesses whether the social media platform Reddit can be effectively used for the

identification and prediction of identities and other sociological questions.

1.1 Goals of the project

The following are some areas that we would like to explore in this thesis:

- A list of different in-groups and out-groups that people self-identify and self-distance from in Reddit.
- An understanding of where Reddit users self-identify.
- An analysis of in-groups and out-groups that are highly correlated with each other. For example, if a person says "I am a man", do they also say "I am not a woman"?
- Evaluating whether self-identification phrases on Reddit can be used in constructing unbiased identity classifiers.

1.2 Document Structure

The rest of the Thesis is structured as follows:

- Chapter 2 provides a Literature Review that goes over research areas where social media has been used as a platform for study. Previous work in building identity models is also briefly outlined.
- Chapter 3 discusses the data that we use. It describes the creation of identity data-sets using our novel method of dependency parsing in addition to the basic regular expression matching. We look at where people self-identify and analyze the co-occurrences between in-groups. We also create a taxonomy of the rich variety of self-identities on Reddit. The work in this chapter was done in collaboration with Ignacio Arranz, Hangfeng He and Dianna Marsala.

- Chapter 4 outlines the method employed in building the test-set used for evaluating our models.
- Chapter 5 compares different methods of sampling negative classes for our training-set to train classifiers. A method to obtain alternate self-identification phrases is also briefly outlined.
- Chapter 6 looks at different types of feature representations.
- Chapter 7 discusses a summary of the results and potential future work.

Chapter 2

Literature Review

Much of the NLP work on group and demographic identities within social media has focused on demographic group prediction, where Twitter has been the social media platform of choice. As of 2017, over 60% of the studies were performed with Twitter data [15]. Blogs were a distant second, and to our knowledge very little research on demographic groups was performed on Reddit [1], even though it has been used as a source of data in recent studies on mental health [22, 21, 56].

Studies that involve predicting demographic attributes of authors have used a wide variety of linguistic and non-linguistic features in creating their models such as profile colors, username and user tweets [32]; profile images and user descriptions [34]; user location and username [13]; following relationship [10]; username, profile photo, friends/followers, date of creation and user tweets [41]. However, in this work, we focus on using only linguistic features, specifically comments made by Reddit users to build models that predict their identities.

In "Depression and Self-Harm Risk Assessment in Online Forums" [56], the paper that won the EMNLP best long paper award of 2017, the authors used the online support communities of Reddit to study and provide a framework for identifying posts that may indicate a risk of

self-harm. As self-harm is closely related to depression, they first identified depressed users through self-reported statements (e.g "I was just diagnosed with depression" and its variants) and matched those with control users. They use all posts and comments made by these authors to create a data-set (Reddit Self-Reported Depression Diagnoses) after a manual annotation step to control for quality. A convolutional neural network model was then trained on this data-set to identify whether an author was depressed based on the language used in the text. They show that the model beats all previous baselines by a large margin, and learns to weight phrases such as "i'm so sorry", "sometimes i", "to scare you", etc. heavily while classifying a person as depressed and at risk of self-harm.

In similar research [22], authors propose social media (specifically Reddit) as a way to characterize and predict shifts from discussion of mental health content to suicidal ideation by focusing on specific subreddits. Participants in Reddits mental health communities who go on to post on the platform's suicide support forum were characterized using a number of linguistic and social interaction based measures. Their results show that transition to suicidal ideation is associated with psychological states like heightened self-attentional focus, poor lingustic coherence and linguistic coordination with the community among others. They develop a Logistic Regression classifier that predicts the tendency of individuals discussing mental health concerns to engage in these characteristic "suicidal" behaviours with a high accuracy.

A study in 2013 [45] analyzed over 700 million words, phrases and topic instances collected from the Facebook messages of 75,000 volunteers. The size of their data enabled them to perform open-vocabulary analysis using phrases and automatically derived topics, thereby not limiting them to the use of *a priori* language categories. Their open vocabulary approach (Using phrases and topics generated by Latent Dirichlet Allocation) results show strong correlations between language and personality, gender or age. They showed that mentions of an assortment of social sports and life activities (such as basketball, snowboarding, church, meetings) correlate with emotional stability, and that introverts show an interest in Japanese media (such as

anime, pokemon, manga and Japanese emoticons: ^_^). Their inclusion of phrases in addition to words provided further insights such as males prefer to precede "girlfriend" or "wife" with the possessive "my" significantly more than females do for "boyfriend" or "husband".

Past work on predicting user demographics have employed different methods of labeling ground truth data. Sometimes user names are matched against databases of popular names to infer user gender [15]. To infer age, searching for patterns such as "Happy ##th/st/nd/rd birthday to me" has been employed [58]. These methods can be followed by manual inspection [36] to ensure the quality of the labels. Other approaches include the annotation of labels through crowd-sourced workers, such as Amazon Mechanical Turk [35], or the use of commercial vendors [57]. Other methods included simple sentences [55] or regular expressions [41].

Similarly to how Twitter handles are used to label gender, recent work has used Reddit-specific user features to identify user characteristics. "Flair" is a set of tags that users self-apply within a specific subreddit that cause an icon or badge to appear next to their usernames. For example, "Eagles" or "Seahawks" are two types of flair found within the "NFL" subreddit. This feature has been used to identify users who disclose their personality type [26] or whether they have bipolar disorders [46]. Other work has used standardized behavior by users in the subreddit ChangeMyView[50], who use the *delta* character (Δ) to express they have changed their view, followed by a description of why it changed.

A similar study [16] to our work analyzed authors use predicates as a means of identifying fine-grained social roles among users on Twitter. They focus on roles that are finer grained than gender and political affiliation like "smoker", "student" and "artist". In order to identify a set of verbs that preferentially select that particular role, they rank verbs according to the point-wise mutual information of that verb appearing with the given role in the web-scale part of speech n-gram corpus Google-V2(Lin et al., 2010). Through this method, they find that the verb "draw" is highly indicative of an "artist", "play" is highly indicative of an athlete, "blog"

is highly indicative of a "blogger" and "cheer" is highly indicative of a "cheer-leader" among other examples. As a quality control step, they require that each tweet with the containing verb be annotated by Mechanical Turk workers as to how likely the author of the tweet belongs to the given social role. They use these positively annotated tweets along with a background set of tweets to train a classification model capable of identifying roles based on a tweet, with some accuracies of around 80%.

The most closely related work to ours is [11], which evaluates whether Twitter contains information to support the prediction of fine-grained categories/social roles like belieber, soldier, pessimist, singer, freshman, etc. They follow two steps in obtaining authors for the selected categories. The first step involves using simple self-identification statements like "I am a ___" and its variants to identify different fine-grained categories and authors associated with them. In the second step, they exploit a complementary signal based on characteristic conceptual attributes of a social role (i.e, match patterns of the form *identification*'s ___ where *identification* $\in$ [artist, doctor, lawyer, swimmer,...]). They identify typical attributes that occur frequently in a possessive construction with that role in the n-gram corpus Google-V2(Lin et al., 2010). For example, with the role of "doctor", they extract terms that match the simple pattern "doctor's ___". They append these results to the end of the phrases "my " and "I have a " and use these concatenated phrases as a second approach of identifying fine-grained categories. For example, for identifying authors for the category "barber", they concatenate the phrase "my" with a frequent occurring possessive construct "scissors", thereby using the concatenated phrase "my scissors" to search for authors belonging to the "barber" category. They additionally include a quality control step where they require Mechanical Turk workers to annotate the tweets containing these phrases based on how likely the author belongs to the given category.

Our work adds depth to their analysis in several ways. We change the self-identification patterns to use dependency parses. We analyze *where* users tend to self-identify by looking at subreddits (something not present in Twitter).

The authors in [56] raise an important issue related to the ethical concerns of using social media data for research, especially since they are often sensitive. Privacy concerns and risk to the individuals in the data should always be considered [25, 48]. The risks associated with the Reddit data-set used in this thesis are minimal as we only use information that is publicly available. This assessment is supported by previous work on Reddit data [20].

Chapter 3

Data Retrieval, Creation and Analysis

Reddit, often called "The front page of the internet" is the 6th most popular website in the United States according to Alexa [8] and the 20th worldwide. Reddit is divided into several "pages" called subreddits, each covering a different topic. Subreddits are managed by "moderators", volunteers who can edit the appearance of a particular subreddit, dictate what types of content are allowed, and even remove posts or content or ban users from that subreddit.

According to a survey conducted by the Pew Research Center's Internet & American Life Project [1], 6% of online adults are Reddit users. This survey also finds that young men are especially likely to visit the site, with 15% of male internet users between the ages of 18 and 29 saying that they use Reddit, compared with 5% of women in the same age range and 8% of men in the age group 30-49. Some of the other findings from the Pew Research Center survey are shown in Figure 3.1.

For Sections 3.1-3.3 and 3.8, the entire Reddit data (2005-2018) was used in the analysis, while for Sections 3.4-3.7, only Reddit data from the year 2017 was used.

reddit usage by demographic group

% of internet users in each group who use reddit

		% who use reddit
All internet users (n=1,895)		6%
a	Men (n=874)	8[b]
b	Women (n=1,021)	4
Race/ethnicity		
a	White, Non-Hispanic (n=1,331)	5
b	Black, Non-Hispanic (n=207)	4
c	Hispanic (n=196)	11[ab]
Age		
a	18-29 (n=395)	11[bcd]
b	30-49 (n=542)	7[cd]
c	50-64 (n=553)	2
d	65+ (n=356)	2
Education attainment		
a	No high school diploma (n=99)	9
b	High school grad (n=473)	4
c	Some College (n=517)	6
d	College + (n=790)	7[b]
Household income		
a	Less than $30,000/yr (n=417)	6
b	$30,000-$49,999 (n=320)	6
c	$50,000-$74,999 (n=279)	7
d	$75,000+ (n=559)	6
Urbanity		
a	Urban (n=649)	7[c]
b	Suburban (n=893)	6[c]
c	Rural (n=351)	2

Source: Pew Research Center's Internet & American Life Project Spring Tracking Survey, April 17 – May 19, 2013. N=2,252 adults ages 18+. Interviews were conducted in English and Spanish and on landline and cell phones. The margin of error for results based on all internet users is +/- 2.5 percentage points.
Note: Percentages marked with a superscript letter (e.g., [b]) indicate a statistically significant difference between that row and the row designated by that superscript letter, among categories of each demographic characteristic (e.g. age).

Figure 3.1: PEW research survey results on Reddit

3.1 Data Description

All public posts and comments made by Reddit users since 2005 are stored in a public database on BigQuery, which is Google's scalable cloud data warehouse. As described on its homepage [3], BigQuery makes it fast and efficient to query your data using SQL, thereby allowing you to easily work with extremely large data-sets. The Reddit data-set was created by

Jason Baumgartner [5] of PushShift.io [6] aided by The Internet Archive [4].

A subset of the 20 fields present in the Reddit data-set is shown in Table 3.1 along with a brief description. These were the main fields used to perform our analysis presented in the later section. Reddit has been growing in popularity since it was founded in Medford, Massachusetts in June, 2005, resulting in a higher volume of comments posted every year. This becomes apparent from Figure 3.2, where the number of comments per year has grown from 48,489,057 in 2010 to 1,239,564,030 comments in 2018.

The data-set is split into multiple files, one for every month of every year. Some statistics of the data-set are displayed in Table 3.2.

Field	**Description**
created_utc	Time when the comment was created (Coordinated Universal Time)
subreddit	The subreddit in which the comment was posted
body	The comment
author	Reddit username of the comment author
author_flair_text	A tag chosen by the author for a particular subreddit/comment
parent_id	the id of the comment/post to which this comment was posted
id	unique identifier for the comment

Table 3.1: Examples of fields in the Reddit data-set on BigQuery

Metric	**Value**
Size (Terabytes)	1.4
Number of Comments	5,075,098,104
Number of unique authors	26,904,699
Number of Years	14 (2005-2018)

Table 3.2: Statistics of the Reddit Data-set

3.2 Main Challenges

Google BigQuery allows users to query up to 1 Terabyte of the data for free every month. For every additional terabyte queried, a $5 fee is charged. This makes it expensive to work

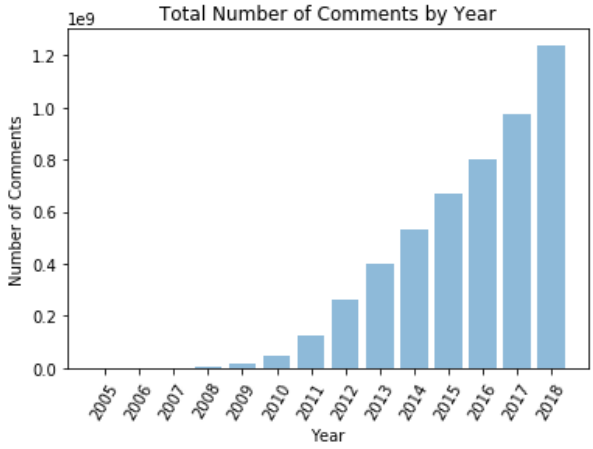

Figure 3.2: Number of Comments per year

and perform several different analysis using the entire 1.4 Terabyte Reddit data-set. Therefore, the slowest tasks were prioritized and run on BigQuery, while relatively faster tasks were run locally on Nlpgrid[1].

Running tasks on Nlpgrid required that the data be stored locally on the server in a consistent and easily accessible format. Once downloaded on Nlpgrid, all the files were converted from formats such as *bzip2, xz* and *zstd* to *gzip* formatted files and stored across 200 files.

The Reddit data-set contains surprisingly many comments made by bots. These bots could negatively impact the quality of our analysis as they echo comments of other users, thereby adding noise to the results. It was important that these comments were removed before we proceeded with any further analysis. Luckily, most bots self-identify, so any authors having comments containing phrases like "I am a bot", "I am a robot" and its variants were removed. 32,759 robots were identified from the Reddit Data-set with over 211,000,000 comments (which accounts for 4.16% of the comments on Reddit)! The top 5 robots with the highest number of comments are displayed in Table 3.3

[1]Nlpgrid is a Massively Parallel(MP) Cluster at the University of Pennsylvania

Reddit User-name	Total Number of Comments
AutoModerator	33,551,744
MTGCardFetcher	1,102,559
MemeInvestor_bot	1,070,663
grrrrreat	882,512
imguralbumbot	825,362

Table 3.3: Robots with the Highest Comments in the Reddit Data

3.3 Identifying Authors' In-groups via Self-Identification

3.3.1 Regular Expression Matching and its Limitations

In [11], the authors rely on variants of a single pattern, "I am a ___" to bootstrap data for training balanced-class binary classifiers using unigrams observed in twitter tweet content. Inspired by this method, the in-groups(out-groups) of authors were initially obtained by doing a simple regular expression matching of their comments to the phrase "I am a(n) ___"("I am not a(n) ___"), where the blank matches the immediate next word. However, this method produced several false positives in the results. False positives are instances where the search pattern detects a sentence as a user self-identifying with a given role, when in reality it is not the case. A simple example is that such a pattern would identify "I am a dog lover" as a "dog", doing so incorrectly. Similarly, "My mom thinks I am a star" would tag the user as a "star".

3.3.2 Our Approach: Dependency-Parse based Matching

In order to reduce the false positives in the results, another step was added after the regular expression matching stage, where dependency parsing (with the open source spaCy package) was used to identify the cases where the verb "am" is the root of the sentence, and find its attribute in the expression. This solves for the issues stated above. An example of this is shown

in Figure 3.3

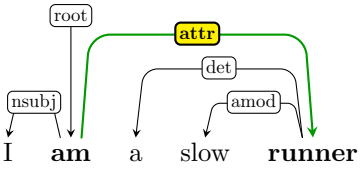

Figure 3.3: The use of attributes to identify in-groups in expressions such as "I am a(n) ___"

Having identified the attribute, a recursive search of the dependency tree at the left and right of the attribute is performed, which identifies in-groups with adjectival modifiers ("**slow** runner"), adverbial modifiers ("**very** empathetic person"), or compounds ("**free speech** absolutist"). This method ensures "**only child**" is not identified as a "child", or an "**odd duck**" is not identified as a "duck".

A similar process is performed with expressions such as "I am not a(n)..." and "I'm not a(n)...", which generates the data-set of out-groups. Most frequent matches for the dependency-based pattern are shown in Table 3.4.

The results were manually inspected to review the quality of the identified cases. 100 of the identified sentences were sampled to classify between positive, not clear, or negative matches. Table 3.5 shows the precision obtained for the listed categories, calculated as the positive cases over the total number of sampled sentences. The prevision for our matches are reported and compared to the prevision reported by [11] for Twitter data. The results show stronger precision, in particular for more fine-grained or specific roles.

#	In-groups	#	Out-groups
207,834	idiot	198,779	expert
204,829	guy	90,212	doctor
133,650	sucker	73,902	lawyer
123,921	girl	44,042	fan of it
108,316	man	33,882	smart man
107,906	atheist	24,033	idiot
102,946	woman	21,844	scientist
80,546	dude	20,598	parent
74,027	student	19,735	fan of them
72,833	american	19,005	mod
61,281	firm believer	18,747	professional
56,882	asshole	18,496	trump supporter
56,882	noob	18,406	native speaker
53,833	member	17,372	huge fan of it
44,243	male	17,137	woman
41,880	person	16,900	native english speaker
38,782	christian	16,696	member
38,727	female	16,380	fan of
36,090	college student	16,303	pro
35,332	engineer	16,041	christian
31,020	expert	15,979	troll
29,845	senior	15,924	big fan of it
29,161	believer	15,269	racist
28,080	teacher	13,721	teacher
27,275	programmer	13,507	asshole
27,040	introvert	13,352	american
26,497	newbie	12,882	programmer
26,458	part	12,854	bot
25,253	liberal	12,799	girl

Table 3.4: The most frequent matches to our dependency-based pattern

3.3.3 Limitations of our approach

Our approach with dependency parsing does not guarantee eliminating false positives completely. The attribute "bit" is still generated as an output. As observed by [11], the results for "lawyer", "engineer" or "doctor" have a high number of false positives due to memes such as the Star Trek reference "Dammit Jim, I'm a doctor, Not an X". Other false positives include sentences in quotes, or lyrics of songs.

Additionally, this method identifies authors who use phrases such as "I am a Jew", but does not identify a phrase such as "I am Jewish", which is more common. Similarly, our method

Role	On Reddit	On Twitter
civil engineer	0.99	–
electrical engineer	0.98	–
engineer	–	0.6
atheist	0.97	0.5
business owner	0.96	–
dad	0.95	–
economist	0.94	–
vegetarian	0.94	0.7
wife	0.93	–
man	0.86	0.8
doctor	0.64	0.2

Table 3.5: Precision of matches obtained for a sample set of in-groups, after manually reviewing 100 comments from each

misses phrases such as "I am muslim" and "I am gay" among others.

3.4 Self-Identification through Flair

The alternative approach for obtaining user identities was the use of flair. Flair is a user defined attribute that causes an icon or badge to be appended to their username on a given subreddit. A total of 2,508,667 users use flair to identify themselves, doing so in 32,464 subreddits out of 247,322 subreddits. In comparison, 1,269,102 users use self-identification phrases in 33,150 subreddits.

This method of self identification represents personal interests of the users. Five out of the top eight subreddits are sports related (NBA, NFL, soccer, hockey, baseball), two are from video-games (Overwatch, Global Offensive) and one of them is about politics (The_Donald, which discusses US president Donald Trump). When looking for flair that may represent user demographics, there are several that can be useful for this purpose, as shown on Table 3.6.

Subreddit	Top flairs
AskMen	Male, Female, Bane, Sup Bud?
Ask Women	Female, Male
Teenagers	14, 15, 16, 17, 18, 19
Christianity	Christian (Cross), Roman Catholict, Atheist, Episcopalian, Eastern Orthodox, United Methodist
Military	United States Army, Army Veteran, US Air Force, US Navy, US Marine Corps, Marine Vet
AskMen Over30	male 30-34, male 35-39, male over 30, male 40-44, male 45-49

Table 3.6: Subreddits with flair categories that identify demographic or fine-grained groups

3.5 Building a taxonomy of users' in-groups and out-groups

To perform a detailed analysis that can be relevant to studies in sociology, a data-set of users that use self identification with a subset of selected categories is built.

When filtering by self-identification categories with at least 300 users in 2017, the result is a remaining 792 categories. These are manually inspected to keep only those which represent relevant identifications of groups. This process excluded **false positives** ("I am a bit", "I am a little", "I am a level") and categories with insufficient information due to the method of extracting information ("I am a fan", "I am a firm believer"). This leaves **272 concrete categories**, which are categorized as related to Age, Gender, Family Role, Hobbies, Jobs, Origin, Politics, Religion, Sexual Orientation, Sports, Wealth and Other. These are given in Table 3.7. The remainder of the analysis in this chapter looks at subsets of these 272 categories to draw insights on the in-groups and out-groups of Reddit users.

3.6 Where do users self-identify?

In looking to understand the forums where users disclose a given type of identity, we look at the subreddits with the highest rate of self-identification statements. Looking at overall volumes is a proxy for subreddits with most activity, so we look at the fraction of users or comments self-identifying with a category.

Taxonomy of ingroups and outgroups by category
AGE adult, 30 year old man, millennial old guy, teenager, old man, teenager, young woman
ANIMALS animal lover, cat person, dog lover, dog person
FAMILY ROLE dad, married man, mom, parent, single dad, single guy, single mom, single parent, wife
GENDER chick, dude, female, girl, grown woman, guy, lady, male, man, trans man, trans woman, woman
HOBBIES cyclist, gamer, PC player, PS4 player, runner
ORIGIN American citizen, Aussie, Australian, Brit, Californian, Canadian, Chinese, dual citizen, European, foreigner, immigrant, Indian, New Yorker, southerner, Texan, US citizen, white American
POLITICS anarchist, Bernie supporter, capitalist, communist, conservative, Democrat, fascist, feminist, leftist, liberal, libertarian, Marxist, Nazi, registered Democrat, Republican, socialist, Trump supporter
PROFESSION accountant, actor, architect, artist, athlete, attorney, baker, barista, bartender, biologist, business owner, carpenter, cashier, chef, chemist, civil engineer, college kid, college student, computer programmer, computer science major, computer scientist, consultant, contractor, cook, cop, CS major, CS student, delivery driver, dentist, designer, developer, doctor, drummer, economist, editor, electrical engineer, electrician, EMT, engineer, engineering student, English major, English teacher, entrepreneur, farmer, filmmaker, firefighter, freelancer, FTM, full time student, game developer, grad student, graduate, graduate student, graphic designer, guitarist, high school student, high school teacher, historian, intern, IT guy, journalist, landlord, law student, lawyer, librarian, lifeguard, manager, mathematician, mechanic, mechanical engineer, med student, medic, medical student, musician, network engineer, nurse, nursing student, paramedic, pharmacist, PhD student, photographer, physician, physicist, plumber, police officer, professor, programmer, project manager, psychologist, rapper, recent grad, recent graduate, recruiter, reporter, resident, RN, SAHM, scientist, server, singer, small business owner, social worker, software dev, software developer, software engineer, soldier, sophomore, student, supervisor, sysadmin, teacher, technician, therapist, trader, truck driver, undergrad, university student, waiter, waitress, web dev, web developer, writer
RACE Asian, black guy, black woman, white dude, white girl, white guy, white male, white man, white person, white woman
RELIGION agnostic, agnostic atheist, Catholic, Christian, Jew, Mormon, Muslim
SEXUAL ORIENTATION lesbian, gay dude, gay guy, gay male, gay man, bisexual woman, straight dude, straight female, straight guy, straight man, straight white male
SPORTS baseball, Cubs fan, basketball, Cavs fan, Celtics fan, Knicks fan, Lakers fan, Spurs fan, NFL Bears fan, Browns fan, Cowboys, fan, Eagles fan, Falcons fan, Giants fan, Jets fan, Liverpool fan, Packers fan, Patriots fan, Pats fan, Steelers fan, Vikings fan, Arsenal fan, Chelsea fan, United fan
WEALTH billionaire, broke college student, broke student, millionaire
OTHER addict, alcoholic, ginger, gun owner, introvert, junior, lefty, nerd, only child, pedophile, recovering addict, righty, smoker, transplant, twin, veteran, vegan, vegetarian, virgin, vet

Table 3.7: Ingroups and outgroups with 300+ users

For each of the 14 categories listed in Table 3.7, we filter out subreddits with fewer than 10,000 comments and look at the top subreddits sorted by:

- The fraction of authors who self-identify as that category.
- The fraction of comments that contain a self-identification statement belonging to that category.

Category	Top Subreddits
Age	Drama, CPTSD, raisedbynarcissists, aspergirls, AsianParentStories, autism, stepparents...
Animals	**Dogfree**, **Pets**, penpals, MonsterGirl, playingcards, **hitmanimals**, **Goldfish**, HungryArtists...
Family Role	NoFapChristians, **MomForAMinute**, **AskParents**, **parentsofmultiples**, **Parenting**...
Gender	**ask_transgender**, **genderqueer**, **Bumble**, **MensLib**, GCdebatesQT, **TransyTalk**, **bisexual**...
Hobbies	usedpanties, **RS3Ironmen**, **scienceofdeduction**, vidme, **ironscape**, copypasta, **GirlGamers**...
Origin	usedpanties, **immigration**, **IWantOut**, **ChineseLanguage**, **cuba**, blog, penpals, **taiwan**...
Politics	**DebateAnarchism**, **Socialism_101**, **DebateCommunism**, **Anarchy101**, **Liberal**...
Profession	penpals, WindowsMR, Instagram, gameDevClassifieds, TrueAtheism, **freelance**, **engineering**...
Race	**Alt_Hapa**, **AsianSubDebates**, **AsianMasculinity**, **hapas**, **socialjustice101**, **blackladies**...
Religion	**religion**, **Christian**, **DebateReligion**, **islam**, **TrueAtheism**, **DebateAChristian**...
Sexual Orientation	**LesbianActually**, **actuallesbians**, **GenderCritical**, **lgbt**, **bisexual**, **LadyBoners**, **amihot**...
Sports	**EvilLeagueOfEvil**, **AroundTheNFL**, **falcons**, **Fantasy_Football**, **soccercirclejerk**...
Wealth	copypasta, videogamedunkey, Instagram, **millionairemakers**, **DaveRamsey**, **fyrefestival**...
Other	**alcoholism**, **introvert**, **alcoholicsanonymous**, **vegetarian**, **REDDITORSINRECOVERY**...

Table 3.8: Top subreddits by fraction of self-identification comments. Bold subreddits are related to that category.

This indicates that self-identification is closely related to the topic being discussed. This shows a strong relation to Reddit's anonymity and absence of usernames. It creates a setting where users are prompted to provide more information in order to add meaning.

These results follow the cooperative principle of conversation [28]. The cooperative principle is divided into 4 maxims, one of them being the maxim of quantity, which states that one tries to be as informative as one possibly can, and gives as much information as is needed, and no more. More contemporary theories, like social information processing theory [53] and media richness theory [23] attempt to explain how people manage relationships in computer-mediated environments (like instant messaging, e-mail, and chat rooms) that do not reproduce visual social cues that are used in face-to-face conversation. Applying these theories to Reddit can help explain why people state their gender identity where it is relevant as context to their opinion, which perhaps would not be necessary if their gender was stated or implicit in their profile.

In the case of political identities, "I am a Republican" is most frequently found in the subreddits "**Republican**", "Alabama", "**Impeach Trump**", "Prematurecelebration", "**ModelUSGov**".

Subreddit	% of comments
AskFeminists	0.17%
OneY	0.17%
debateAMR	0.15%
Feminism	0.14%
GenderCritical	0.14%
ftm	0.14%
koreanvariety	0.13%
feminisms	0.13%
TwoXChromosomes	0.09%
PurplePillDebate	0.08%
RedPillWomen	0.07%
MensRights	0.07%
FeMRADebates	0.07%
DigitalCartel	0.07%
againstmensrights	0.07%
TiADiscussion	0.07%
FemmeThoughts	0.07%
TheBluePill	0.07%
asktransgender	0.06%
genderqueer	0.06%
everyansshouldknow	0.06%

Table 3.9: Top subreddits by rate of male identification ("I am a man"). Bolded subreddits are gender related.

For democrats, "**moderatepolitics**", "maryland", "**hillaryclinton**" and "oklahoma" are top subreddits. The 5 bolded subreddits out of the top 10 listed are very clearly forums for political discussion.These examples help draw insights on how users tend to self-identify in a way that is most relevant to the discussion.

3.7 Co-occurrence across groups

3.7.1 Measuring Co-occurrence

We want to understand which groups have a high correlation, or are identified simultaneously by many users. This can help us understand whether a certain gender has a stronger or weaker affiliation towards a political view, or what jobs they have. Figure 3.4 shows the number of unique self-identifications made by 'X' Reddit user. Over 1.2 million users self-report

at least 1 unique identity of the form "I am a(n) ___". We use point-wise mutual information [43] to measure co-occurrence of different groups individuals associated themselves with. Mathematically:

$$\text{pmi}(x_1, x_2) = \log \frac{p(x_1, x_2)}{p(x_1)p(x_2)}$$

In this case x_1 and x_2 are two different in-groups, and we want to know if the probability of them occurring jointly is higher than what we would expect if they were independent and unrelated.

Having user i and group j, we estimate the probabilities:

$$p(j) = \frac{\sum_i \mathbb{1}_{user_{i,j}}}{\sum_i \mathbb{1}_{user_i}}$$

$$p(j_1, j_2) = \frac{\sum_i \mathbb{1}_{user_{i,j_1,j_2}}}{\sum_i \sum_j \sum_j \mathbb{1}_{user_{i,j,j}}}$$

We selected instances with more than 20 co-occurrences, to ensure we did not include use cases that were unusual due to being a combination of two groups with very low probabilities.

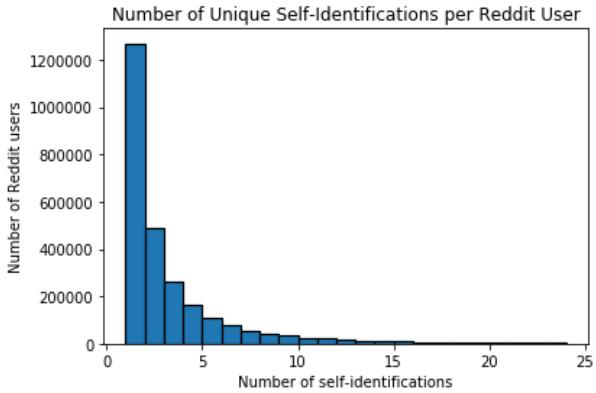

Figure 3.4: Number of Reddit Users that self-identify 'X' number of times

3.7.2 Co-occurrence between in-groups

We selected major demographic categories, and for its main labels looked at co-occurrence of self-identification across patterns. The major set of co-occurrences took place between fans of teams from the same city, so we compare NFL and NBA teams[2]. Table 3.10 shows how Browns fans and Cavs fans (for the Cleveland Browns NFL team and Cleveland Cavaliers NBA team) have the highest co-occurrence ratio.

PMI	Cavs Fan	Celtics Fan	Knicks Fan	Lakers Fan	Spurs Fan
Bears fan					
Browns fan	5.2				
Cowboys fan					3.1
Eagles fan					
Falcons fan	1.7	1.1		1.5	0.4
Giants fan		1.3	2.9		
Jets fan	0.8	0.2	3.8	0.6	
Packers fan	1.4			0.2	
Patriots fan	0.7	3.8		1.5	1.4
Steelers fan	2	1.8			0.1
Vikings fan	0.3		0.3	1.7	

Table 3.10: Co-occurrence between NFL teams and NBA teams. Empty values in table had negative PMI

We perform error analysis for users who identify as part of two groups known to be mutually exclusive. Table 3.11 illustrates examples picked randomly from users that simultaneously self identify as “republican” and “democrat”.

For the first three cases users seem to self identify as both. The last example contains two different comments stated by the user in different conversations. The second sentence is not entirely clear, and would benefit from more context on the entire conversation, which we have kept outside of the scope of this work. For a list of other in-group co-occurrences, refer to

[2]Spurs Fan could potentially be referring to Mauricio Pochettino’s Tottenham Hotspurs instead of Gregg Popovich’s San Antonio Spurs

"**i'm a republican** by heart but i vote with logic and always vote to give money back to the poor... so **i'm democrat.**"
"I always considered myself a conservative and **I'm a registered republican**. But I believe we should have a single payer healthcare system [....]
Im a lifelong northeast democrat, two time obama voter who woke up one day like, wtf, **I'm a republican now**...PA, MI, WI etc all feel the same way.

Table 3.11: Examples of users who were identified as both republicans and democrats

Table 3.12.

3.7.3 Co-occurrence of in-groups with out-groups

We also analyzed the categories of self-identification that had highest co-occurrence with categories of self-distancing. The main results shown on Table 3.13 help draw interesting insights. A huge part of the cases are all political labels. The majority of these cases are of identities that could be regarded as being part of a same larger category ("communist, Marxist and socialist", "Christian and catholic"). Users seem to be looking to clarify their specific ideological affiliation within a larger group. **This suggests that users tend to self-distance from in-groups that are similar but not the same to the ones they are part of.**

3.8 Creating Author Representation Data-set

In order to make the analysis in the subsequent chapters easier, we create an Author Representation data-set in which every author is represented by a random subset of comments made by them on Reddit. The challenge for this task was to ensure that the random subset wasn't time biased (i.e, most of the comments from the subset belong to a particular year) especially given the size of the corpus.

A constraint was added to limit the total number of words in the random subset of comments

I am	I am	PPMI	I am	I am	PPMI
bisexual	queer	8.70	bisexual	libertarian	3.14
bisexual	gay	7.96	agnostic	gay	3.03
bisexual	trans	7.79	bisexual	democrat	2.88
gay	trans	6.69	atheist	bisexual	2.75
gay	queer	6.11	gay	muslim	2.74
queer	trans	6.02	bisexual	jew	2.72
bisexual	woman	4.45	gay	mormon	2.68
agnostic	bisexual	4.11	bisexual	lesbian	2.59
bisexual	mormon	3.91	lesbian	trans	2.56
bisexual	man	3.83	democrat	gay	2.53
gay	man	3.80	catholic	gay	2.47
gay	jew	3.61	bisexual	republican	2.44
gay	lesbian	3.17	bisexual	muslim	2.33
bisexual	libertarian	3.14	bisexual	catholic	2.11
agnostic	gay	3.03	gay	libertarian	2.05
bisexual	democrat	2.88	gay	republican	2.05
atheist	bisexual	2.75	trans	woman	1.99
gay	muslim	2.74	atheist	gay	1.64
bisexual	jew	2.72	lesbian	queer	1.56
gay	mormon	2.68	gay	woman	1.48
bisexual	lesbian	2.59	queer	woman	1.43
lesbian	trans	2.56	man	trans	1.26
democrat	gay	2.53	catholic	queer	1.15
catholic	gay	2.47	agnostic	trans	0.91
bisexual	republican	2.44	jew	trans	0.85
bisexual	muslim	2.33	agnostic	queer	0.82
bisexual	catholic	2.11	agnostic	mormon	0.81
gay	libertarian	2.05	muslim	queer	0.78
gay	republican	2.05	man	queer	0.65
trans	woman	1.99	agnostic	atheist	0.41
atheist	gay	1.64	mormon	muslim	0.35
lesbian	queer	1.56	mormon	trans	0.30
gay	woman	1.48	catholic	republican	0.24
queer	woman	1.43	democrat	queer	0.21
man	trans	1.26	libertarian	trans	0.20
catholic	queer	1.15	jew	queer	0.17
agnostic	trans	0.91	muslim	trans	0.14
jew	trans	0.85	catholic	trans	0.09
agnostic	queer	0.82	libertarian	queer	0.05
agnostic	mormon	0.81	agnostic	catholic	0.01

Table 3.12: Point-wise mutual information for 80 co-occurring self identification pairs with the highest PPMI values

for each author to a maximum of 20,000 words. This maximum word limit was chosen in order to make sure that the subsequent analysis run-times were not very long, while also making sure

I am	I am not	PPMI	I am	I am not	PPMI
trans man	trans woman	7.93	lesbian	trans woman	4.18
trans woman	trans man	7.49	capitalist	libertarian	4.17
agnostic atheist	agnostic	5.81	leftist	liberal	4.16
anarchist	capitalist	5.74	communist	liberal	4.16
communist	capitalist	5.59	fascist	anarchist	4.15
socialist	capitalist	5.48	paramedic	firefighter	4.13
leftist	centrist	5.07	white woman	mom	4.12
capitalist	socialist	5.05	anarchist	leftist	4.12
leftist	capitalist	5.05	socialist	fascist	4.11
drummer	guitarist	4.97	mom	single parent	4.11
lesbian	trans man	4.93	leftist	fascist	4.11
communist	anarchist	4.88	centrist	conservative	4.06
gay man	trans woman	4.84	wife	mom	4.05
mom	single mom	4.82	capitalist	conservative	4.04
cat person	dog person	4.79	capitalist	communist	4.04
capitalist	anarchist	4.78	socialist	liberal	4.04
undergrad	grad student	4.70	centrist	lefty	4.00
socialist	anarchist	4.62	bisexual woman	lesbian	3.98
capitalist	leftist	4.55	nursing student	nurse	3.95
sysadmin	network engineer	4.50	centrist	socialist	3.94
grad student	undergrad	4.49	leftist	socialist	3.94
centrist	leftist	4.43	grown woman	mom	3.93
fascist	socialist	4.41	fascist	communist	3.93
capitalist	fascist	4.39	communist	socialist	3.89
trans woman	gay man	4.38	centrist	fascist	3.87
guitarist	drummer	4.38	agnostic	capitalist	3.87
communist	fascist	4.37	communist	leftist	3.83
socialist	communist	4.33	fascist	libertarian	3.81
mechanical engineer	civil engineer	4.32	anarchist	liberal	3.78
fascist	leftist	4.30	white woman	dude	3.78
libertarian	anarchist	4.29	landlord	plumber	3.77
leftist	anarchist	4.28	bisexual woman	dude	3.75
anarchist	fascist	4.27	gay dude	lady	3.74
anarchist	communist	4.26	liberal	centrist	3.73
electrical engineer	mechanical engineer	4.24	musician	guitarist	3.72
parent	single dad	4.23	fascist	conservative	3.72
leftist	communist	4.22	libertarian	capitalist	3.71
socialist	centrist	4.22	young woman	mom	3.71
anarchist	socialist	4.20	feminist	trans woman	3.70
dog person	cat person	4.18	leftist	libertarian	3.60

Table 3.13: Point-wise mutual information for 80 self identification and self distance pairs with the highest PPMI values

that that there was sufficient valuable information retained per author.

Chapter 4

Creating a Data-set of Reddit Users who do not Explicitly Self-Identify

One of our objectives in this project is to evaluate whether we can use Reddit users who self-identify as a particular identity in building binary classifiers that predict that specific identity. For instance, we evaluate whether we can use all comments of users who say "I am a man" and its variants as positive training examples for building a binary Logistic Regression classifier that predicts whether an author is a "man".

During a discussion with Prof. Emily Falk [2], an Associate Professor of Communication at the Annenberg School for Communication, she raised an important point stating that in order to build robust identity prediction models, an important question that we would first need to answer is whether there is bias in users who self-report their identity. A person who says "I am a man" may not be completely representative of the entire male population in terms of the language they use or the topics they discuss. Similarly, this applies to people who self-disclose

their non-identities ("I am not a ___").

Therefore, in addition to evaluating the model's performance on users who self-identify (A validation/test set created from users who self-identify with "I am a ___"), we would need to evaluate the model's performance on users who don't explicitly self-identify with the phrase "I am a ___" and its variants. In this chapter, we explore a method used to create this test-set of users who do not explicitly self-identify.

4.1 Attempting to Create a Test-Set from a Survey

A Qualtrics survey was created using identities hand-picked from the most frequent self-identification groups obtained from Reddit and posted as a HIT (Human Intelligence Task) on Mechanical Turk (MTurk). The questions were of the form, "Do you belong to any of the following categories?", followed by a list of in-groups, each with 3 options; "Yes", "No" and "Not Sure", and each worker was paid $1 for completing the survey. The identities included in the survey are shown in Table 4.1. The following worker restrictions were placed on the MTurk HIT via MTurk's built in qualification system:

- **Reddit Account Holder**: True
- **Number of HITs approved**: Greater than or equal to: 50
- **HIT Approval Rate**: Greater than or equal to: 90
- **Location: is one of**: Australia, Canada, United Kingdom, United States

The MTurk workers were required to fill in their Reddit User-name at the beginning of the survey as the goal was to collect gold-standard data by linking the survey answers to the respondent's Reddit comments. We used these users from the survey to validate predictions of our identity classification models. In order to control for quality, we created a post on Reddit to which the survey takers were required to post the unique id generated at the end of the

survey. It was made clear that the workers would be paid only if the Reddit user-name used to comment the unique-id on the Reddit post matched those given in the survey.

However, MTurk shutdown our survey as they found that it violated their terms of service since we were collecting Reddit user-names which they considered to be personally identifiable information. We subsequently revised the survey to exclude collecting Reddit user-names. The data we gathered in the second survey could not be used for creating the gold-standard data, but could be used to get an approximation of the true underlying distribution of each identity in Reddit.

222 workers completed the survey before the survey was shutdown. Out of these workers, the Reddit user-names of only 120 workers were presnt in our Reddit Data-set [7], and none of them used self-identification phrases of the form "I am a ___" on Reddit. The survey results for some of the in-groups collected in the first MTurk survey are displayed in Figure 4.1.

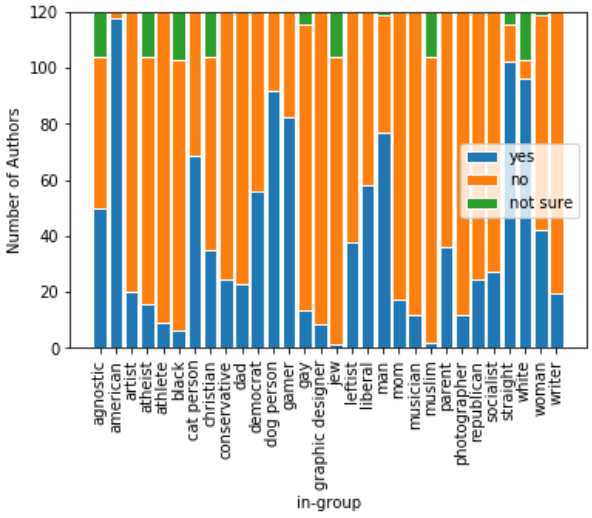

Figure 4.1: Test-set statistics from the MTurk Survey

A dude	A chick	A female	A girl
A grown woman	A guy	A lady	A male
A man	A trans man	A trans woman	A woman
An American citizen	Aussie	Australian	A Brit
A Californian	A Canadian	Chineese	A dual citized
European	A foreigner	An immigrant	An Indian
A New Yorker	A southerner	A Texan	An Asian
A black guy	A black woman	A white dude	A white girl
A white male	A white man	A white person	A white woman
A latino	A latina	A dad	A married man
A mom	A parent	A single dad	A single guy
A single mom	A single parent	A wife	A billionaire
A broke college student	A broke student	A millionaire	An anarchist
A Bernie supporter	A capitalist	A centrist	A communist
A conservative	A Democrat	A fascist	A feminist
A leftist	A liberal	A libertarian	A Marxist
A Nazi	A registered Democrat	A Republican	A socialist
A Trump supporter	A lesbian	A gay dude	A gay guy
A gay male	A gay man	A bisexual woman	A straight dude
A straight female	A straight man	A straight white male	An agnostic
An agnostic atheist	A Catholic	A Christian	A Jew
A Mormon	A Muslim	An adult	A 30 year old man
A millennial	An old guy	An old man	A teenager
A young woman	Accountant	Actor	Architect
Artist	Athlete	Attorney	Baker
Barista	Bartender	Biologist	Business Owner
Carpenter	Cashier	Chef	Chemist
Civil Engineer	College Kid	College Student	Computer Programmer
Computer Science Major	Computer Scientist	Consultant	Contractor
Cook	Cop	CS major	CS student
Delivery Driver	Dentist	Designer	Developer
Doctor	Drummer	Economist	Editor
Electrical Engineer	Electrician	EMT	Engineer
Engineering Student	English Major	English Teacher	Entrepreneur
Farmer	Filmmaker	Firefighter	Freelancer
FTM	Full Time student	Game developer	Grad student
Graduate Student	Graphic Designer	Guitarist	High school student
High school teacher	Historian	Intern	IT guy
Journalist	Landlord	Law student	Lawyer
Librarian	Lifeguard	Manager	Mathematician
Mechanic	Mechanical Engineer	Med student	Medic
Medical student	Musician	Network Engineer	Nurse
Nursing Student	Paramedic	Pharmacist	PhD student
Photographer	Physician	Physicist	Plumber

Police Officer	Programmer	Professor	Project Manager
Psychologist	Rapper	Recent Grad	Recent Graduate
Recruiter	Reporter	RN	SAHM
Scientist	Server	Singer	Small Business Owner
Social Worker	Software Dev	Software Developer	Software Engineer
Soldier	Sophomore	Student	Supervisor
Sys Admin	Teacher	Technician	Therapist
Trader	Truck Driver	Undergrad	University Student
Waiter	Waitress	Web Dev	Web Developer
Writer	Cavs Fan	Celtics Fan	Knicks Fan
Lakers Fan	Spurs Fan	Clubs Fan	Bears Fan
Browns Fan	Cowboys Fan	Eagles Fan	Falcons Fan
Giants Fan	Jets Fan	Packers Fan	Patriots Fan
Steelers Fan	Vikings Fan	Arsenal Fan	Chelsea Fan
United Fan	Liverpool Fan	An animal Lover	A cat person
A dog lover	A dog person	A cyclist	A gamer
An ironman	A PC gamer	A PC player	A PS4 Player
A runner			

Table 4.1: List of Identities included in the Qualtrics Survey posted on Mechanical Turk

Chapter 5

Selecting Negative Training Examples

In order to train binary classification models to predict in-group membership for users who don't explicitly self-identify, we need both positive training data (examples of users in that group) and negative training data (examples of users who do not belong to that group). While the positive training data came from people self-identifying using the phrase "I am a(n) ___", we had to consider a source for negative training data without introducing too much of a bias in the model.

In this chapter, we evaluate two approaches of obtaining negative training data for each of the in-groups. We look at the top features considered by the model in predicting a particular in-group from the text, and evaluate the accuracy of this model on the test-set.

For both the methods in this subsection, we process the training data in the following way:

- We consider equal class distribution
- Maximum number of training examples per class is set at 3000

- We under-sample from the majority class in order to maintain equal class distribution
- All forms of punctuation are removed
- We use unigram, bigram and trigram features

5.1 Method A: From Self-Distancing Authors

We consider authors who self-distance themselves from a certain in-group by using the phrase "I am not a(n) ___" as training examples for the negative class.

When we look at the negative examples obtained using this method, we realize that there is a lot of noise. For example, for the "Photographer" in-group, we notice that several authors use phrases such as "I am not a good photographer","I am not a professional photographer" and its variants. Therefore, considering self-distancing phrases with adjectives before the in-group introduces a lot of incorrect negative examples. This can be easily remedied by only looking at self-distancing phrases without adjectives.

While tuning the classifier, we penalize false positives and false negatives equally during the model prediction phase. We therefore plot the ROC curve for each classifier, and select the threshold for the Logistic Regression Model (default threshold is 0.5) which maximizes the sum of the true positive rate (TPR) and the true negative rate (TNR), where:

$$TPR = \frac{TP}{TP + FN} \tag{5.1}$$

$$TNR = 1 - \frac{FP}{FP + TN} \tag{5.2}$$

We evaluate these set of classifiers against the classifiers built in the next section to determine the best method for sampling training examples for the negative class.

5.2 Method B: From Random Sampling

We consider a random set of authors who haven't self-identified using the phrase "I am a(n) ___" towards the considered in-group as training examples for the negative class.

Take the example of building a model that classifies male authors. Suppose 50% of the training data accounts for the positive class where authors use the self-identification phrase "I am a man", the remaining 50% of the training data accounts for the negative class where authors don't explicitly self-identify as a "man". Assume that the true underlying distribution of Male Authors in Reddit is 65%. Therefore, under that assumption, 65% * 50% (=32.5%) of the authors in the negative training examples are actually males, while 17.5% (50% - 32.5%) are actually not males. Therefore, it would be incorrect to penalize False Positives and False Negatives equally during model prediction and tuning. Ideally, we would penalize False Negatives higher than False Positives as the negative data contains incorrectly labelled examples.

We would expect a perfect model to predict 65% of the negative class examples as Positive (FPR = 0.65), thereby coming up with the following equation:

$$FPR = \frac{FP}{FP + TN} = 0.65 \tag{5.3}$$

or to generalize,

$$FPR = \frac{FP}{FP + TN} = td_{ig} \tag{5.4}$$

where td_{ig} is the true underlying distribution of the in-group in Reddit.

We approximate the true underlying distribution of each in-group in Reddit as the proportion of authors that identified as that in-group in the MTurk survey (Test-set); refer to Figure 4.1. We tune the regularization coefficient (λ) of the Logistic Regression model using a

10-fold cross validation system and select the coefficient that satisfies the following equation:

$$\min_{\lambda}\left(\frac{FP_{\lambda}}{FP_{\lambda}+TN_{\lambda}}-td_{ig}\right) \tag{5.5}$$

Here, $FP_{\lambda}(TN_{\lambda})$ is the number of False Positives (True Negatives) obtained by the Logistic Regression Model with the regularization coefficient λ.

The test accuracies are compared for two Random Negative Sampling methods in Figure 5.1; one where the regularization coefficient $\lambda = \frac{1}{C} = 1$ is not optimized and one where it is according to Equation 5.5. On average, the test accuracies for the model where λ is tuned is higher, implying that this model is able to generalize better.

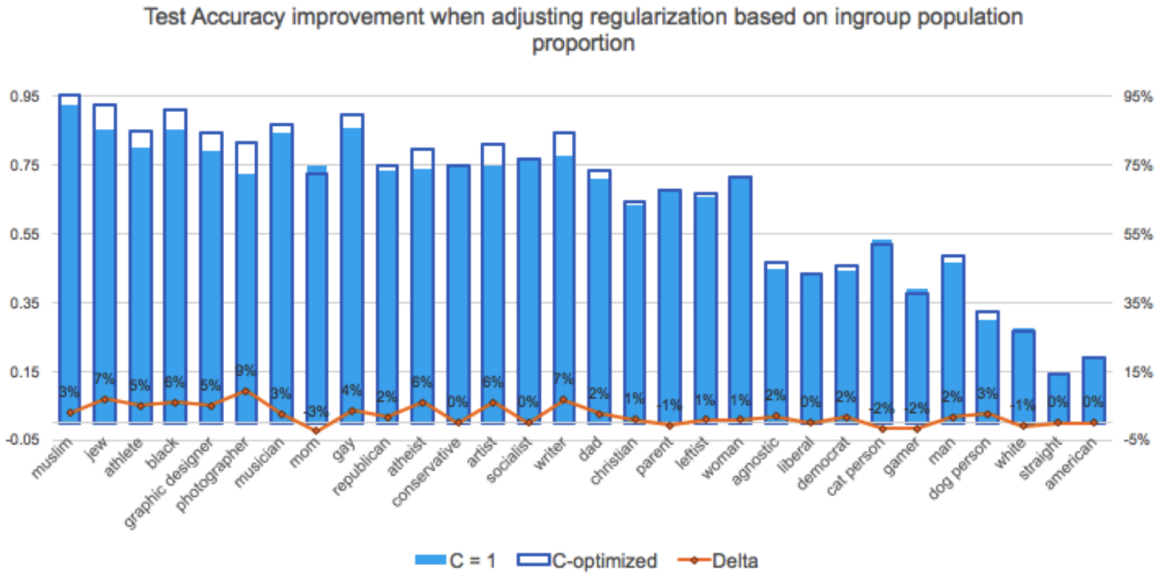

Figure 5.1: Test Accuracies for Random Negative Sampling Methods with and without tuning regularization coefficient

The cross validation and test accuracies are compared between the Random Negative Sampling (RNS) method with tuned regularization coefficient λ and the self-distance sampling method in Figures 5.2 and 5.3. The RNS method has higher cross validation accuracies on

average. However, the test accuracies for some in-groups are higher with the RNS method, while they are lower for other in-groups.

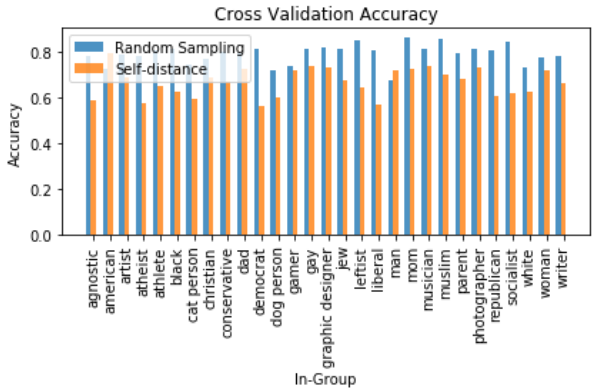

Figure 5.2: Cross Validation Accuracy for Different Negative Class Sampling Methods

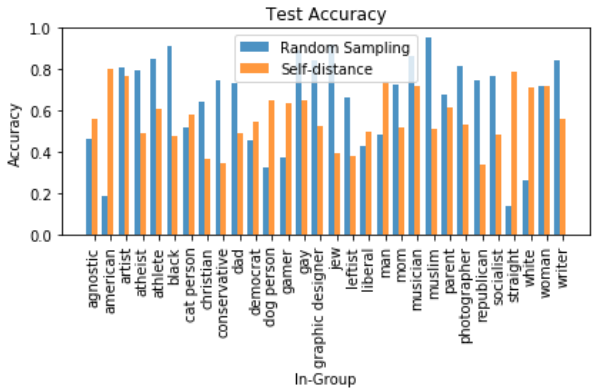

Figure 5.3: Test Accuracy for Different Negative Class Sampling Methods

In order to understand why the two models perform better for different identities, we plotted the test accuracies obtained using each negative sampling method for a particular identity against the identity's true underlying distribution on Reddit. This is shown in Figure 5.4

We notice a strong correlation between the test-accuracies for each negative sampling method and the proportion of Reddit users for each in-group. In-groups that were uncommon (low pro-

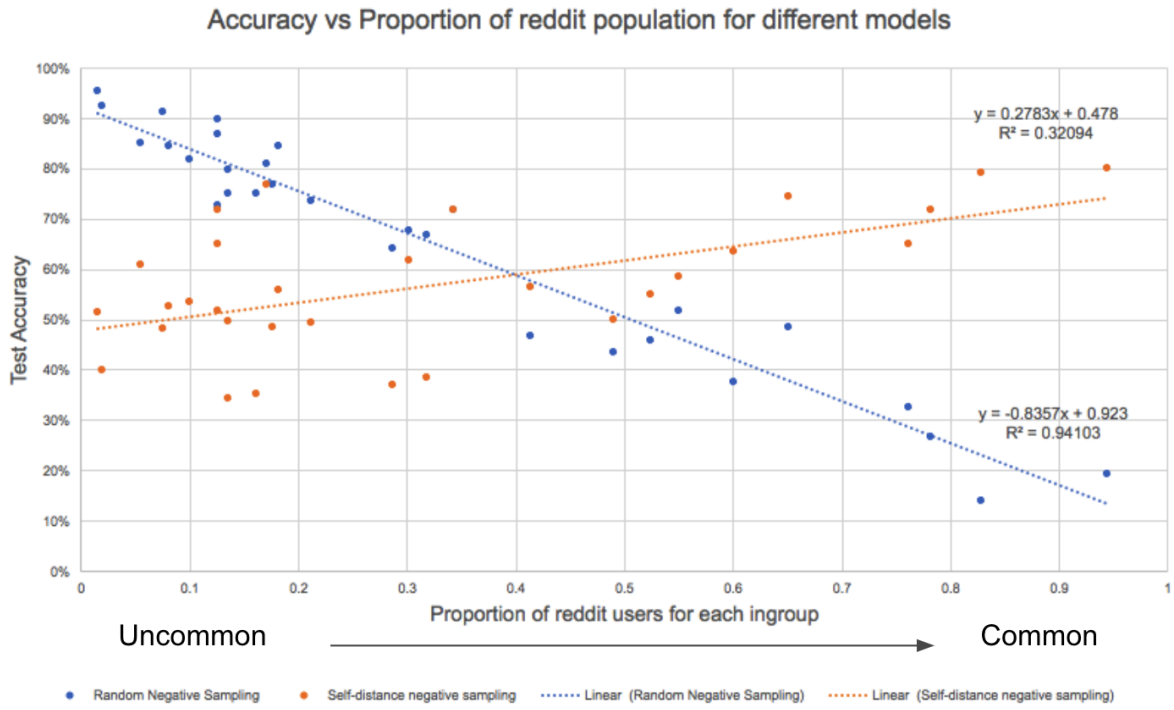

Figure 5.4: Test Accuracy for Different Negative Class Sampling Methods

portion of Reddit users) like "jew","muslim","black","athlete",etc. performed much better if the negative training examples were randomly sampled, while in-groups that were very common (high proportion of Reddit users) like "american","dog person","straight","white" performed much better if the negative training examples were sampled from authors who self-distance themselves from that in-group.

We therefore define a heuristic by which if the in-group is less than 40% of the entire Reddit population, Random Negative Sampling should be used. On the other hand, if the in-group is more than 40% of the entire Reddit population, negative training examples should be sampled from authors who self-distance themselves from that in-group.

The top features for the best model for each in-group based on this heuristic is shown in

Table 5.1.

Self-Identification	Top positive words
agnostic	using, post, ashamed, hour, affect, leg, camp, let, feel good, homes, statements, country,...
american	**american**, favorite, **im american**, college, realize, ive, looking, way, humor, past, im, pm,...
artist	**artist**, **art**, **artists**, **drawing**, **paint**, **draw**, **painting**, **design**, im, **digital**, **graphic**, **artwork**,...
atheist	**atheist**, **religion**, **religious**, **atheists**, **beliefs**, **atheism**, christian, bible, fucking, gay, church,...
athlete	**athlete**, **sport**, **sports**, **athletes**, **athletic**, college, **gym**, **weight**, **muscle**, **coach**, **lifting**,...
black	**black**, **black people**, **im black**, white, **racist**, **race**, white people, woman, **black guy**,...
cat person	**cat person**, **kitten**, things, **kitty**, new, **love cats**, better, fair, yes, story, okay, **kitties**, map,...
christian	**christian**, includes, perfect, real, **christians**, proudly, **bible**, **religion**, **church**, believe, support,...
conservative	**conservative**, liberal, christian, **conservatives**, libertarian, political, went, getting, like,...
dad	watermelon, **dad**, **wife**, old, **daughter**, happily, hs, **feel old**, rave, lincoln, love, drain,...
democrat	**democrat**, **registered democrat**, way, registered, like, year, big, putting, goddamn, im, realize,...
dog person	really, **puppy**, got, op, vid, little, game, sorry, **breed**, reason, **dogs**, pm, thing, better, buddy,...
gamer	**gamer**, **game**, **games**, **gaming**, **pc**, **steam**, going, thanks, im, disable, plus im, fix, problem,...
gay	**gay**, **im gay**, **gay man**, **lgbt**, **sexuality**, **gay guy**, **gays**, cute, male, straight, **gay people**,...
graphic designer	**designer**, **graphic**, **design**, **graphic designer**, **graphic design**, **photoshop**, **logo**,...
jew	**jew**, **jewish**, **jews**, **israel**, **israeli**, **hebrew**, **judaism**, **im jewish**, orthodox, **holocaust**,...
leftist	simpler, practical, fucking, day, **leftist**, totally, time, trump, fit, bombs, women, car, bad, temps,...
liberal	**im liberal**, **liberal**, pretty, went, look, second, taken, ask, best, chance, things, matter, say,...
man	**man**, **simple man**, ampx200b, im simple, gay, fuck, left, **year old man**, real, **grown man**,...
mom	**mom**, **husband**, like, analogy, im, **baby**, **single mom**, **kids**, **children**, **son**, got, thank, thats,...
musician	**musician**, **music**, **guitar**, **band**, **musicians**, **instruments**, **instrument**, im, **musical**, **song**,...
muslim	**muslim**, **muslims**, **islam**, **religion**, **quran**, **arabic**, **allah**, **islamic**, **religious**, **arab**, country,...
parent	**parent**, bi, qualify, structures, maybe, options, amp, finishing, data, says, didnt, im, just,...
photographer	**photographer**, **photography**, **photos**, **camera**, **photo**, **lens**, **shoot**, **photographers**, **shots**,...
republican	**republican**, **republicans**, **trump**, **president**, democrats, conservative, obama, political, vote,...
socialist	du, **socialist**, like, nuclear, theres good, fund, massive, chad, youre, car, theres, bones,...
straight	ive, amazing, probably, ownership, wait, best, didnt, life, argument, wrong, words, east,...
white	key, good, large, **white dude**, **im white**, probably, day, willing, girlfriend, exact, shit, problems,...
woman	**woman**, duh, delicious, im, toys, boy, just, straight, guess ill, canadian, cute, youre, dice,...
writer	**writer**, **writing**, **write**, **writers**, **fiction**, **creative**, **wrote**, **novel**, **published**, **written**,...

Table 5.1: Most positively weighted features per in-group

et

We also look at the the top words for each identity based on the log odds ratio. The log odds ratio is defined as follows:

$$score(phr) = \log_{10} \frac{P_{phr,p}}{P_{phr,n}} \tag{5.6}$$

$$P_{phr,p} = \frac{freq_p(phr)}{\sum_{W \in C_p} freq_p(W)} \tag{5.7}$$

$$P_{phr,n} = \frac{freq_n(phr)}{\sum_{W \in C_n} freq_n(W)} \tag{5.8}$$

where $P_{phr,*}$ is the probability of phrase phr in the corpus $*$, C_* is the set of unique words (vocabulary) in the corpus $*$, and $freq_*(phr)$ is the frequency of the phrase phr in the corpus $*$. The negative examples n are selected based on the heuristic mentioned previously. The top ngrams/phrases for each identity based on the log odds ratio is shown in Table 5.2.

5.3 Obtaining Alternate Self-Identification Phrases

We created positive training instances via a relatively simple pattern based on "I am a ___". Additional alternate self-identification patterns might also be useful to augment the training data-set for every in-group. A method to obtain these alternate self-identification patterns is described in this section.

For each of the 30 in-groups that we considered, we looked at the occurrences of phrases with first person personal pronouns of the form "my ___", "As a ___, I ", "our ___", "I have a(n) ___" and its variants in both of the in-group's positive and negative training examples. The negative examples were selected according to the heuristic defined in the previous section. We ranked the phrases by the log odds ratio defined in Equation 5.6.

This ordered list still contained phrases that weren't particularly representative of any in-group. In order to filter these, the top 50 phrases with the highest *score* were taken for each of the 30 in-groups and annotated by workers on MTurk. Each HIT was designed to include an in-group along with 20 out of its top 60 phrases, each with options; "Highly likely", "Likely", "Hard to Tell" and "Unlikely". Each worker was paid $0.1 per HIT and every HIT was annotated by

3 different workers. The following worker restrictions were placed on the MTurk HIT:

- **Number of HITs approved**: Greater than or equal to: 50
- **HIT Approval Rate**: Greater than or equal to: 90
- **Location: is in**: United States

The reason for including only workers located in the United States was because of the presence of certain in-groups like "Democrat" and "Republican" which may be unfamiliar to people outside the United States, as are phrases like "my 401k".

Once the annotated results were obtained, we weighted each option as: "Highly likely" = 1, "Likely = 0.5", "Hard to Tell" = 0 and "Unlikely" = -1. After summing up the annotation scores for every phrase, we filtered out phrases with a score lesser than 2. This was to ensure that there was at least one "Highly likely" annotation for that phrase. A list of filtered phrases for each in-group is displayed in Table 5.3. The phrases "white", "agnostic" and "straight" did not have any phrases after filtering. This is probably because there aren't any phrases of the form searched that are characteristic of these in-groups.

For the "mom" and "dad" in-groups, we enforced the condition that the author's comments had to include some form of self-identification related to gender ("woman" for "mom" and "man" for "dad"). This is because a lot of phrases like "my kid", "my son", "my daughter" and its variants indicated that the author was a "parent", but did not indicate the gender ("mom" or "dad").

Table 5.4 shows the number of authors obtained using variants of the phrase "I am a(n) ___", and the increase in the number of authors for each identity when we rely on alternate self-identification statements (Table 5.3).

Self-Identification	Top words based on log odds ratio
agnostic	india, on top of, my wife, coffee, summer, an entire, indian, shop, idk, because she, gym, pm...
american	oliver, john oliver, my mom, california, boyfriend, ton of, off of, texas, my family, dating, tonight...
artist	drawing, no no, artists, makeup, penis, 3d, instagram, no no no, painting, tumblr, portfolio, tattoo...
atheist	atheist, bible, beliefs, belief, muslims, atheism, liberal, the bible, muslim, christians, islam, moral...
athlete	32, calories, muscle, sec, workout, protein, lifting, athlete, lift, squat, the gym, fitness, injury...
black	black people, white people, racism, african, askwomen, bigot, blacks, the black, bernie, in america...
cat person	kitty, my cat, anime, the cat, tbh, xd, trans, pets, dat, artist, je, van, the rules, det, tumblr, british...
christian	bible, the bible, christians, christianity, sin, of god, beliefs, belief, the church, atheist, catholic...
conservative	conservative, liberal, republican, clinton, democrats, republicans, liberals, federal, conservatives...
dad	my son, my kids, wife and, parent, my daughter, the kids, my wife and, texas, radio, lunch...
democrat	review, will probably, attractive, partner, was more, photos, reminds me, france, last night, cheese...
dog person	my dog, you love, puppy, breed, na, hahaha, request, 31, ign, hitler, your dog, linked, terrorist...
gamer	xbox, console, ps4, weapon, gameplay, dlc, of the game, fps, enemy, nintendo, combat, controller...
gay	lgbt, drag, rudy, queen, sexuality, attracted, lesbian, attracted to, identity, penis, gays, bernie...
graphic designer	designer, graphic, logo, photoshop, instagram, print, portfolio, creative, font, marketing, drawing...
jew	israel, jews, jewish, israeli, jew, palestinians, muslim, muslims, palestinian, arab, judaism, clinton...
leftist	capitalism, liberal, hillary, clinton, socialism, bernie, election, socialist, democrats, liar, economic...
liberal	le, boyfriend, er, suggestion, cheese, error, hi, tea, shoes, lawyer, hadn, ve always, it sounds like...
man	deck, the team, nyx, na, the ball, of the game, in the game, pussy, wins, fuckin, battery, 300, stats...
mom	my husband, my son, pregnant, my daughter, pregnancy, my kids, the baby, parent, babies...
musician	guitar, bass, bands, the music, jazz, musician, the song, recording, piano, chord, tracks, instrument...
muslim	islam, muslim, muslims, allah, quran, islamic, israel, prophet, pakistan, the quran, isis, saudi...
parent	parent, my son, james, pregnant, my kids, my daughter, lebron, the kids, birth, babies, the baby...
photographer	lens, photography, photographer, instagram, lenses, canon, the camera, cameras, exposure, digital...
republican	republican, republicans, clinton, democrats, conservative, bernie, federal, liberals, vote for...
socialist	capitalism, socialism, socialist, workers, bernie, economic, capitalist, democratic, clinton...
straight	agent, bsa, the bsa, jojo, nurse, hospice, my boyfriend, uj, jolyne, anime, client, estate, canada...
white	champ, jets, the jets, load, incredible, last night, the water, bathroom, the ball, pot, restaurant...
woman	my husband, trans, 32, my boyfriend, makeup, pregnant, therapy, ugh, women are, my sister...
writer	wat, writer, 32, author, writers, james, comic, fiction, horror, novel, creative, drama, uh, artist...

Table 5.2: Top ngrams obtained for each identity using log odds ratio

Alternate self-identification phrases for in-groups
american “my president”, “my 401k”, “as an american, i”
artist “my portfolio at”, “my painting”, “my own work”, “my tutorial videos”,“my art style”, “our painting”,“my art and”,“my deviantart”,“as an artist, i”,“my song”...
atheist “as an atheist i”,“our atheism”,“my atheism”
athlete “my deadlift”,“our sport”,“my lifting”,“my cardio”,“my pr”,“our swing”,“our coaches”,“our ball”,“my first meet”,“my football”...
black “my black ass”
cat person “our kitties”,“my cat loves”,“i have two cats”,“my cat does”,“my other cat”,“my cat will”,“my kitties”,“my cat would”,“my kitty”
christian “our bible”,“as a christian, i”,“our churches”,“my priest”,“my savior”,“my parish”,“my bible”,“my church is”,“as a christian, i”,“my pastor”...
conservative “as a conservative, i”,“our border”,“my christian”
dad “my first kid”,“i have a son”,“i have two kids”,“my eldest”,“my son in”,“my daughter and”,“as a parent, i”,“my daughter has”,“my kids in”,“my son has”...
democrat “as a democrat i”,“our democratic”,“our liberal”
dog person “our animals”,“my pups”,“my dog for”,“my dogs have”,“my german shepherd”,“my dog does”,“my dog loves”,“my first dog”,“my dogs and”,“my other dog”...
gamer “my sims”,“my xb1”,“i have a ps4”,“my games are”,“my gaming pc”,“my favourite game”,“my ps”,“i have the game”,“my first pc”,“my steam account”...
gay “my gayness”,“my coming out”,“my straight friends”
graphic designer “my professional designs”,“my graphic design”,“my designs”,“as a designer, i”, “my graphic”
jew “our jewish”,“our rabbi”,“as an israeli, i”,“my synagogue”,“as a jew, i”,“my rabbi”,“my bar mitzvah”
leftist “our liberal”,“our cultural”,“my liberal”
liberal “our union”,“our democratic”
man “my ex wife”,“my wife says”,“my girlfriend”,“as a man, i”
mom “my youngest is”,“my son loves”,“my daughter has”,“my son would”,“my kid was”,“our toddler”,“my first pregnancy”,“my water broke”,“my oldest son”,“my first child”...
musician “my creative”,“my mixes”,“our singing”,“my bass”,“my music is”,“our tracks”,“my first guitar”,“my pedal”,“my violin”,“our melody”...
muslim “my muslim”,“as a muslim, i”,“our prophet”,“my hijab”,“my ummah”,“our muslim”,“my mosque”
parent “my children to”,“my twins”,“my daughter”,“my four year”,“my son will”,“my kid has”,“my kids love”,“my first pregnancy”,“my son to”,“my kids will”...
photographer “my 35mm”,“our lenses”,“my lens”,“my favourite lens”,“our cameras”,“my tripod”,“my raw”,“my first camera”,“my nikon”,“my d750”...
republican “as a republican i”,“as a christian, i”
socialist “as a socialist i”
woman “my husband”,“my first boyfriend”,“my husband will”,“my boyfriend does”,“my bf is”,“our bra”,“my periods”, “my ob”,“as a girl, i”,“my cervix”...
writer “my first novel”,“our thesis”,“my first draft”,“our scripts”,“my scripts”,“my novel”,“my poetry”,“my script”,“my editor”,“my readers”...

Table 5.3: Top alternate phrases for identifying in-groups

Identity	Unique Authors for "I am a(n) ___"	Unique Authors for Alternate Self-identification Phrases
agnostic	15802	0
american	124552	85637
artist	43075	152062
atheist	127409	21764
athlete	7276	169423
black	9754	2000
cat person	4544	94597
christian	52958	118478
conservative	23697	63055
dad	20464	393690
democrat	24990	30532
dog person	5777	179359
gamer	72412	307235
gay	25838	12503
graphic designer	11384	31541
jew	12514	9284
leftist	7757	63104
liberal	52834	30282
man	330841	140613
mom	24717	229827
musician	25231	304525
muslim	15668	26794
parent	23738	516878
photographer	22392	144088
republican	20233	16460
socialist	14167	436
straight	84698	0
white	50691	0
woman	157475	288976
writer	38907	165770

Table 5.4: Number of Unique Authors on Reddit obtained for each Identity based on Self-Identification Phrases

Chapter 6

Selecting Features for Logistic Regression

In this chapter, we evaluate and compare different types of features extracted from the training data to build binary Logistic Regression models to predict in-groups for Reddit users.

- Bag of n-grams(binary) with lemmatization
- Bag of n-grams(frequency) with lemmatization
- Probability of different topics discussed using the LIWC data-set (Closed Vocabulary Approach)
- Probability of different topics discussed using Topic Modeling (Open Vocabulary Approach)

For all the methods in this chapter, we initially process the training data in the following way:

- We sample positive training examples from authors who self-identify using the phrase "I am a(n) ___".

- We sample negative training examples according to the heuristic defined in the previous chapter (If the in-group is less than 40% of the entire Reddit population, Random Negative Sampling should be used. Otherwise negative training examples should be sampled from authors who self-distance themselves from that in-group).
- We consider equal class distribution by under-sampling the majority class. The random baseline is therefore 50%.
- Maximum number of training examples per class is arbitrarily set at 3000.
- All forms of punctuation are removed

A common technique implemented in pre-processing text before any downstream task in Natural Language Processing is to lemmatize or stem words in order to reduce inflectional forms and sometimes derivationally related forms of a word to a common base-form. This is done in order to decrease the total number of unique words in our dictionary while retaining as much information as possible. The dictionary size (number of features) directly relates to the performance of the machine learning model in the downstream task, in the sense that extremely big dictionaries could slow the model down with excess unnecessary information, while extremely small dictionaries might improve the run-time of the model, but as a result of having sacrificed valuable information. Lemmatization and stemming are a way of optimizing both the model run-time and the information retained. Lemmatization relies on a lexical knowledge base like word-net to obtain the correct base form of words. However, stemming is a more crude form where words are truncated to remove inflections. Table 6.1 shows how a few words are stemmed and lemmatized differently.

word pair	**Stemmed pair**	**Lemmatized pair**	**Better Result**
(goose,geese)	(goos,gees)	(goose,goose)	Lemmatization
(meanness,meaning)	(mean,mean)	(meanness,meaning)	Lemmatization
(iphone,iphones)	(iphone,iphone)	(iphone,iphones)	Stemming

Table 6.1: Stemming and Lemmatization Examples

There are clearly advantages and disadvantages with each method:

- As lemmatization relies on a lexical knowledge base, the results are usually a lot more accurate. However, for words that are not contained in the lexical knowledge base, they are reduced to themselves, which might not always be the best solution. This can be seen with the example of the words "iphone" and "iphones", both getting reduced to themselves rather than both getting reduced to "iphone".
- As stemming is a more crude form where words are chopped without looking at parts of speech, there might be several incorrect results like "goose" and "geese" getting reduced to "goos" and "gees" rather than both getting reduced to the same base. However, stemming performs better on words like "iphone" and "iphones" which are typically not present in a lexical knowledge base.

6.1 Binary Bag of n-gram Features

This is a very common method for extracting features from text. The document is first prepared by lemmatizing every word, following which it is tokenized into 1 word (unigram), 2 word (bigram) and 3 word (trigram) phrases. The document is then represented by a vector, the size of which is equal to the number of unique n-grams (vocabulary). Each element in the document vector is either a 1 or a 0 denoting the presence or absence of the n-gram at the particular index. The 20,000 most frequently occurring n-grams in the training-set are only considered. A simple logistic regression model is trained using these set of features.

We will compare this model to the model generated in the next section where frequency bag of n-gram features are used.

6.2 Frequency Bag of n-gram Features

This method of obtaining features is similar to the binary bag of n-gram features except for the fact that instead of each element in the vector representing the presence(1) or absence(0) of an n-gram, it denotes the normalized frequency (probability) with which the n-gram was observed in the document. The 20,000 most frequently occurring n-grams in the training-set are only considered. The performance of a logistic regression model trained on these set of features is compared to that trained on binary features in Table 6.2

The cross validation accuracies for the models trained on the binary features are higher. However, from the Test-set accuracies, frequency features perform better for some in-groups while binary features perform better for others suggesting that there isn't a better feature representation among the two.

6.3 LIWC Features

In this section, we build a document representation based on categories defined in pre-constructed word-category lexicons. A popular word-category lexicon is the Linguistic Inquiry and Word Count (LIWC) data-set developed by researchers with interests in social, clinical, health and cognitive psychology. The 73 language categories in LIWC were created to capture people's social and psychological states. As Yla R. Tausczik and James W. Pennebaker [51] state:

> " Empirical results using LIWC demonstrate its ability to detect meaning in a wide variety of experimental settings, including to show attentional focus, emotionality, social relationships, thinking styles, and individual differences"

Using LIWC categories to engineer features for each document is a closed-vocabulary approach as we rely on *a priori* word category human judgments. The LIWC data-set contains some categories that encapsulate large topics (like family, money, space, etc.) and some cat-

Self-Identification	**Cross Validation Accuracy Binary**	**Cross Validation Accuracy Normalized Frequency**	**Test Accuracy Binary**	**Test Accuracy Normalized Frequency**	**Test Majority Baseline**
agnostic	0.56	0.53	0.52	0.54	0.58
american	0.85	0.80	0.79	0.82	0.96
artist	0.78	0.72	0.80	0.78	0.85
atheist	0.78	0.68	0.80	0.76	0.86
athlete	0.80	0.76	0.86	0.85	0.92
black	0.81	0.79	0.91	0.95	0.96
cat person	0.62	0.60	0.53	0.54	0.58
christian	0.76	0.68	0.62	0.61	0.74
conservative	0.80	0.77	0.74	0.77	0.82
dad	0.79	0.74	0.7	0.80	0.87
democrat	0.57	0.56	0.53	0.44	0.53
dog person	0.60	0.58	0.60	0.51	0.78
gamer	0.76	0.70	0.57	0.58	0.70
gay	0.82	0.78	0.85	0.82	0.89
graphic designer	0.82	0.75	0.81	0.72	0.95
jew	0.81	0.75	0.90	0.80	0.98
leftist	0.85	0.80	0.67	0.65	0.65
liberal	0.58	0.57	0.51	0.48	0.55
man	0.68	0.66	0.70	0.62	0.63
mom	0.85	0.86	0.69	0.78	0.89
musician	0.81	0.76	0.89	0.89	0.9
muslim	0.85	0.80	0.97	0.96	0.99
parent	0.77	0.74	0.65	0.66	0.77
photographer	0.83	0.75	0.78	0.80	0.90
republican	0.81	0.77	0.78	0.75	0.85
socialist	0.84	0.80	0.76	0.79	0.77
straight	0.49	0.55	0.79	0.46	0.89
white	0.63	0.61	0.66	0.63	0.96
woman	0.77	0.74	0.70	0.68	0.63
writer	0.78	0.72	0.84	0.78	0.85

Table 6.2: Cross Validation and Test Accuracies for Binary and normalized Frequency Bag of ngram Identity-Models

egories that encapsulate different parts of speech (articles, auxverbs, pronouns, etc.). It can therefore be used to analyze both topics and language style of documents.

Each document in the training-set is represented as a vector of size 73, with each element representing the probability of the LIWC category at that index being discussed in the document. The probability of a LIWC category in a document is defined as:

$$p(category|document) = \frac{\sum_{word \in category} (freq(word, document))}{\sum_{word \in vocab(document)} (freq(word, document))} \tag{6.1}$$

While analyzing most of the LIWC categories (except for "FOCUSFUTURE", "FOCUSPAST" and "FOCUSPRESENT"), we stem the text of all authors as the words for these LIWC categories in the data-set are stemmed.

Using these features, we build a logistic regression model for every in-group and compare the cross-validation accuracies to models trained on binary features as a from of reference. They are provided in Figure 6.1

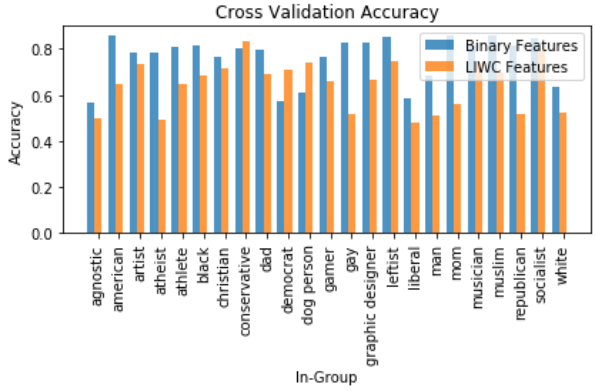

Figure 6.1: Cross Validation Accuracy for comparing binary ngram features vs LIWC topic probability features

From the results, it can be inferred that using LIWC topic probabilities as features performs as badly as random (cv accuracy of around 0.5) for certain in-groups like "white", "man",etc. This suggests that there aren't many topics discussed with a higher probability within these

groups compared to their counterparts.

On the other hand, in-groups like "conservative", "artist", "socialist","musician" etc. that have cross validation accuracies much higher than random (0.5) suggest that certain topics are discussed with a higher probability within these groups.

The top 3 positively weighted LIWC topics by the logistic regression model for some in-groups are shown in Table 6.3.

Identity	**Topic**	**Top Words**
conservative	RELIG	god, christian, hell, cathol, church, jesu, moral, belief, holi,...
conservative	FAMILY	famili, parent, wife, dad, babi, mom, brother, son, mother...
conservative	WE	we, our, us, let, ourselv, weve, let, we'r, we'v
artist	I	i, my, me, myself, im, mine, id, ive, idk, imma, ikr, ima, ili, idc,...
artist	FEEL	feel, hard, cool, hand, hair, skin, felt, pain, hot, fire, touch,...
artist	ADJ	as, more, thank, will, than, play, same, help, most, great, artist,...
socialist	ANGER	fuck, shit, kill, hate, war, fight, stupid, attack, hell
socialist	POWER	up, over, down, great, help, govern, best, polit, power, war, kid,..
socialist	NEGEMO	:, fuck, shit, bad, problem, long, wrong, kill, hate
musician	HEAR	say, sound, said, song, listen, hear, heard, phone, voic, speak,...
musician	WE	we, our, us, let, ourselv, weve, we'r, we'd
musician	FRIEND	guy, friend, dude, follow, date, girlfriend, contact, buddi, confid,...
muslim	RELIG	islam, god, allah, christian, hell, quran, belief, jew, holi, faith,...
muslim	FOCUSPRESENT	is, are, can, be, have, do, think, get, know, has, now, see, want,...
muslim	FAMILY	famili, parent, marri, brother, bro, sister, dad, mom, wife,...

Table 6.3: Top 3 positively weighted LIWC categories for some identities

6.4 Topic Modelling Using LDA

In this section, we create a list of 500 topics from the Reddit data-set using a machine learning model for topic modelling called Latent Dirichlet Allocation(LDA). Since we use the Reddit data-set to find topics rather than rely on a predefined set of topics, this is an open vocabulary approach or a data driven approach.

LDA discovers latent topics by identifying groups of words in the corpus that frequently occur together within documents. The basic idea is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words. α and β are the Dirichlet priors:

- α: parameter on the per document topic distribution. High α means a document may have many topics, low α means a document has only one or a few topics.
- β: parameter on the per topic word distribution. High β means that each topic will contain a mixture of most of the words. Low β means that each topic will contain a mixture of just a few of the words.
- θ: the topic distribution for a document.
- Z is used to notate each topic which is assigned to each word (w).

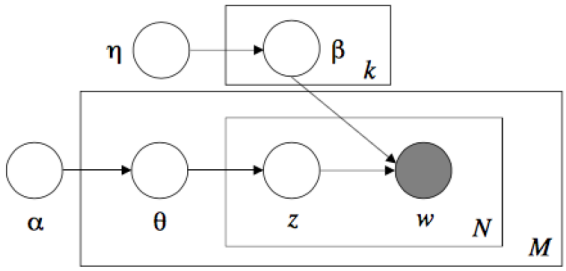

Figure 6.2: Latent Dirichlet Allocation plate notation

We fed the LDA model 20 million comments randomly sampled from the Reddit corpus to generate a list of 500 topics. The comments were directly fed into the LDA model without any pre-processing (ex. removing stop-words, lowercasing, stemming/lemmatizing,etc). Each document (author) in the training-set was then represented as a vector of size 500, with each element representing the probability of the corresponding topic at that index being discussed

in the document. The probability of a topic in a document is defined as:

$$p(topic|document) = \sum_{word \in topic} p(topic|word) * p(word|document) \tag{6.2}$$

Here, $p(word|document)$ is the normalized word use by that document and $p(topic|word)$ is the probability of the topic given the word, which is provided by the LDA model.

A logistic regression model is built using these document-topic-probability feature representations for every in-group and the cross validation accuracies are compared to models trained on Binary Features as a form of reference. They are provided in Figure 6.3.

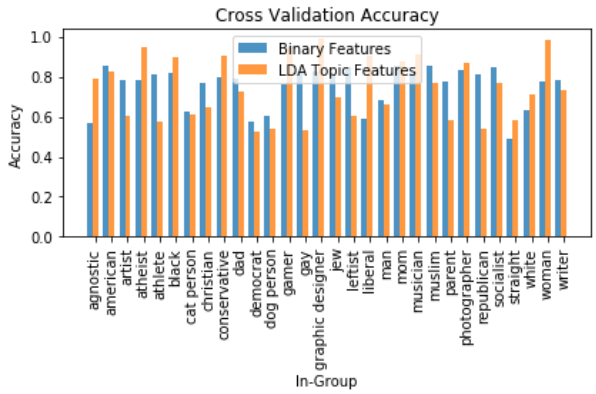

Figure 6.3: Cross Validation Accuracy for comparing binary ngram features vs LDA topic probability features

Similar to the last section, cross validation accuracies for some in-group models are near 0.5 (random) suggesting that there aren't any topics discussed with a higher probability within these groups. On the other hand, certain in-groups like "woman", "atheist", "graphic designer", "photographer", etc. have extremely high cross validation accuracies suggesting that certain topics are discussed with a higher probability within these groups.

The top 8 positively weighted LDA topics are represented for some in-groups in Tables 6.4, 6.5

and 6.6.

Topic1	**Topic2**	**Topic3**	**Topic4**	**Topic5**	**Topic6**	**Topic7**	**Topic8**
gun	people	people	ain't	love	john	cat	ass
guns	feel	culture	shit	i'd	she's	cats	face
people	person	american	fuck	i'm	love	pet	kick
crime	abuse	white	fuckin	hate	show	dog	shit
mass	it's	country	tho	absolutely	rachel	love	fuck
control	anger	world	gud	hear	girl	cute	fucking
shooting	behavior	america	git	gotta	nick	pets	guy
carry	angry	americans	dat	awesome	sarah	kitty	punch
laws	things	it's	bitch	prefer	named	cage	kicked
violence	don't	cultural	ass	great	names	dogs	pain
don't	hurt	countries	gotta	man	season	vet	bitch
firearms	emotional	western	fam	haha	amy	kitten	beat
weapons	bad	history	bout	amazing	jane	animal	man
murder	feelings	native	y'all	loved	lisa	food	back
violent	life	society	lol	happy	emily	rat	dude
shootings	abusive	race	boi	personally	morgan	animals	kicking
firearm	control	black	gonna	fell	girls	rats	slap
ban	kind	cultures	man	yeah	oliver	litter	yeah
illegal	fear	don't	nigga	idea	dean	mine	bad
crimes	can't	population	damn	honestly	episode	pig	lol

Table 6.4: Top 8 positively weighted Topics for the in-group "woman"

Topic1	**Topic2**	**Topic3**	**Topic4**	**Topic5**	**Topic6**	**Topic7**	**Topic8**
israel	dead	bitcoin	family	feel	relationship	class	chicken
jews	horse	wallet	mother	i'm	don't	classes	sauce
jewish	alive	btc	father	it's	it's	math	add
middle	body	block	son	feeling	things	school	make
saudi	zombie	coins	wife	don't	feel	exam	cook
people	horses	transactions	daughter	life	time	study	salt
arabia	walking	transaction	husband	makes	life	year	rice
east	bodies	segwit	married	i've	you're	college	cooking
israeli	zombies	mining	parents	things	make	students	oil
world	it's	network	kids	time	doesn't	taking	meat
state	back	chain	dad	happy	person	courses	recipe
land	brain	miners	friends	people	good	professor	pan
arab	grave	blockchain	brother	felt	work	major	cheese
palestinians	corpse	fees	years	can't	situation	test	pepper
jew	die	blocks	mom	love	love	semester	eggs
peace	goat	cash	sister	hope	talk	grade	put
palestinian	died	coin	children	pain	trust	studying	oven
country	buried	exchange	child	sad	feelings	time	hot
countries	he's	fork	life	make	sounds	i'm	garlic
palestine	left	ethereum	members	bad	can't	student	beef

Table 6.5: Top 8 positively weighted Topics for the in-group "graphic designer"

Topic1	**Topic2**	**Topic3**	**Topic4**	**Topic5**	**Topic6**	**Topic7**	**Topic8**
party	bag	watch	attack	birth	eye	dps	you're
vote	box	it's	block	child	eyes	boss	argument
brexit	put	nice	giraffe	abortion	nose	tank	point
labour	bags	great	heavy	control	blood	class	don't
government	stuff	love	hit	baby	surgery	raid	i'm
election	carry	watches	dodge	life	skin	fight	it's
voted	boxes	mine	light	it's	it's	healer	that's
people	pack	good	game	women	face	play	doesn't
parties	plastic	case	combo	pregnant	cut	time	make
leave	back	condition	character	don't	ear	wow	wrong
referendum	pocket	i've	play	sex	glasses	healing	fact
tories	inside	i'm	attacks	woman	body	group	arguing
corbyn	bring	wrist	parry	pregnancy	blind	legion	isn't
majority	empty	time	good	children	vision	gear	talking
support	small	wear	it's	babies	ears	bosses	can't
parliament	full	strap	guard	people	penis	classes	making
leader	backpack	pretty	characters	born	pain	good	arguments
political	don't	gold	damage	body	contact	spec	logic
country	trash	i'd	move	mother	throat	mythic	literally
tory	things	dial	moves	choice	scar	content	people

Table 6.6: Top 8 positively weighted Topics for the in-group "photographer"

Chapter 7

Conclusion and Future Work

Reddit represents a rich source of data for researchers interested in sociological or linguistic phenomena related to self-identification. We have shown that it's possible to identify hundreds of self-identified in-groups and out-groups as wide ranging as *dog lovers*, *plumbers*, *feminists* and *trans men* with high precision. Leveraging the fact that Reddit is organized into discrete topic forums, we have shown that self-identification happens most often in places where that group membership is relevant (e.g. self-identifying one's gender happens most often in discussions of gender and sexuality). By analyzing users who self-identify with multiple in-groups and out-groups, we show interesting co-occurrence patterns. Finally, we define a heuristic that should be followed for sampling negative training examples when building binary classification models for any identity/in-group. We also show that using topic probability features provide better performance for certain in-groups.

Some of the future work could:

- look at combining various feature representations (LIWC topic probabilities, LDA topic probabilities and ngrams (unigrams, bigrams and trigrams) among others) while building identity classifiers, thereby identifying the optimum set of features.

- look at augmenting the positive training examples for each in-group with alternate self-identification phrases (some of which are described in Chapter 5).
- compare the attitudes/sentiments of these different in-groups towards various topics and entities (Controversial topics like abortion, climate change, animal testing,...; Products; Laws; People;...).

Bibliography

[1] 6% of online adults are reddit users. https://www.pewinternet.org/2013/07/03/6-of-online-adults-are-reddit-users/. Accessed: May 2, 2019.

[2] Associate professor emily falk personal website. https://www.asc.upenn.edu/people/faculty/emily-falk-phd/. Accessed: May 2, 2019.

[3] Google bigquery homepage. https://cloud.google.com/bigquery/. Accessed: April 7, 2019.

[4] The internet archive. https://archive.org/details/2015_reddit_comments_corpus. Accessed: May 2, 2019.

[5] Jason baumgartner twitter. https://twitter.com/jasonbaumgartne. Accessed: May 2, 2019.

[6] pushshift.io main page. https://pushshift.io/. Accessed: May 2, 2019.

[7] Reddit post for the public dataset
. https://www.reddit.com/r/datasets/comments/3bxlg7/i_have_every_publicly_available_reddit_comment/. Accessed: May 2, 2019.

[8] Top sites in united states. https://www.alexa.com/topsites/countries/US/. Accessed: May 2, 2019.

[9] M. D. Adrian Benton, Raman Arora. Learning multiview embeddings of twitter users. 2013.

[10] J. C. Aron Culotta, Nirmal Kumar Ravi. Predicting the demographics of twitter users from website traffic data. 2015.

[11] C. Beller, R. Knowles, C. Harman, S. Bergsma, M. Mitchell, and B. Van Durme. I'm a Belieber: Social roles via self-identification and conceptual attributes. In *ACL*, 2014.

[12] V. Beretta, T. Cribbin, D. Maccagnola, and E. Messina. An interactive method for inferring demographic attributes in twitter. 2015.

[13] S. Bergsma, M. Dredze, T. W. Benjamin Van Durme, and D. Yarowsky. Broadly improving user classification via communication-based name and location clustering on twitter. 2013.

[14] J. D. Burger, J. Henderson, G. Kim, and G. Zarrella. Discriminating gender on twitter. 2011.

[15] N. Cesare, C. Grant, and E. O. Nsoesie. Detection of user demographics on social media: A review of methods and recommendations for best practices. *CoRR*, abs/1702.01807, 2017.

[16] B. V. Charley Beller, Craig Harman. Predicting fine-grained social roles with selectional preferences. 2014.

[17] M. D. Choudhury, S. Counts, and E. Horvitz. Predicting postpartum changes in emotion and behavior via social media. 2013.

[18] B. CY, C. M, P. L, M. LC, T. A, and B. JS. Publicly available online tool facilitates real-time monitoring of vaccine conversations and sentiments. 2016.

[19] T. S. David Bamman, Jacob Eisenstein. Gender identity and lexical variation in social media. 2014.

[20] F. C. David E. Losada. A test collection for research on depression and language use. 2016.

[21] M. De Choudhury and E. Kiciman. The language of social support in social media and its effect on suicidal ideation risk. In *ICWSM*, 2017.

[22] M. De Choudhury, E. Kiciman, M. Dredze, G. Coppersmith, and M. Kumar. Discovering shifts to suicidal ideation from mental health content in social media. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, pages 2098–2110, New York, NY, USA, 2016. ACM.

[23] A. R. Dennis and S. T. Kinney. Testing media richness theory in the new media: The effects of cues, feedback, and task equivocality. *Information systems research*, 9(3):256–274, 1998.

[24] E. Di Minin. A framework for investigating illegal wildlife trade of social media with machine learning. 2019.

[25] S. L. S. Dirk Hovy. The social impact of natural language processing. 2016.

[26] M. Gjurković and J. Šnajder. Reddit: A gold mine for personality prediction. In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, pages 87–97. Association for Computational Linguistics, 2018.

[27] J. J. Greg P.Griffin. Where does bicycling for health happen? analyzing volunteered geographic information through place and plexus. 2015.

[28] P. Grice. Logic and conversation. In *Cole, P.; Morgan, J. Syntax and semantics. 3: Speech acts.*, pages 41–58. Academic Press, 1975.

[29] S. H. The association of suicide-related twitter use with suicidal behaviour: a cross-sectional study of young internet users in japan. 2015.

[30] P. S. Hamidreza Alvari, Soumajyoti Sarkar. Detection of violent extremists in social media. 2019.

[31] S. Hassanpour, N. Tomita, T. DeLise, B. Crosier, and L. A. Marsch. Identifying substance use risk based on deep neural networks and instagram social media data. 2019.

[32] U. A. Jalal S. Alowibdi and P. Yu. Empirical evaluation of profile characteristics for gender classification on twitter. 2013.

[33] D. JC, H. H, K. AE, C. L, and A. J. The use of social media by state tobacco control programs to promote smoking cessation: a cross-sectional study. 2014.

[34] I. W. Jisun An. Demographics and hashtag use on twitter. 2016.

[35] W. Liu and D. Ruths. What's in a name? using first names as features for gender inference in twitter. *AAAI Spring Symposium - Technical Report*, pages 10–16, 01 2013.

[36] P. S. Ludu. Inferring gender of a twitter user using celebrities it follows. *CoRR*, abs/1405.6667, 2014.

[37] X. Ma, J. Hancock, and M. Naaman. Anonymity, intimacy and self-disclosure in social media. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 3857–3869. ACM, 2016.

[38] C. R. Marc-André Kaufhold. Cultural violence and peace in social media. 2019.

[39] D. J. McIver1, J. B. Hawkins, R. Chunara1, A. K. Chatterjee, A. Bhandari, S. H. J. Timothy P Fitzgerald, and J. S. Brownstein. Characterizing sleep issues using twitter. 2015.

[40] D. G. Myers and H. Lamm. The polarizing effect of group discussion: The discovery that discussion tends to enhance the average prediscussion tendency has stimulated new insights about the nature of group influence. *American Scientist*, 63(3):297–303, 1975.

[41] M. Pennacchiotti and A.-M. Popescu. A machine learning approach to twitter user classification. volume 11, 01 2011.

[42] K. W. Prier, M. S. Smith, C. Giraud-Carrier, and C. L. Hanson. Identifying health-related topics on twitter: An exploration of tobacco-related tweets as a test topic. 2011.

[43] C. Rajski. A metric space of discrete probability distributions. *Information and Control*, 4(4):371–377, 1961.

[44] K. S. Salathé M. Assessing vaccination sentiments with online social media: implications for infectious disease dynamics and control. 2011.

[45] A. Schwartz, J. Eichstaedt, M. Kern, L. Dziurzynski, S. Ramones, M. Agrawal, A. Shah, M. Kosinski, D. Stillwell, M. Seligman, and L. Ungar. Personality, gender, and age in the language of social media: The open-vocabulary approach. 2013.

[46] I. Sekulić, M. Gjurković, and J. vSnajder. Not just depressed: Bipolar disorder prediction on reddit. 2018.

[47] J. W. P. J. S. Shlomo Argamon, Moshe Koppel. Automatically profiling the author of an anonymous text. 2009.

[48] S. T. Simon Suster and W. Daelemans. A short review of ethical challenges in clinical natural language processing. 2017.

[49] H. Tajfel. Social identity and intergroup behaviour. *Information (International Social Science Council)*, 13(2):65–93, 1974.

[50] C. Tan, V. Niculae, C. Danescu-Niculescu-Mizil, and L. Lee. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the 25th international conference on world wide web*, pages 613–624. International World Wide Web Conferences Steering Committee, 2016.

[51] Y. R. Tausczik and J. W. Pennebaker. The psychological meaning of words: Liwc and computerized text analysis methods. In *Journal of Language and Social Psychology*, 2010.

[52] J. C. Turner. Self-categorization theory and social influence. *The phychology of group influence*, pages 233–275, 1989.

[53] J. B. Walther. Interpersonal effects in computer-mediated interaction: A relational perspective. *Communication research*, 19(1):52–90, 1992.

[54] E. A. F. W. Xin Chen, Yu Wang. A comparative study of demographic attribute inference in twitter. 2015.

[55] S. Xu, C. Markson, K. L Costello, C. Y Xing, K. Demissie, and A. Llanos. Leveraging social media to promote public health knowledge: Example of cancer awareness via twitter. *JMIR Public Health and Surveillance*, 2:e17, 04 2016.

[56] A. Yates, A. Cohan, and N. Goharian. Depression and self-harm risk assessment in online forums. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2968–2978, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[57] Q. You, S. Bhatia, T. Sun, and J. Luo. The eyes of the beholder: Gender prediction using images posted in online social networks. pages 1026–1030, 12 2014.

[58] F. A. Zamal, W. Liu, and D. Ruths. Homophily and latent attribute inference: Inferring latent attributes of twitter users from neighbors, 2012.