Anonymous Username Classification with Username Strings and an SVM

Dawn Chandler

Simon Fraser University

dschandl@sfu.ca

301249853

LING 807

2019-12-06

APA Style

3604 words

Table of Contents

Introduction

For decades now, the Internet has been a place where people can take a break from their real world lives. If they want to, they can escape into another world and assume another identity: from online virtual worlds like Second Life, to social media and forums like Twitter or Stack Overflow. Anonymity has allowed for a host of social activities to emerge online: trolling, online dating, taboo discussions, illegal activity, and more. Username choice and anonymity shapes the discussions on countless platforms online, most of which require users to make an account and an identity in order to make posts or use their services. In most cases, this identity does not have to reflect the user's real world personhood. Consequently, the usernames that people use online are typically more than just unique identifiers: they are a way to express oneself with as much or as little anonymity as desired.

Since anonymity allows for unique opportunities online, there is much research that examines the relationship between anonymity and online behaviour. Datasets produced from online platforms tend to be very large, and usernames are one of the most salient ways that individuals express their identity online, so being able to automatically classify usernames by anonymity would be useful to this type of research. In this paper, we describe our contribution to the problem of anonymous username classification. We trained an SVM model to classify usernames as Anonymous or Identifiable using only features from the username. Our results are promising, especially considering that our work was a test of the feasibility of doing anonymous username classification given username strings only.

Related Work

Anonymous Account Classifiers

Peddiniti, Ross, Cappos (2017) is the only other known work on automated anonymous account classification. This is despite the fact that information classification is a broad and prolific field of machine learning. Peddinti et al.'s (2017) goal was to investigate the feasibility of automating labelling Anonymous and Identifiable Twitter account classification, rather than username classification. They trained two classifiers: one for Anonymous and Non-anonymous accounts, and one for Identifiable and Non-identifiable accounts.

Their goal for building these account classifiers was to automate identifying Anonymous and Identifiable accounts, so that they could metadata about those types of accounts as features for identifying *sensitive* (LGBT, mental illness, pornography, etc.) Twitter accounts, as follower anonymity patterns are indicative of sensitive accounts.

Peddinti et al. (2017) tested several methods for identifying accounts as Anonymous or Identifiable. Their first approach was name list membership. Twitter accounts have First Name and Last Name fields. They used checked these two fields against lists of first names and last names from the US Census and US Social Security Administration. Accounts that had valid First Name and Last Name fields were labelled as Identifiable. However, this method had issues with common English names that are also common words, such as Crystal or May. There were also issues with parsing First Names or Last Names that contained more than just a name, for example First Name: DoctorAlice. They tried applying structural constraints, but these were ineffective. Their conclusion was that "simply checking for name occurrences from first and last name lists did not give more than 58% precision for anonymous and 70% precision for identifiable accounts" (Peddinti et al., 2017). They were seeking high precision for their purposes.

Their successful approach was to build a machine learning classifier using the Weka toolkit. The researchers chose a random forest model with 100 trees and 10-fold cross validation. Initially, they used all the Twitter account features in their training. Then they pruned their model to use only the 16 features that provided the greatest information gain. These salient features were: “name popularity rankings, word occurrences in Scrabble word lists (word lists without proper nouns and names), and Twitter account profile properties” (Peddinti et al., 2017). Twitter account profile properties include number of followers, includes a URL in the profile, and enabled protected privacy feature. None of these features were extracted from the username.

Peddinti et al. (2017) achieved very usable results with the machine learning classification method. For their purposes, sought high precision. They were able to achieve high precision at the cost of low recall by tuning their cost parameters. They were able to do this using a cost scheme to penalize misclassification. "Unlike spam or abuse detection techniques, our sensitive account detector does not necessitate identifying all of the anonymous or identifiable accounts – What is most important is identifying a significant fraction with low error rates. Hence, high recall values are not absolutely necessary [...]" Peddinti et al. (2017). They were able to achieve 0.90 and 0.93 precision for Anonymous and Identifiable classification, respectively.

Anonymity Research

There are a number of studies which have examined the role of anonymity online (Peddinti, Korolova, Bursztein, & Sampemane, 2014; Bernstein et al., 2011; Postmes, Spears, Sakhel, & de Groot, 2001; Correa, Silva, Mondal, Benevenuto, & Gummadi, 2015). Even in work where anonymity was not initially considered, an anonymous username classifier could

easily allow anonymity to be studied as a variable. Gautam & Taboada (2019) investigated classifying online opinion article comments by toxicity and constructiveness. One point for further research that they identified was exploring the relationship between the anonymity of a user and the toxicity of a comment. Being able to provide researchers with a reliable method of automatic anonymous username classification would allow them to perform more analyses on a broader range of datasets, as not all online platforms have user accounts that are marked as anonymous in the metadata. It would also significantly reduce the need for manual annotation and allow larger datasets to be studied.

Objective

We wanted to classify usernames using only the username, unlike Peddinti et al. (2017), who used many features from a user's Twitter account to build their Anonymous and Identifiable account classifiers. Being able to classify an account based only on the username is akin to what humans are able to do naturally. This approach would also be more extensible to other platforms, as each platform tends to have different account features, but usernames across platforms tend to have the same limitations in length and type of characters. Finally, if an anonymous account classifier were used to find correlations with other features in a dataset, the classifier would have to be independent of features from the dataset. For example, training an anonymous account classifier using features from the posts of a user and then examining the relationship between anonymity and the posts of a user could produce a kind of circular reasoning. For these reasons, we chose to use only at the username itself in building our machine learning classifier, and not any other features that could not be extracted from the user account. We were unsure whether

this username focused approach would yield results, so we viewed this project as a feasibility study.

Approach

Dataset

The dataset we used was the SFU Opinions and Comments Corpus (SOCC), also used in Gautam & Taboada (2019). This dataset contains 5 years' (2012-2016) worth of comments and their metadata scraped from The Globe and Mail website, Canada's main English daily news outlet. The SOCC contains 10,339 opinion articles (editorials, columns, and op-eds) and 663,173 comments. We did not calculate the number of usernames, but we estimate it at around 300,000. From this dataset, we used the raw comments data, which included the comment author (i.e. username), the comment, and many metadata features, such as the comment ID, whether the comment author was a moderator, and whether the comment was edited. Since our objective was to classify usernames (as opposed to accounts) as anonymous or not, we ignored all fields except for comment author.

For our humble research project, we randomly selected 1,500 unique usernames from the dataset. We were limited in terms of human and computing power; otherwise, we would have used more of the data. We manually annotated the usernames as either Identifiable or Anonymous. Any username that was not Identifiable was marked as Anonymous. We did not include an Unclassifiable category, nor did we remove usernames which were difficult to classify.

See Figure 0 for examples of Identifiable and Anonymous usernames from the dataset. See Appendix A for our classification manual. The vast majority (88%) of usernames were Anonymous by our definition.

Anonymous	**Identifiable**
TopGearFan	G. Barry Stewart
Vicki1990	AdamGrant42
Al from Kitimat	Elliotgrace
Jennifer_M	Jehuda Ben-Israel
gadzooks66	Karen Johnson

Figure 0: Examples of username classification.

Implementation

Seeing that Peddinti et al.'s (2017) statistical machine learning classifier was more effective than simple rule based methods, we decided to follow suit and build our own classifier based on a statistical machine learning model. We used the Python machine learning library scikit-learn to train a state vector machine (SVM) model. See Appendix B for the code and documentation.

SVMs require predefined features in order to train on data and predict labels. We extracted several features from the usernames that we thought might provide information gain in our task. They are listed in Figure 1. We were unable to quantitatively determine which of these features provided more information gain than others.

In order to determine values for # of Tokens, # of Scrabble Words, Contains First Name, and Contains Last Name, we had to segment each username. Segmentation was done through rules, delimited by spaces, then underscores, then dashes, then periods, then the CamelCase

convention. Numbers were not segmented into their own tokens, but this should be done in future work. For # of Scrabble Words, we checked each token's membership in the SOWPODS Scrabble word list (SOWPODS, n.d.). For Contains First Name and Contains Last Name, we checked each token's membership against first name and last name wordlists compiled by Novelle (2013).

We included the name features because a username cannot be identifiable without containing a first name and last name. For # of characters, # of Tokens, # of Scrabble Words, and Contains Digits, we suspected, based on our manual inspection of the data, that Anonymous and Identifiable usernames would have different values for each. Anonymous usernames tend to contain more words than Identifiable usernames, tend to contain digits more often than Identifiable usernames, and can be shorter or longer than Identifiable usernames.

Feature	**Value Type**	**Example Value**
(Comment Author)	(String)	(Alberta_Rob)
# of Characters	Integer	11
# of Tokens	Integer	2 # ['alberta', 'rob']
# of Scrabble Words	Integer	1 # 'rob'
Contains First Name	Boolean	True # 'rob'
Contains Last Name	Boolean	False
Contains Digits	Boolean	False
(Is Anonymous)	(Boolean)	(True)

Figure 1: Features extracted from usernames. All example values correspond to the username Alberta_Rob. Comment Author was not used as a training feature, but is provided in order to give a clearer picture of our data. Is Anonymous was the label for each data point.

SVMs are one of the most accurate statistical machine learning methods, such as decision trees and logistic regression. (Navlani, 2018) SVMs work by projecting the data points into a higher dimensional space and drawing a hyperplane with the maximal amount of margin between the data points closest to the hyperplane (called support vectors) of each class, in this case Anonymous and Identifiable. We used an 80/20 train/test split to train scikit-learn's LinearSVC classifier (one of the library's implementations of an SVM) on the features we extracted from each username.

Results

The performance of our model on the test data (300 usernames) is summarized in Figure 2.

Label	Precision	Recall	F1-score
Anonymous	0.945	0.977	0.961
Identifiable	0.786	0.595	0.677

Figure 2: Performance measures of our SVM model on Anonymous and Identifiable usernames.

Discussion

Evaluation

These performance measures show that classification using our model, which represents an initial effort, is more accurate than flipping a coin, at the very least. For classifying Identifiable usernames (Figure 3), the precision and overall performance of our model is not as high as the random forest results achieved by Peddinti et al. (2017). However, our model achieved higher precision than their simple rule based classification method, in spite of the fact that they optimized it for precision. For classifying Anonymous usernames (Figure 4), our model performed much better than either of Peddinti et al.'s (2017) methods. This is probably due to an

imbalance in our data, since only 12% of our 1,200 training examples were Anonymous usernames.

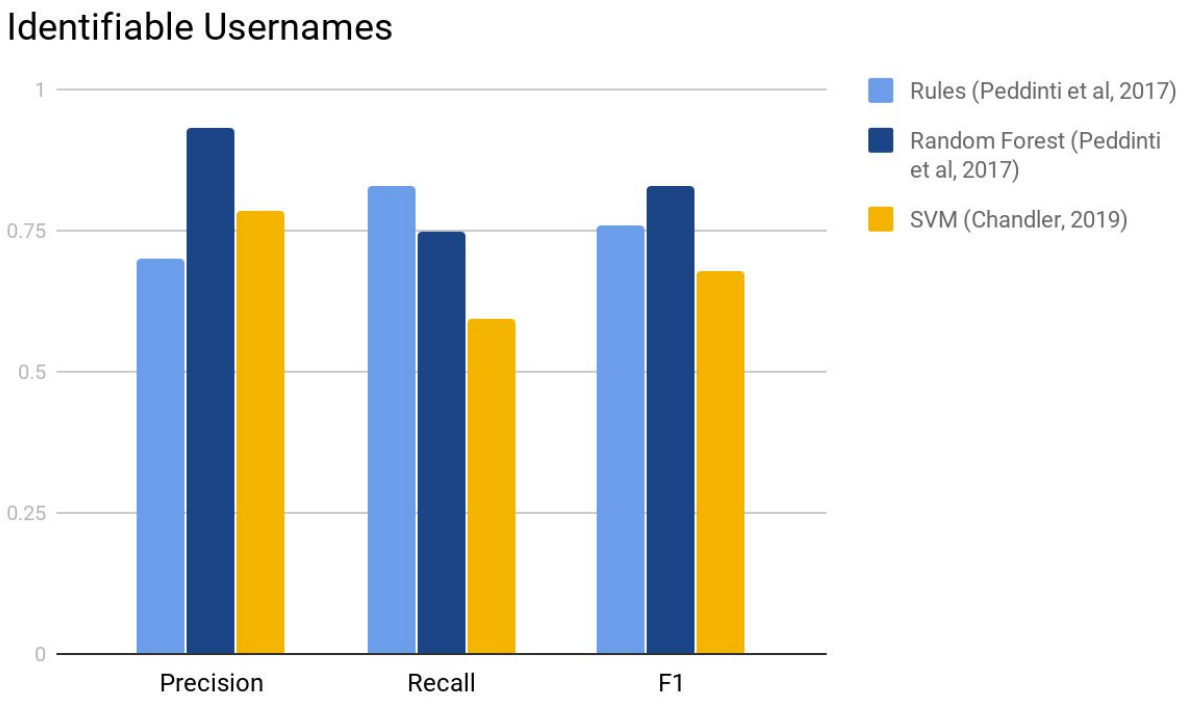

Figure 3: Comparison of precision, recall, and F1-score for the methods from Peddinti et al. (2017) (light and dark blue) versus our SVM model (orange/yellow) on classifying Identifiable usernames.

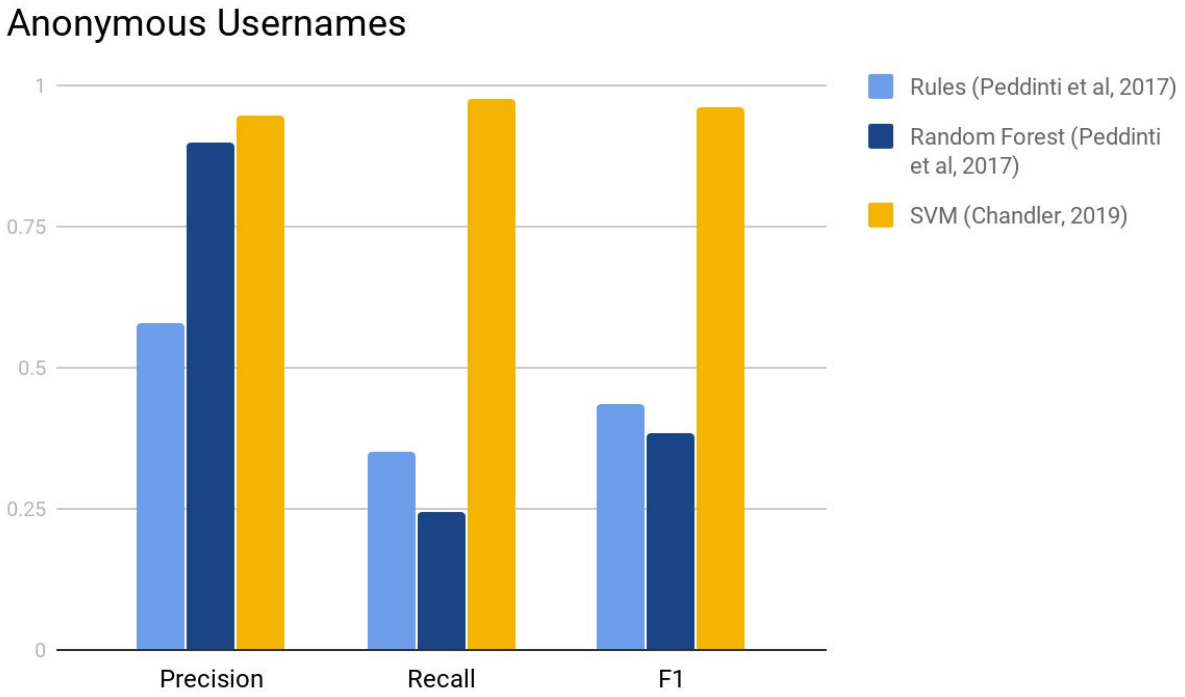

Figure 4: Comparison of precision, recall, and F1-score for the methods from Peddinti et al. (2017) (light and dark blue) versus our SVM model (orange/yellow) on classifying Anonymous usernames.

It should be noted that Peddinti et al. (2017) trained their model on significantly more data (100,000 accounts), pruned their features based on information gain, and implemented a cost scheme. Given our limited resources, we were unable to undertake similar efforts. Fundamentally, our goals were different. Peddinti et al. (2017) sought to maximize the precision at the expense of recall, which we did not. They also sought to classify Identifiable and Anonymous accounts, as opposed to usernames. Consequently, many of their data features came from Twitter accounts. By contrast, all our features were extracted from usernames only. Peddinti et al. (2017) also did not perform any segmentation to determine their features, as Twitter accounts contain separate First Name and Last Name fields, which they used for name

list and Scrabble word list membership checking. Finally, the SOCC data is different from Peddinti et al.'s (2017) Twitter data in several respects. Since Twitter and The Globe and Mail comment section represent different platforms, users choose different types of usernames. Moreover, 72% of the Twitter accounts in their dataset were Identifiable and only 15% were Anonymous. (The rest were unclassifiable.) This data imbalance is the inverse of ours: we had very few Identifiable usernames in our dataset.

In context, our results are not much worse than Peddinti et al.'s (2017). Consequently, we have reason to believe that our approach shows promise.

Analysis

False Negatives.

The SOCC dataset contained many features which proved challenging for our classifier. Most of the false negatives (i.e. Identifiable usernames misclassified as Anonymous) seem to have contained an uncommon first or last name which was not present in the first name or last name datasets. One other issue is that, currently, the segmenter does not have a way of segmenting usernames not delimited by spaces, capitals, or other methods. Nor does the function that determines list membership does not check substrings of tokens, so it is not surprising that tedwilson2 was a false negative.

Username
Stiv Ramone
tedwilson2
Gaius Cassius
Guido Sartucci

MarcGoulet

Figure 5: Examples of false negative errors. False negatives are usernames whose true label is Identifiable and whose predicted label is Anonymous.

False Positives.

The false positives (i.e. Anonymous names misclassified as Identifiable) are more difficult to explain. They all contain either one first name or last name, such as david travels or Enrique. But there are many other usernames in the dataset that share these features that were not misclassified, such as colin7 and BCDave1. It could be the case that the number of characters and number of tokens played a role in the classifier's decision, as only one false negative was one token long.

Username
Enrique
I'm Howard
Sojourner Soo
Sandman in Sechelt
Richard _S

Figure 6: Examples of false positive errors. False positives are usernames whose true label is Anonymous and whose predicted label is Identifiable.

Annotation Errors.

There may have been annotation errors made during the manual classification process, especially since there were several Identifiable usernames that contained non-English names. For example, we had to deliberate whether before we decided that that Amguada Kickboote was probably not a real name. The name lists include many non-English names, but they are still Euro-centric as a consequence of the datasets from which they were compiled.

In order not to confuse the classifier, we had to annotate Louis Riel, Michael Moore, Peter Pan, Sweeney Todd, Winston Churchill, and others imitative usernames as Identifiable. While it is possible that a living person named Louis Riel commented on an article on The Globe and Mail website, it seems more likely that some user, seeking anonymity, commented using the name of a deceased figure from Canadian history. Labelling these imitative usernames as Anonymous would have resulted in more false positives. Unfortunately, this is a limitation of statistical machine learning classifiers.

Unbalanced Data.

As mentioned, our data is very unbalanced, with only 12% of usernames being Identifiable. Our F1-score for classifying Anonymous usernames was 0.961, significantly higher than our F1-score for classifying Identifiable usernames: 0.677. The most likely explanation for this difference is the imbalance of data. The model did not see enough examples of Identifiable data (only 144 usernames) during training, or it has learned that by guessing that an unclear username is Anonymous, it will most likely be correct. Correctly classifying Identifiable usernames clearly poses a challenge.

Further Research

While our approach shows promise, there are many areas where it could use improvement. We suggest the following areas to optimize for future researchers who want to build off of our work in anonymous username classification.

Data.

We recommend annotating many more usernames than 1,500, as our limited means did not allow us to label even 1% of the SOCC data. In particular, our SVM classifier would likely

have performed better had it been given more than 144 Identifiable examples to train on. Removing usernames from the training set that are difficult to classify would also likely result in performance gains.

Being able to resolve the imitative usernames classification issue without increasing the false positive rate for Identifiable usernames would be technically challenging. As imitative usernames represent only a very small portion of usernames, we do not recommend making exploring this issue a priority.

Training a classifier on annotated data which has multiple classes of anonymity (ex. Highly Anonymous, Anonymous, Identifiable, Highly Identifiable) could provide more insightful data for future research into the relationship between anonymity and online behaviour.

Features.

The word list could have been better tuned to this dataset. Many users, such as ARareTorontoConservative and Sandman in Sechelt, include their location in their username. Adding Canadian cities and provinces to the Scrabble word list would have increased the accuracy of finding the number of words in a username. However, this would mean redefining the # of Scrabble Words feature.

The segmentation of usernames could be improved. Digits should become their own tokens instead of doing AdamGrant42 => ['Adam', 'Grant42'] and then removing digits when checking against word lists. This can be fixed with a regular expressions rule. Currently, the segmentation is rule based and requires knowledge of delimiters. Being able to segment usernames in a more sophisticated way, such a n-grams, could improve the accuracy of segmentation.

Peddinti et al. (2017) used name popularity ranking as a feature. We use Contains First Name and Contains Last Name as Boolean features, but popularity rankings of names could be more useful. However, this would likely reduce the variety of non-English names that we currently have using Norvell's (2013) dataset. The biggest issue with names may be that our definitions of anonymity and identifiability are based on the presence of a first and last name. However, this requirement is not reflected in the extracted features nor model. One simple solution would be to replace Contains First Name and Contains Last Name with a new Boolean feature Contains First and Last Name.

Since much of Natural Language Processing research uses word or character level embeddings of text, it could be promising to apply those to anonymous username classification. A character level embedding could be taken of the whole username, and pre-trained word embeddings could be used to represent relevant tokens in the username.

Model.

Finally, experimenting with different statistical machine learning classifiers, such as naive Bayes or random forest, would also be worth exploring. Unsupervised classification, such as k-neighbours, might also be worth investigating, as usernames follow certain patterns which differ from platform to platform. The results of unsupervised learning might reveal different types of anonymity. Neural machine learning models are also worth exploring, particularly if character or word embeddings are integrated into the features of the data. Since there is very limited research into anonymous username classification, it is unknown which type of model would be most suited to this task.

In general, there were many cost schemes and tuning of hyperparameters that we could have done had we had more time. Typically, machine learning research does this kind of experimentation, but our work was more so a proof of concept on the feasibility of classifying anonymous usernames using only the username itself.

Conclusion

From the SOCC dataset, we extracted features from 1,500 usernames and manually annotated them as Anonymous or Identified. Using an SVM classifier, we were able to demonstrate promising results in labelling these usernames based only on features extracted from the username strings. While our model generally did not perform better than Peddinti et al.'s (2017), our datasets and objectives were different: they optimized precision over recall whereas we did not, and we classified usernames based on username strings only. Seeing as our approach was a proof of concept, there is much room for improvement. With sufficient resources, we have reason to believe that a new model based on our baseline could achieve competitive or state-of-the-art results in anonymous username classification. Making such a tool available to other researchers would allow for new insights into the relationship between anonymity and behaviour online.

References

Bernstein, M. S., Monroy-Hernández, A., Harry, D., André, P., Panovich, K., & Vargas, G. G. (2011). 4chan and/b: An analysis of anonymity and ephemerality in a large online community. *In Proceedings of the 5th International AAAI Conference on Weblogs and Social Media (ICWSM)*. Retrieved from https://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/viewFile/2873/4398

Correa, D., Silva, L., Mondal, M., Benevenuto, F., & Gummadi, K. (2015). The Many Shades of Anonymity: Characterizing Anonymous Social Media Content. *In International AAAI Conference on Web and Social Media.* Retrieved from https://www.aaai.org/ocs/index.php/ICWSM/ICWSM15/paper/view/10596

Gautam, V., & Taboada, M. (2019). Constructiveness and toxicity in online news comments. Retrieved from http://www.sfu.ca/~mtaboada/docs/research/Constructive_News_Comments_Report.pdf

Kolhatkar, V., Wu, H., Cavasso, L., Francis, E., Shukla, K., & Taboada, M. (2018, January 18). SFU Opinion and Comments Corpus. Retrieved December 4, 2019 from https://researchdata.sfu.ca/islandora/object/islandora%3A9109

Navlani, A. (2018, January 12). Support Vector Machines in scikit-learn. Retrieved December 5, 2019, from https://www.datacamp.com/community/tutorials/svm-classification-scikit-learn-python

Norvelle, E. (2013). Name Databases. Retrieved December 4, 2019 from https://github.com/smashew/NameDatabases

Pal, S. (2018, November 15). Scikit-learn Tutorial: Machine Learning in Python. Retrieved December 5, 2019, from https://www.dataquest.io/blog/sci-kit-learn-tutorial/

Peddinti, S. T., Korolova, A., Bursztein, E., & Sampemane, G. (2014). Cloak and swagger: understanding data sensitivity through the lens of user anonymity. *In Proceedings of the 35th IEEE Symposium on Security & Privacy* (pp. 493–508). San Jose, CA: IEEE. doi: https://doi.org/10.1109/SP.2014.38

Peddinti, S. T., Ross, K. W., & Cappos, J. (2017). Mining anonymity: identifying sensitive accounts on Twitter. Retrieved from https://arxiv.org/abs/1702.00164

Peddinti, S. T., Ross, K. W., & Cappos, J. (2014). "On the Internet, nobody knows you're a dog": a Twitter case study of anonymity in social networks. *In Proceedings of the Second ACM Conference on Online Social Networks (COSN)* (pp. 83–94). Dublin. doi: https://doi.org/10.1145/2660460.2660467

Postmes, T., Spears, R., Sakhel, K., & de Groot, D. (2001). Social Influence in Computer-Mediated Communication: The Effects of Anonymity on Group Behavior. *Personality and Social Psychology Bulletin, 27*(10), 1243–1254. doi: https://doi.org/10.1177/01461672012710001

SOWPODS Scrabble Word List. (n.d.). Retrieved December 4, 2019 from https://www.wordgamedictionary.com/sowpods/

Appendix A

Classification Manual

These are the guidelines used to classify usernames from the SOCC dataset as Identifiable or Anonymous. This is a binary classification: we did not label degrees of anonymity or identifiability, nor did we label data as unknown or unclassifiable. When we were not sure which username a category was, we labelled it as Anonymous.

Identifiable Usernames

Definition.

Assuming that:

1. the user used their real name when creating their username, and
2. you had unlimited search access to all government databases,

you **WOULD BE ABLE** to trace the user's identity in the real world using only their username. (See Figure i For examples.)

By our estimate, only about 12% of the usernames in the SOCC are Identifiable; the rest are anonymous. The purpose of annotating the usernames is to pick out the Identifiable ones. As such, the default label should be thought of as Anonymous.

Identifiable Username	**Reason**
Karen Johnson	There is a first name and last name.
G. Barry Stewart	There is a given name and last name, plus an initial.
AdamGrant42	There is a first name and last name, plus some digits.
Elliotgrace	There is a first name and what could be a last name.

Jehuda Ben-Israel	There is a non-English first name and last name.

Figure i: Examples of Identifiable usernames from the SOCC dataset, including the reason for the label.

Anonymous Usernames

Definition.

The definition of Anonymous is any username that is not Identifiable, but we have included it in words below regardless.

Assuming that:

1. the user used their real name when creating their username, and
2. you had unlimited search access to all government databases,

you **WOULD NOT BE ABLE** to trace the user's identity in the real world using only their username. (See Figure ii for examples.)

The vast majority of usernames in the SOCC dataset are Anonymous. We used this label for data which we were unsure about.

Anonymous Username	**Reason**
gadzooks66	There are apparently no names in this username.
TopGearFan	There is no name. It is a reference to a TV show.
Vicki1990	There is a first name and presumably year of birth, but that's not enough to uniquely identify a person.
Al from Kitimat	There is a first name and presumably area of residence, but that's not enough to uniquely identify a person.
Jennifer_M	There is a first name and presumably last

	name initial, but that's not enough to uniquely identify a person.

Figure ii: Examples of Anonymous usernames from the SOCC dataset, including the reason for the label.

Imitative Usernames

If a username is likely referring to a known figure and not themselves, it should still be labelled as Identifiable so as not to confuse the classifier. For example, Winston Churchill and Peter Pan should be labelled as Identifiable. The only exception to this are names which are highly unlikely to be used by real people, such as Lady Gaga or Frodo Baggins. See Figure iii for examples.

Username	**Label**	**Reason**
Michael Moore	Identifiable	There is a first name and last name.
Sweeney Todd	Identifiable	There is a first name and last name.
John A. Macdonald	Identifiable	There is a first name and last name, plus an initial.
Frodo Baggins	Anonymous	The first name and last names were invented by the character's creator. They do not reflect the names of any real persons.
Lady Gaga	Anonymous	"Lady" is a title, not a name. The last name does not reflect the name of any real persons.

Figure iii: Examples of Identifiable and Anonymous imitative usernames from the SOCC dataset, including the reason for the label.

Appendix B

Code & Documentation

Please refer to the Anonymous Username Classifier GitHub repo.