The Communicative Effects of Anonymity Online

Caleb Johnson
Natural Language Processing
Salt Lake City, UT
calebdeejohnson@gmail.com

Abstract

An ever-increasing number of Americans have an active social media presence online. As of March 2020, an estimated 79monthly users of some sort. Many of these online platforms allow users to operate anonymously which could potentially lead to shifts in communicative behavior. I first discuss my compilation process of the Twitter Anonymity Dataset (TAD), a human-classified dataset of 100,000 Twitter accounts that are categorized by their level of identifiability to their real-world agent. Next, I investigate some of the structural differences between the classification levels and employ a variety of Natural Language Processing models and techniques to shed some light on the behavioral shifts that were observed between the levels of identifiability.

Introduction

Over the past half century, emerging communication advancements have helped foster a dramatic shift in the human experience. Adoption and improvement in computer networking understanding have helped to lead to an ever-closer world, and somewhat more recently, have led to the increasing interconnectedness in the form of social media platforms. As of March 2020, an estimated 79% of Americans were active monthly social media users and that number is predicted to increase of the coming years. Two of the most commonly adopted of these sites, Facebook and Twitter, observed 2.38 billion and 330 million average monthly users respectively as of the end of 2019 alone [1].

Some of these social media platforms, such as Facebook, historically have held a Real-Name policy. This policy requires users to register their real names when creating an account on their platform [2]. A 2014 Facebook press release identified their reasoning as "[people's desire to] know whom they are connecting with", as well as to improve the content quality of the site by "decreasing spam, bullying, and hacking" [3]. This decision has also led to a large amount of controversy and unrest as many users worry of privacy breaches as their online activity is stored and analyzed.

Twitter, on the other hand, does not impose a Real-Name policy, and simply requires their users to register with a unique pseudonym [4]. This policy has led to many Twitter users creating accounts that have no observable tie to their real name or identity. Some users opt for this in an effort to create a unique online persona or identity, while others are seeking to maintain their anonymity during their online activity.

Social media communication is showing no signs of diminishing, as the number of Americans actively using social media is expected to increase to over 82% by 2022. The effects of this communication, and the potential consequences of doing so anonymously, are and will continue to be an important consideration for maintaining the exchange of ideas that the internet allows.

Current Research

The effects of anonymity have long been the subject of psychological and sociological studies and papers. APS Fellow Philip Zimbardo of Stanford University famously conducted an experiment on the effects of anonymity in regard to harmful decision making in 1969. Zimbardo found that people who felt greater anonymity were twice as likely to comply in behavior that led to the harming of another person [5]. In the exploding world of natural language research there has been some research conducted in the field of anonymity online.

The most interesting previous research on this topic was conducted by a group of professors and students at New York University and was titled "On the Internet, Nobody Knows You're a Dog: A Twitter Case Study of Anonymity in Social Networks" [6]. In the paper, the group analyzes the correlation between connecting with controversial groups online in regard to anonymity. They provided presumably the first data-driven analyses on the topic of online anonymity and behavior.

While their research was very fascinating, I had some concerns with the parameters by which they created their dataset as well as the reach of their project. They categorized accounts solely based upon how much of a full name was listed and did not take into account other factors such as profile picture, location specificity, biographical clues, and verification status. This decision was well described and had its merits, but I believe that more robust classification criteria, while potentially more subjective, will provide us with more true results. In addition, they employed Amazon's Mechanical Turk program to

get human classification on their data [7]. While the program is an effective way to crowd-source large amounts of data classification, I did worry about inconsistencies in the data and format as a result of many untrained classifying agents.

The New York University paper also did fascinating research into the behavior of anonymous agents online and observed the material with which they engaged. They observed what accounts they followed, but I am curious more in the ways in which they communicate. Through natural language processing techniques, I hope to gain a better understanding of how they think rather than simply with what they engage.

Dataset Creation

For the purpose of the research I wished to conduct, I decided to compile my own dataset that modeled and credited some of the structure of the New York University dataset. I believed that their standards were a bit too liberal on what constituted an identifiable user and wanted to create a more nuanced dataset for this distinction. In addition, I had concerns on how to handle nonhuman and group accounts.

I utilized the developer tools credentials provided by the Twitter API to stream tweets to create my dataset. I pulled accounts that had recently tweeted (a non-retweet [8]) and had at least 100 tweets to the account. This was in an attempt to build a dataset of users that had at least some footprint on the forum and voiced opinions that were solely their own and not retweeted from other sources. Additionally, I limited my sampling to accounts that were in English as my model was predicated on the analysis of language and this would be difficulty if we were dealing with different languages and I only logged accounts that had a location listed in the United States. One of the subsets of my analysis was how free users felt they were to express opinions and I hoped to select users that at least somewhat similar cultural backgrounds and freedoms of expression. I pulled tweets in blocks of 20,000 at 6 times slots throughout the day, that were chosen by means of a random number generator, in an effort to avoid noise based on the demographics that would be online at certain times of day. I pulled around 120,000 accounts to consider, but due to decisions to not include certain account types in the classification paradigm the final dataset was just over 100,000.

The most salient consideration in the classification process was how easily the user could be identified by their displayed account information. For instance, if a user were to post something highly controversial or upsetting to someone, how easily could the offended party track down who the user was based on the information provided in their account. We only considered information in the account's profile, so any identifying features in the actual content of their tweets was not considered in this classification. This made the categorization slightly more subjective, but I believe that the human intuition behind the classification allowed us to come to more accurate results. For instance, a "Dan Smith" may be harder to identify than a "Janice Stubblefield" because of the frequency of same name occurrences. Likewise, the specificity of the user's biography may be weighted heavily in being able to identify a certain user, while a link to one's LinkedIn profile would be heavily weighted in another user's case. In all cases, we had no way to ensure the authenticity or correctness of any and fields and had to exercise best judgement and intuition in the classification process.

Account Data Considered

- **Name:** The name the user opted to provide was very influential in our classification. We had no way to verify the authenticity of the name claims, but the field still provided insight into how identifiable the user was attempting to be.

- **Username:** Similar to the name, however potential inclusion of birth year, occupation, and further insight into their full legal name often appeared in this field.

- **Profile Picture:** Users are provided the option to upload a picture to represent them, and in absence of this a generic Twitter egg is supplied. This field was not restricted to a photo of themselves and could be any image that they selected.

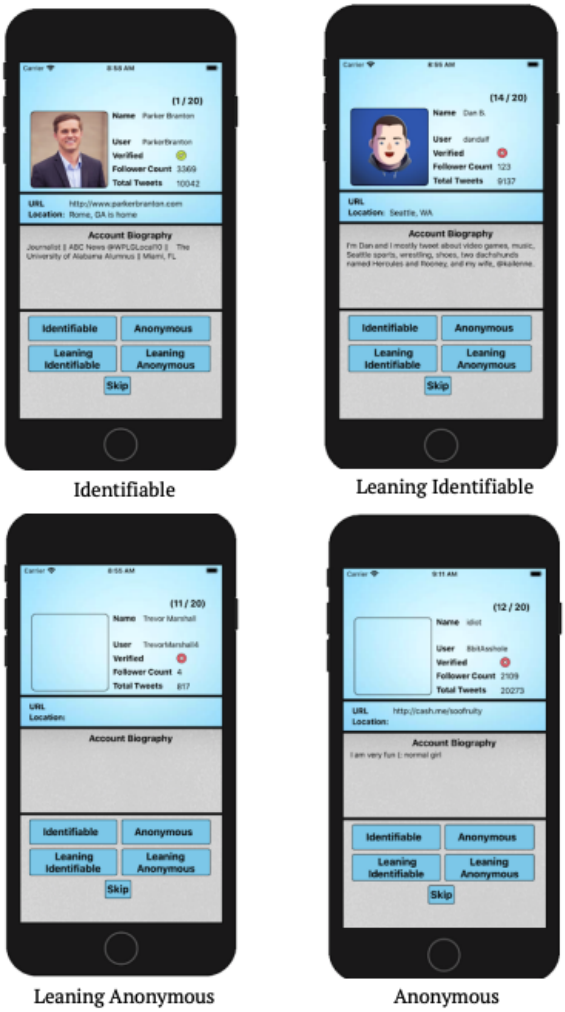
- **Location:** Users were provided the optional field to provide a location for where they were based. This field is not fixed by geo-tagging the user and thus has no guarantees or standard for what it can be listed with. That said, it was still fairly influential depending on the specificity of the location. Someone who provided a location of "Meridian, Idaho" is far more identifiable than someone who simply provided "The Midwest".

- **URL:** Users were provided the optional field to provide a URL link to a different online platform. Users who provided references to a Facebook account, LinkedIn account, or similar site that imposes a real-name policy were more heavily weighted.

- **Verification Status:** A Twitter account can be verified by the Twitter company to ensure users that an account of public interest is authentic. While the criteria to be substantial enough to be "verifiable" is murky, the status provided was influential in ensuring that a user could be identified to their real-world agent.

- **Biography:** The efficacy of the biography section was at times subjective, but many users opted to provide their employer, job title, identifiable aspects of their life, and more in their biography that allowed us to feel more comfortable in our classification.

Figure 1
Classification Levels Example



Non-Human Accounts

While accounts that were tied to popular culture references or definitively ambiguous details were allowed, we did attempt to remove accounts that were clearly affiliated with groups of people rather than individuals from the dataset. One of central goals was to see the behavior of individuals under anonymity, so we utilized a skip button to avoid having to classify groups. Many accounts are tied to real-world companies and organizations rather than individuals, so we skipped those accounts instead of attempting to classify them. This means they were not counted towards the 100,431 accounts that we classified in our dataset. In total, our classifiers opted to skip over 4,000 accounts and thus we can make a prior estimate that they composed about 4

Classifiers and Verification

In total, seven different people participated in the creation of the dataset. In an effort to provide consistent classification across agents, we distributed training slides that walked through basic criteria for classification with example accounts so that we could be on a more level understanding when classifying. We also had a group text where we discussed potential cases that may be ambiguous to classify and discussed group standards for when such cases occurred. The deepest thanks go to the family and friends who assisted me in the classification process, and in particular, to my aunt Sandra who classified over 50,000 unique Twitter accounts alone. This was the lion share of the process and she cannot be credited enough for its completion.

In addition, verification batches were distributed to all classifiers to quantify how closely we were classifying to one another. My own classifications were used as the benchmark and we observed how frequently we were classifying in harmony. Our analysis grouped both identifiable and leaning identifiable together as a form of "positive" classification and anonymous and leaning anonymous together as a form of "negative" classification. In total, we classified in unison approximately 91.2classification 93.5identifiable (meaning we had much more firm criteria for what it meant to be identifiable). The confusion matrix for our classifications can be viewed in Table 1.

Full Dataset

After classification, we cycled through the full 100,431 accounts and pulled up to 20 non-retweets from each for our full working dataset. This resulted in a dataset of 1.5 million unique tweets that would leverage for the rest of our analysis, with the proportion breakdown of levels mirroring the smaller account dataset very closely.

	Observed Positive	Observed Negative
Expected Positive	93.5%	6.5%
Expected Negative	9.3%	90.7%

Table 1
Classification Similarity Table

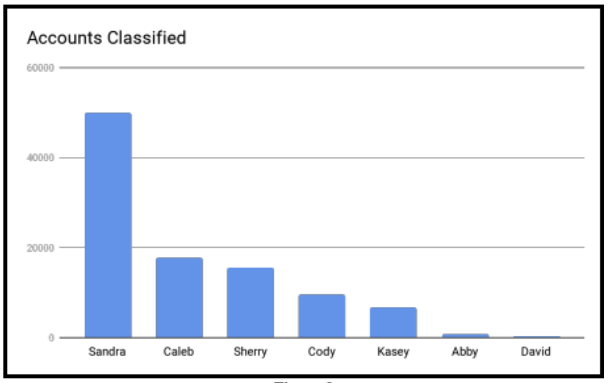


Figure 2
Classifier Count Breakdown

Results

Our classification criteria resulted in a majority of users being classified as definitively anonymous from their real-world agent. This is due to a belief that there would simply be no means to connect the user to their real-world agent based on the information provided in their profile. Leaning anonymous received by far the next largest share of the users, and in fact

the two levels of anonymous users accounted for 84.2platform. Only 8.7criteria. A graphical representation of this proportion breakdown can be seen in Figure 3.

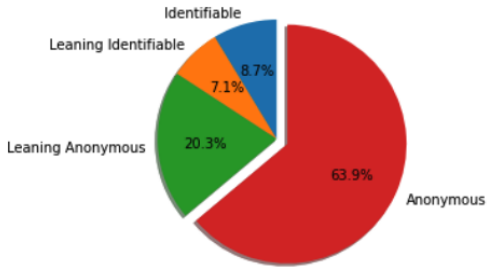


Figure 3
Classification Level Percentages

Tweet Structure

Some of the most surface level analysis of the communicative behaviors was the metadata on the tweets. We observed distinct trends in the structure of the tweets between the levels of identifiability which we found noteworthy. In this section we will observe data about the tweets themselves, including average tweet lengths, usage frequencies, and more.

Tweet Length

We observed a distinct increase in the average length of a tweet based upon the level of identifiability of the user. Users who were distinctly identifiable used an average of 88.5 characters per tweet while anonymous users only averaged 71.9 characters per tweet. A graphical representation of this breakdown can be seen in Figure 4. While we cannot make any definitive claims on the cause of this, we did observe a few trends that may help explain this.

One potential cause of this was the increased usage of superlatives among anonymous users. Anonymous users were 34(conditioned by containing words ending in '-est', 'most', and/or 'wrong') than identifiable users were. This habit of speaking more in absolutes could lead to more concise, less nuanced opinion sharing that may lower the average tweet length overall. Further potential explanation of this trend is the increased prevalence of quantifier words in identifiable users' tweets. For instance, we noticed distinct increases in the frequencies of the phrases 'I think' and 'my opinion' as users became increasingly identifiable. 0.91identifiable users conditioned their thoughts with 'I think', while anonymous users only used this condition 0.49other potential quantifier words, and at rates that proved statistically significant even when controlling for the increased average characters between the levels. Identifiable users also used the phrases 'I feel' and 'my feelings' more often, which increasingly may imply that users may have felt the need to condition their thoughts, while anonymous users felt more freedom to voice opinions in a concise manner. A frequency breakdown of this trend can be found in Table 2

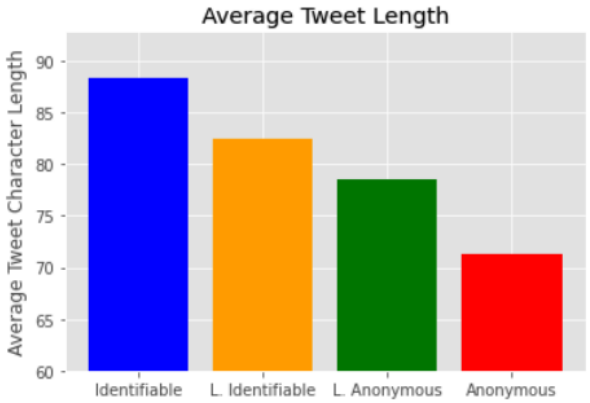


Figure 4
Average Tweet Lengths

	Frequency of 'My opinion'	Frequency of 'I think'
Identifiable	0.25%	0.91%
Leaning Identifiable	0.21%	0.72%
Leaning Anonymous	0.17%	0.56%
Anonymous	0.14%	0.49%

Table 2
Frequencies of Opinion Quantifiers

Vulgarity

The Office of Communication (commonly known as Ofcom) is the government-sanctioned regulatory authority for media in the United Kingdom. In their assessment of the appropriateness of language, they ordered vulgar words and phrases by four levels of severity: mild, medium, strong, and strongest [9].

In the data we observed only marginal increases in the usage of mild and medium vulgarity between anonymous and identifiable users. The distinction became much more profound among vulgar phrases and words classified as strong and strongest, however. The data found anonymous users were two and a half times more likely to use the strongest level of vulgarity in their tweets. In particular, the f-word saw the greatest relative increase among all the words and phrases labeled as vulgar by Ofcom with a relative increase of over three times.

Emoji Analysis

Emojis are a form of descriptive characters that users utilize to help convey feelings and sentiments in their tweets. They can represent basic emotions such as joy, anger, or sadness as well as images such as food, flags, and more. An Emojipedia analysis of 68 million tweets in May of 2020 found that 19.84% of tweets on the platform contained at least one emoji. Their analysis observed an over 4% increase in the frequency of Emoji use from their previous analysis in August of 2018, showing an increased use of emojis by the general Twitter user over that period [10].

Utilizing the 50 most frequently used emojis, I considered

which emojis saw their relative frequency change the most between the two levels of identification. One of the most striking observations was the underlying connotations of the most polarizing emojis. Emojis that appeared more frequently in identifiable tweets were largely positive, with emojis representing a high-five, a thumbs up, and a large smile being the most starkly frequent relative to anonymous users. In contrast, anonymous users used emojis that represented a larger emotion spectrum more frequently, with emojis representing fear, naughtiness, and crying being the most frequently used relative to identifiable users.

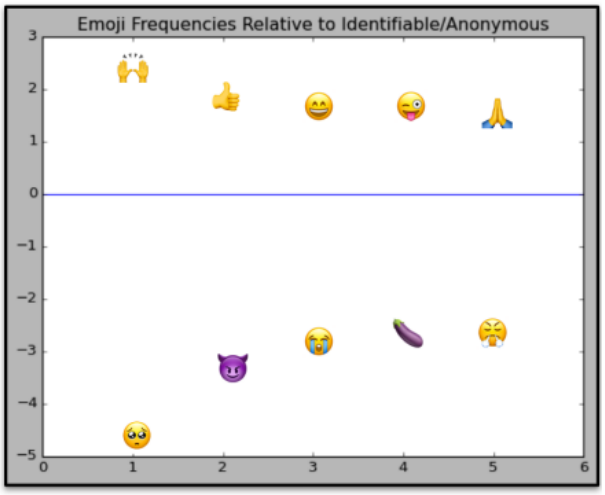


Figure 5
Emoji Relative Frequencies

Sentiment Analysis

Human language is very complex and circumstantial, but we hoped to shed light on some of the quantifiable distinctions in communication patterns between the identification levels of users. Sentiment analysis is a natural language processing technique that attempts to systematically categorize the prevailing sentiment of sections of text as either positive or negative. In our case, it would be able to provide a binary classification for whether the tweet had more of a positive or negative net polarity in its sentiment.

Sentiment Classification

Python has many libraries that allow us to explore the sentiment scores of bodies of text. We explored three different models from different libraries and considered their relative strengths. For each model, we prepared the data in the manner that best worked with the model's parameters. The three models we used were NLTK's sentiment analyzer, TextBlob's sentiment framework, and CoreNLP's sentiment trainer.

Sentiment Model

In an attempt to obtain better predictive results in our sentiment analysis, I implemented a max-voting ensemble model between our three classifications. In this case we ran all three sentiment analysis models on each tweet, and as it was a binary classification of positive or negative, we were then able to select the mode classification for the tweet based upon the three classifications. Each of the models had been trained on different data, and my hope was to reduce any potential bias from a single model in the sentiment classification process. The ensemble model is depicted in Figure 6. Each of the sentiment models were weighted equally and our final prediction was simply whichever binary classification received the majority.

Sentiment Results

We observed a distinct positive trend in the average sentiment scores of tweets as users became more identifiable. If the sentiment distribution of tweets were perfectly balanced, we would expect an average polarity of 0, however all four levels ended up with a positive average polarity score. Identifiable users had an average polarity score of 0.27, demonstrating a definite trend towards more positive sentiment in their communication, while anonymous users posted a lesser positive average of 0.13. A graphical representation of this trend can be seen in Figure 7. This classification lines up with basic frequency reports, as anonymous users were 29% more likely to use the words with the roots 'hate' or 'anger', while identifiable users were 35% more likely to use terms with roots of 'love' or 'like'.

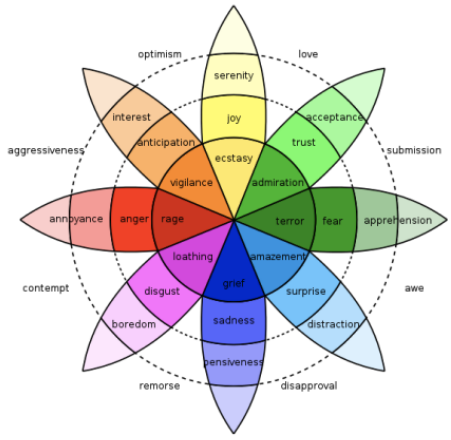


Figure 8
Plutchik's Wheel of Emotion

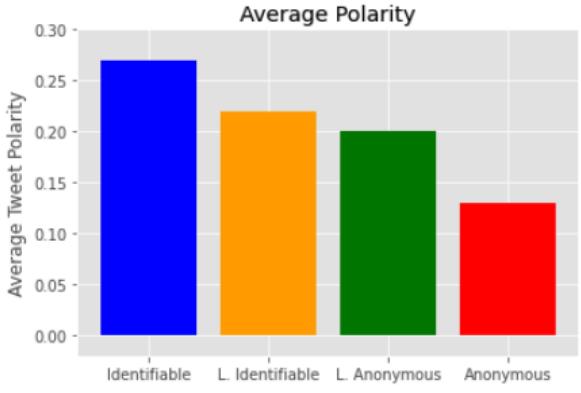


Figure 7
Tweet Sentiment Polarities

Emotion Classification

Emotion classification in text can be viewed as an evolution on the more established field of sentiment analysis. While

sentiment analysis attempts to determine the polarity of a sentence or tweet, whether it be "positive" or "negative", emotion classification is more fine-grained and less standardized. The most common form of emotion classification is a multiclass approach with a set of discrete emotions. The set of discrete emotions chosen for a specific classifier is generally some subset of renowned psychologist Robert Plutchik's wheel of emotion (namely: joy, trust, fear, surprise, sadness, disgust, anger, and anticipation), with often an ambiguous "neutral" class included as well. In Plutchik's understanding, emotions have cross-cultural counterparts that have a level of robustness to a model and allow for a more multi-leveled understanding of a text.

Figure 8 shows Plutchik's famous wheel of emotions. Plutchik's believed that emotions can be expressed at different intensities and can mix with one another to form different emotions, demonstrating the complexities of modeling human language in discrete terms [11].

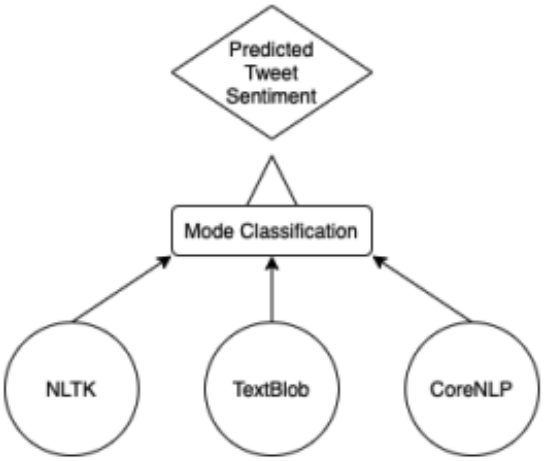


Figure 6
Sentiment Ensemble Model

Training Dataset

The dataset I eventually decided to train my model off of was a highperformance training set developed by Elastic labs that employed a 6-level classification set of anger, fear, joy, love, sadness, and surprise. The dataset was hand-labelled with two levels of authentication that agreed on prior different classifications and was built over the classification of over a million tweets.

Data Preprocessing

Before we can run any of our more advanced models on the data, like with sentiment analysis, we will want to convert the text into a format more conducive to analysis. Natural language is complex, but we can remove some of the ambiguity by simplifying the input language.

Our first step was to tokenize the data, which is where we split our tweet into smaller units called tokens. We defined tokens as single words as this seemed to be the most useful semantic unit for processing. This step also throws away certain characters, such as punctuation and quotation marks, as they are not as interesting to our analysis. Next we stemmed and lemmatized the data, which is the process of reducing inflectional forms to a more common base form. For instance, words such as tweet, tweeting, tweeted, and tweets can all be reduced to the base form of tweet so as to simplify analysis and group words by their common form. We removed especially common terms from the dataset, such as 'the', 'of', and others as their frequency in all text does not make them insightful or interesting for analysis.

Next we converted our text into a numerical format that is better suited to computation. To accomplish this, we created an index-word mapping, where we assigned each of our unique stemmed words to a unique index value sequentially. This allows our data to appear as a sequence of numbers that can be mapped back to real language. Finally, we vectorized our data by converting them into tensors, a data structure that is well suited for linear algebra and heavy computation. We were finally able to feed this vectorized format of our data to the model for classification after it had been trained on our training dataset.

Model Architecture

After trial and error, I eventually opted to build my model around a gated recurrent units (GRUs) paradigm which is a recurrent neural network that utilizes a gating mechanism. The model is very similar to the more common Long Short-Term Memory (LSTM) model in that it includes a forget gate and helps overcome the vanishing gradient problem that often plagues recurrent neural networks [12]. The model performed faster and still maintained accuracy because of the succinct nature of tweets when compared to the LSTM.

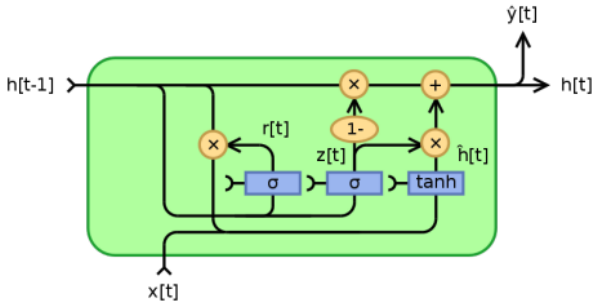


Figure 9
Gated Recurrent Units Model

Model Classification Performance

The entire training dataset was passed through the GRU model 10 full times to help provide better classification accuracy. When compared to the validation dataset, our model was able to correctly classify the expected emotion 92% of the time, and in particular excelled at classifying sadness and joy (97% and 95% respectively), which is noteworthy as they turned out to be two most commonly observed emotions by a significant

margin. The greatest difficulty for the model was in distinguishing between the emotions love and joy and then fear and surprise, which is understandable when comparing the pairs of emotions proximity to one another on Plutchik's wheel of emotion. The confusion matrix can be viewed in Figure 10.

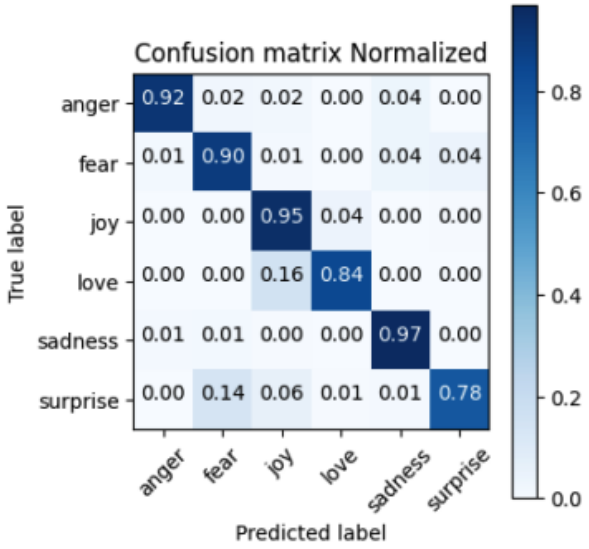


Figure 10
Emotion Model Confusion Matrix

Emotion Classification Results

There is a distinct change in the emotion profiles between identifiable and anonymous users. We observe that identifiable users exhibit predominately joy, while the levels of anger, fear, and surprise both saw much lower frequencies. This aligns with our sentiment analysis that observed a more positive average polarity and our frequency analysis that saw users more likely to use words using joyful connotations like 'like' and 'love'. It may be interpreted that users felt some reservation in sharing their true emotions, as they did when sharing opinions without qualifying terms, and thus the less balanced emotion profile is in a sense dulled version of the true. A graphical representation of identifiable user's emotion profile can be seen in Figure 11.

Anonymous users observed a much broader emotion profile, and in particular saw stark rises in the frequencies of anger, fear, and surprise as compared to their identifiable counterparts. A graphical representation of anonymous users' emotion profile can be seen in Figure 12.

This trend follows our more simplistic emoji frequency analysis where the fearful, sad, naughty, and upset emojis all appeared with greatly increased frequency. This aligns with the results of our frequency analysis of emotionally charged words and our sentiment analysis, showing a distinct prevalence of robust emotions. In contrast to the identified users, this broader emotion spectrum by anonymous users may be closer to the true emotion profile they experience as they are not inhibited by real-world consequences for the way they express themselves.

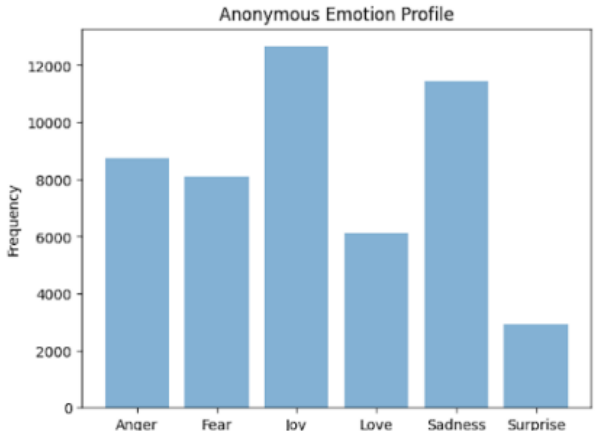


Figure 11
Anonymous Emotion Profile

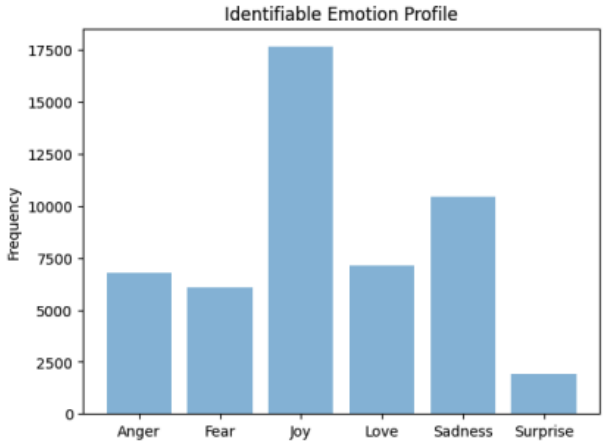


Figure 12
Identifiable Emotion Profile

Identifiable and anonymous users exhibited distinct emotion profiles when conditioned on containing controversial topics such as politics, religion, and sports. While the exact discrepancies are fairly subjective, and only minimal understanding can be obtained without a better controlled experiment, it is worth noting the overarching trend that the levels of joy classifications decrease dramatically between identifiable and anonymous users, while anger, sadness, and love all saw dramatic increases in every single controversial category observed. This is interesting to consider in context of free speech, as all classified accounts had a listed location in the United States. Despite being theoretically free to voice opinions, as long as they do not endanger public safety, users still preferred to express a wider range of emotions when engaging in controversial subjects as an anonymous user.

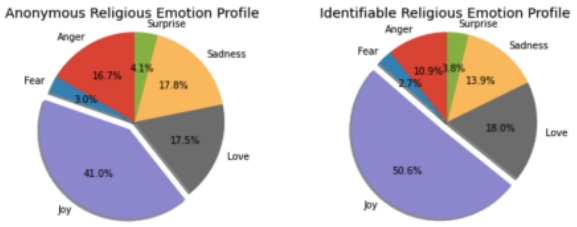


Figure 13
Religious Emotion Profile

Conclusion

Users online exhibit distinct communicative trends depending on their level of identifiability. As the world continues to grow closer together online, with over 80% of Americans predicted to have an active social media platform by 2022, it is important to understand the effects that anonymity has on human behavior.

The broader emotion profile of anonymous users may suggest less inhibition when voicing their opinions online. Users seemed to feel greater safety in expressing fear, sadness, and surprise while they communicated online, as well as to engage in more controversial matters. Anonymity leads to thoughts being expressed more concisely and in more absolute terms.

The knowledge of accountability for their words seems to cause users to qualify their opinions and to further explain and expound their thoughts before submitting them online. They also exhibit more joy than anonymous users, though it may be argued that this is a mask that is concealing their true emotion profile.

As society grows, and younger generations continue to prefer online communication to that of in-person, governments and industries will need to keep in mind the benefits and pitfalls of allowing their users to masquerade anonymously. It may provide insight into users' true feelings, but it may just as well allow users to show only their sharpest edges without accountability.

Acknowledgements

I would like to sincerely thank Dr. Nancy Fulda, my faculty advisor; Dr. Christopher Archibald, my faculty reader; and Dr. Seth Holladay, my faculty coordinator for their kindness, patience, and support throughout this whole process. I would not have been able to have completed this project without them. In addition, the work of everyone who helped me classify my data was invaluable and I never would have finished this project without their assistance and sacrifice. In particular, my aunt Sandra, who classified over 50,000 accounts, cannot be thanked and credited enough.

I am also deeply grateful for Vika Filimoeatu, Julie Radle, and the entire BYU Honors Program for their guidance and assistance throughout my entire undergraduate education. This program has helped me grow more than I ever would have imagined when I enrolled three years ago.