Proceedings of the Eighteenth International AAAI Conference on Web and Social Media (ICWSM 2024)

A Semantic Interpreter for Social Media Handles

Azwad Anjum Islam and Mark A. Finlayson

Florida International University
Knight Foundation School of Computing and Information Sciences
11200 SW 8th Street, Miami, FL 33199, USA
{aisla028, markaf}@fiu.edu

Abstract

A handle is a short string of characters that identifies a user or account in a social media platform and is unique within the scope of the platform. Though usually of limited length, a handle can often be the most information-dense string in a social media user profile, potentially containing clues to the user's name, age, location, demographics, or group affiliations. Despite this, the handle has been frequently set aside in work related to inferring user information from their social media profiles. We present a technique for semantic parsing of handles, which seeks to extract relevant information from the handle string. The technique is rule-based and relies on a set of tokenization rules and a variety of external databases (e.g., of names or places) to provide potential interpretations of handles in terms of names, locations, dates, indices, years, ages, positive/negative sentiments, and acronyms. We evaluate an implementation of the technique for English against existing corpora as well as manually evaluate parses of randomly sampled handles, showing that our method achieves good results in both tokenizing the handles (84.9% chance that the correct tokenization is in top three parses while 97% chance that one of the top three parses are at least reasonable) and providing overall "optimistic" interpretation performance of 90.1% accuracy and 0.89 F_1. We also evaluate performance on each of the semantic aspects we interpret (name, location, index, year, age, sentiment, acronym). The technique not only allows us to extract additional information about a user from their handle but also allows us to measure trends in how handles are constructed on specific social media websites. We find that 59% of the handles in our data contain at least part of a person's name, and over 69% of the handles are indicative of the user's gender identity in some way. While our implementation targets English, it can be easily adapted to other languages given the appropriate databases. We release both our code and annotated evaluation data to aid other researchers in validating or extending our work.

1 Introduction

In the context of social media, a handle is a string chosen by the creator of an account to uniquely identify the account within the scope of the platform. Alternatively referred to as the *username*, *user id*, *display name*, *alias*, *screen name*, or *nickname* (Hämäläinen 2022), handles are rarely chosen at random: important clues about how a user wishes to present themselves are revealed in the handle text. For example, two different handles—`AlexBurnsNYT` and `NewYorkBoy`—paint two very different pictures in our mind, although they might plausibly describe the same person. The first suggests a professional male journalist working for the New York Times, while the second suggests a youth, also male, but presenting themselves less seriously, perhaps with pride in being from New York. This type information is often not available from other sources (such as the account name or profile) and can be invaluable to understanding how the associated posts are to be interpreted. While this information can be useful for many purposes—for example, marketing and advertising—in our work we have been particularly interested in how handles are crafted to project a certain identity to improve the effectiveness of influence in disinformation campaigns, where a malicious actor takes on a false in-group identity to appear as a more convincing, trustworthy, or authentic source of information. However, to be able to assess the use of such disguises, we first need to be able to break down a handle into its constituent semantic components. This first task is the focus of the work reported here.

Despite containing valuable clues about user demographics and identity, handles have often been left aside as a source of information in social-media-centric research. While some researchers, in their attempts to infer attributes of social media users, have used handles to supplement more explicit information (such as profile name, profile description, search history, previous posts, etc. (e.g., Pennacchiotti and Popescu 2011; Burger et al. 2011; Volkova, Wilson, and Yarowsky 2013; Liu and Ruths 2013; Nguyen et al. 2013) focused solely on handles is rare (see §2 for the few studies we have identified). While leveraging multiple sources of information allows for inferring more about a user with greater confidence, in many situations, especially due to users' growing concerns with privacy, handles are the only piece of information readily available.

In this work, we present a rule-based parsing technique that can propose interpretations of a handle that assign possible meanings to components of the handle string, including identifying information such as first name, last name, name initial, gender identity, possible country of origin, specific location, age, year, expressed sentiment, and organizational acronyms. The technique relies only on the handle string and requires no input from any additional sources, such as the

account name or profile. We used a rule-based technique—instead of more recently popular fine-tuned or generative large language model (LLM) approaches—for reasons discussed in Section 2. We release both our code and annotated evaluation data[1] to aid other researchers in validating or extending our work.

The paper is structured as follows: we first review other work that has—to some extent—attempted to interpret handles (§2), establishing the context for the proposed technique. Next we describe our parsing strategy in detail, showing first how handles are split into parts and then how each type of semantic information is identified (§3). We then present several evaluations to assess the effectiveness and performance of the proposed technique that use both manual coding and comparison with independent datasets (§4), followed by the results (§5). We then present a general discussion on the work, including its unique usefulness, limitations, and potential use and misuse (§6). Finally, we conclude with a list of our contributions (§7).

2 Related Work

Hämäläinen (2022) reviewed studies of handles and grouped them into three major types, as follows:

1. Qualitative studies typically focus on the analysis of handle semantics, exploring the underlying motivations behind name choices, and examining their relationship to the owner's identity (e.g., Bechar-Israeli 1995; Hogan 2013; Gatson 2011; van der Nagel 2017). The authors and venues of qualitative studies are usually situated in disciplines such as onomastics (the study of history and origin of names), linguistics, and various branches of the humanities. The number of handles examined in these studies usually numbers a few hundred, and often involves time-consuming and laborious (though valuable) qualitative interviewing of actual users.

2. Experimental studies primarily rely on research data derived from empirical experiments (e.g., Back, Schmukle, and Egloff 2008; Heisler and Crabill 2006; Silva and Topolinski 2018). These studies are typically published in behavioral, psychological, and cognitive science venues and focus on examining the communicative aspects of handles, seeking to uncover the assumptions that can be inferred about users' personalities based on their choices, or examining the effectiveness of different types of handles in various contexts of online communication. Again, the number of unique handles examined in these studies are small, from a few 10s to a few 100s.

3. Computational studies use various computational methods to analyze large sets of handles (1000s to millions) to obtain information about users. These studies are usually published in computer science and data science venues. Our work falls in this category, and we will make use of several corpora from this body of work to evaluate our method, as described in Sections 4 and 5.

[1]Code and data can be found at https://doi.org/10.34703/gzx1-9v95/FIY3KZ, and is provided subject to an ethics agreement and under the Creative Commons CC-BY 4.0 License.

There are a large number of studies that focus on inferring demographic information of social media users, but the majority of them either ignore handles completely or make use of them only as a supplement to other features.

Burger et al. (2011) used a supervised machine learning approach to predict the gender of the user for a particular Twitter profile using as features such as the profile full name, profile description, content of the tweets, and profile handle. They used character n-grams to decompose handles into distinct features and achieved a maximum accuracy of 92% with a Balanced Winnow2 classifier. However, they also experimented with using only the handle feature to predict the gender of the user, achieving an accuracy of 77.1%. Both results were significantly higher than the baseline random prediction model (54.9%). In this work, they produced a large gender-labeled Twitter dataset of approximately 184,000 accounts which was developed by collecting additional information about users from their profiles in other blogs or websites of the internet. Unfortunately, this dataset seems to be no longer available.

A particularly interesting study was conducted by Jaech and Ostendorf (2015) which also use handles alone to predict the gender of users in a dating website, as well as the preferred posting language of users in Twitter. The authors used the *Morfessor* algorithm (Creutz and Lagus 2007; Virpioja et al. 2013) for morphological decomposition of the handles instead of character n-grams. The study shows that using a morphological segmentation algorithm slightly outperforms character 3-grams and 4-grams. Also, while the character n-gram models do not benefit from semi-supervised learning, the morphological segmentation-based system does, which is shown by a 10% relative reduction in error rate over the baseline n-gram models.

Knowles, Carroll, and Dredze (2016) developed a tool named *Demographer* which predicts gender solely from the profile name using a combination of name lists and a linear SVM classifier. To train the classifier the authors used gender-labeled name lists from the U.S. Social Security Administration which contained approximately 68k unique names. Since this data only contains American names, one would expect that the classifier would not work well outside that context. For evaluation, gender-labeled names were extracted from Wikidata, as well as names from publicly posting Twitter accounts from the datasets developed by Burger et al. (2011). The study reported an F_1 score of 94.97 for the names in Wikidata and an F_1 score of 90.80 for the gender annotated Twitter dataset. Wood-Doughty et al. (2018) built on this work by making use of character-level neural models to predict the gender and ethnicity of Twitter users. The authors reported a modest increase of 1–2% in accuracy over the SVM implementation of Knowles, Carroll, and Dredze (2016).

More broadly, the field of natural language processing has in the past few years moved toward neural approaches based on fine-tuned large language models (LLMs) or, even more recently, generative LLMs that provide answers via prompting. In this work we opted for a rule-based approach for at least three reasons. First, we only have a very small set of gold annotated data, and did not have the resources to

produce more than this set for development and evaluation. For each category of information we sought to extract from handles, we only have a few 10s to at most one hundred or so examples, so it seemed that a fine-tuned LLM approach was unlikely to be productive. Second, for generative approaches we explored using prompting to produce handle parses (trying a variety of prompt forms, often including example parses), and found that while state-of-art systems like ChatGPT (OpenAI 2021) were able to handle easily tokenizable handles such as `AlexBurnsNYT`, it struggled or did not produce correct or reasonable output for many subtle handles—for example, handles that included non-English names such as `jotahlozano`—which our system can easily handle. It also produced different output in different runs, making the output unpredictable. It seems plausible that more effort (or more data) might produce better performing systems of this type; in that case, our rule-based approach provides a clear, explainable baseline against which such future possible techniques can be compared. Third, we saw no principled way of producing explainable confidence scores for interpretations, which is an advantage of our approach (although, to be clear, we do not deeply evaluate our confidence scores here because that is highly dependent on the specific datasets used and overall system architecture in which the handle parser is integrated). A comparison in performance between our system and ChatGPT is provided in (§5).

3 Parsing Strategy

Our approach to handle parsing can be divided into two broad steps: first, break the handle into its constituent parts (tokenization), and second, interpret the parts generated in the first step (interpretation). We will call the constituent parts of a handle the *tokens*. Finding the correct token boundaries can be challenging, as handles don't necessarily follow any particular structure or patterns, and often contain intentional misspelling of words (e.g., `amaaaaanda`), substitution of letters by other characters such as digits (e.g., `B3EL1VE_20`), and other anomalies. Therefore, each handle is tokenized in multiple ways, resulting in multiple possible *tokenizations*. All these tokenizations are then processed in the interpretation step. During interpretation, each token within each tokenization is analyzed in parallel by different modules (e.g., implementing interpretation procedures for names, places, acronyms, etc.) in order to find possible *token interpretations*. Individual token interpretations are also assigned confidence scores. Individual token interpretations and confidence scores can then be combined into overall *tokenization interpretations* (of which there may be several for each tokenization). Finally, tokenization interpretations are ranked based on their overall scores. All implementations and experiments described below were run on commercial off-the-shelf laptop compute hardware as can be found in any recent (4 years old) laptop configuration. The overall workflow of the system is shown in Figure 1.

3.1 Tokenizing Handles

Handles can be constructed in many different ways, and are often made up of non-standard words or tokens. Furthermore, some tokens may have reasonable interpretations corresponding to distinct tokenizations, and it is ambiguous as to the "correct" tokenization. For example, the handle `ravisherman` can be interpreted as `Ravi S Herman`, `Ravi Sherman`, or simply `Ravisher man`, all being reasonable. Therefore, instead of trying to find the objectively correct tokenization, we try to find all possible reasonable ways the handle can be tokenized, resulting in multiple tokenization candidates, and rely on the interpretation step to pick the better tokenization. We use the following strategies (in various combinations, discussed below) to tokenize handles:

- **Camel case, underscores, numbers**: This strategy assumes token boundaries within a handle are marked by capitalized letters (camel case), underscores, and numbers. For example: `AdamSmith94_FR` ⇒ `adam`, `smith`, `94`, `fr`.
- **Underscores and numbers**: Not all handles use camel case, and capital letters may not always denote token boundaries. Therefore this strategy assumes token boundaries are indicated only by underscores and numbers. For example: `BeLiEvE20` ⇒ `believe`, `20`.
- **Underscores**: Sometimes users use digits to represent letters (Example: Using "4" to represent "A"), where the digits do not necessarily denote token boundaries. Therefore we use a strategy that considers only underscores as denoting token boundaries while ignoring numbers. For example: `BEL19VE_RF` ⇒ `bel19ve`, `rf`.
- **Continuous capital letters**: Three or more consecutive capital letters often denote an acronym or abbreviation. These are identified as separate tokens to detect acronyms within handles. Example: `AlexBurnsNYT` ⇒ `alex`, `burns`, `nyt`.
- **Word segmentation**: Many handles do not use capital letters, symbols, or digits to indicate token boundaries. To handle these cases we use word segmentation approaches (Norvig 2009) to detect potential tokens. For example: `alexmorganofficial` ⇒ `alex`, `morgan`, `official`. In our implementation we use the `WordSegment`[2] python module. It provides a pre-trained module for English that uses word unigram and word bigram data to segment words derived from the Google Trillion Word Corpus (Brants and Franz 2006). The word unigram data file includes the most common 333,000 words from the corpus. Similarly, word bigram data includes the most common 250,000 word pairs. Different data can easily be loaded into the module to enable segmentation in different languages.
- **Multi-word expressions**: Sometimes two or more tokens represent one single entity—such as `newyork` or `atlanticcity`—and so those tokens should be interpreted as a unit in interpretation stage. In such cases we seek to find a token which is made up of portions of the original handle, but with spaces inserted (e.g., `new york` or `atlantic city`). To find these cases we

[2]https://github.com/grantjenks/python-wordsegment. Licensed under the Apache License, Version 2.0.

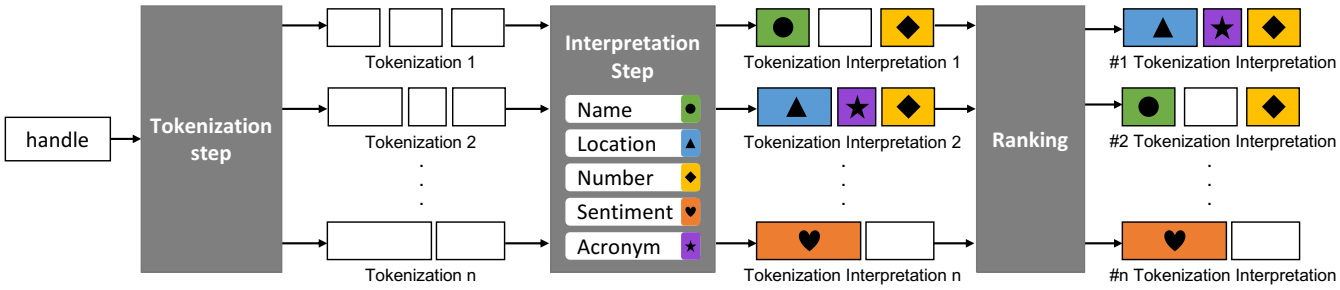

Figure 1: Workflow of the system

use multi-word expression detection (Baldwin and Kim 2010). In our implementation we use the multi-word expression capability provided by `nltk` (Bird, Klein, and Loper 2019), using a list of possible multi-word expressions extracted from the data files that are used in different interpretation modules. This ensures we can capture any multi-word expression that is interpretable by the system.

- **No tokenization**: Sometimes an handle in its entirety represents an entity that can be interpreted in the interpretation stage. For those cases, keeping a tokenization that retains the entire handle as a single token is important.

The tokenization strategies above are not mutually exclusive; some of them may be combined in sequence to produce compound tokenization strategies. There are two combination strategies:

1. **Word segmentation post-processing:** Tokenizations that are generated by strategies based on camel cases, underscores, numbers, and continuous capital letters may still contain tokens that can be broken down further using word segmentation approaches. Therefore, we apply the **word segmentation** strategy on top of these tokenizations to create possible new tokenizations. For example: `adamsmith_fr` $\Rightarrow$ `adamsmith, fr` (tokenization based on underscore) $\Rightarrow$ `adam, smith, fr` (new tokenization based on word segmentation).
2. **Multi-word expression post-processing:** Multi-word expression tokenization is applied to all tokenizations generated by the other strategies to generate possible new tokenizations. For example: `newyorkBoy94` $\Rightarrow$ `newyork, boy, 94` (tokenization based on camel case and number) $\Rightarrow$ `new, york, boy, 94` (new tokenization based on word segmentation) $\Rightarrow$ `new york, boy, 94` (new tokenization based on multi-word expression detection).

Multiple strategies can produce the same tokenization; in which case any duplicates are discarded.

3.2 Token Interpretation

The tokens generated in the tokenization step are then assigned token interpretations in the interpretation step (possibly multiple interpretations per token). The system runs each token through each interpretation module to find possible interpretations. As a consequence a single token can have more than one interpretation, generated by different modules. The system has interpretation modules for names, locations, numbers (years, ages, and indices), sentiment, and acronyms. Each token interpretation is assigned a confidence score from 0 to 1, which may be derived in different ways for each module. If a token cannot be interpreted by any of the modules, it is marked as *uninterpreted* with a score of 0. The overall confidence score of a tokenization interpretation is then computed as the average of the individual token confidence scores. Note that while we did some experimentation with different ways of computing individual and overall confidence scores, we did not investigate this very deeply because it is highly dependent on the quality of the specific databases used and the overall system in which the handle parser is integrated. There probably is much more optimization that can be done for confidence scores, but we set this aside as a less important problem.

3.3 Interpreting Names

The name module is the most significant part of the interpreter, both in terms of complexity and amount of data. Names also give clues to a user's claimed gender, possible country of origin, and race or ethnicity. The name module uses two different approaches to identify names. First, the module searches for each token in a dataset that contains common first and last names from all over the world; this dataset also associates first names with a gender along with a confidence score for the gender label, and both first and last names with common countries of origin and rankings that indicate how popular the names are in different countries. Second, to capture cases of names that are not present in the data, the module uses a character n-gram-based statistical model that classifies if a token is structurally similar to other known names.

Name Dataset The name module uses a dataset curated by Remy (2021) that contains of 730K first names and 983K surnames from 106 countries gathered from user profiles of Facebook, a social media platform[3]. It is important to note that using social media profiles (where names are self-reported by the user) will introduce some noise, as many of these profiles use fake names to conceal users' identities.

[3]This dataset and its associated code uses the MIT License.

For first names the dataset associates the name with the most commonly reported gender and its confidence score, and for both first and last names the most commonly reported countries of origin as well as rankings indicating the popularity of the names in different countries. Remy also provided a python module named `names-dataset` that enables retrieval of all information in the dataset. It is worth noting that while this dataset was generated from a larger Facebook data dump containing additional information that can be considered Personally Identifiable Information (PII), the dataset used in this work was stripped of any such PII.

Some entries that appear in this dataset are also common dictionary words (e.g., *the*, *and*, *grace*, etc.). Some of these entries are artifacts of the way users construct their profile names in social media. To mitigate the effect of these artifacts, we perform the following simple check to filter out tokens that could potentially be interpreted as a word rather than a name. A token is not interpreted as a name if it satisfies all of the following conditions:

- The token is a dictionary word.
- The token is not a popular name. This is determined by the ranking information provided in the Remy data. We consider a token to be a popular name if it is among the 500 most popular names for any country. The cut-off could be modified to fit particular use cases.
- The token appears in the English language more times than any of the 10 most popular American names in the last 100 years appears in the English language. We use the word unigram data from the Google Trillion Word Corpus (Norvig 2009) to determine how many times a token appears in the English language, while we use the Social Security Administration database (Social Security Administration 2023) to determine the most popular names.

This allows us to rule out artifacts such as *and* or *do* as names, while keeping tokens such as *grace* which are not as popular words, or *mark* which are popular words but are also popular names.

Race and Ethnicity Dataset Names can also give strong clues as to race and ethnicity, which for certain national contexts is captured in available datasets. To demonstrate this, we use a dataset curated by Rosenman, Olivella, and Imai (2022) from the Southern United States. This dataset approximates the racial distributions over five racial categories: White, Black, Hispanic, Asian, and Other, that are associated with first, middle, and last names from the voter files of six US Southern states (Alabama, Florida, Georgia, Louisiana, North Carolina, and South Carolina). If a name is not found in the datasets, the module refrains from making any estimation of race or ethnicity. Note that this portion of the interpretation module is only illustrative, as the data we have is quite limited in geographical scope. The system can easily be extended with additional race or ethnicity datasets to improve coverage.

Detecting Missing Names No matter how large a dataset of names is, there will always be names that are missing. To address this problem, our approach makes use of a statistical model to classify if a token is structurally similar to other known names. To illustrate how this would work, we first gathered positive and negative examples. First, we used 35k American names obtained from the Social Security Administration database (Social Security Administration 2023) as positive examples. We used 35k non-name negative examples, half of which were taken from the word unigram data of most frequent English words (Norvig 2009). Dictionary words that were also present in the list of names (such as *rose*, *lily*, etc.) were excluded from negative examples to avoid any overlaps. The other half of the negative examples were randomly generated strings. The length of these strings followed a normal distribution with the same mean and standard deviation as those of the positive examples. We then extracted the trigrams from all the names and computed the probability of that trigram being found in a name and the probability of it being found in a non-name. We used additive smoothing to eliminate zeros in the probability distributions (Lidstone 1920). Then, for any string to be classified as name or non-name, we computed the probability of it being a name by multiplying the name probabilities of all the constituent trigrams. We did the same for the non-name probabilities, and then compared the two overall probabilities scores, marking the string as name or non-name based on which probability was higher. We evaluated this model using random 10-fold cross validation, which produced an accuracy score of 89% (standard deviation of 2.3 percentage point across 10 runs) and an F_1 score of 0.89 (standard deviation of 0.03). This statistical model does not provide any gender or origin information about the name in question, and also does not indicate if a string is a first or last name. Like the case of race and ethnicity, this model was illustrative only, having been trained using only American names and English words, so it likely does not generalize well for names of other cultures and languages.

Grouping Multiple Names Together The name interpretation module interprets names token by token. However, the name of a person usually consists of multiple parts that can be represented over more than one token in a handle. To capture the full name of a person, it is important to be able to group together the individual tokens. To accomplish this, after interpreting individual tokens as first names and surnames, the system uses the following grammar to group them together as a full name. Tokens that consist of single letters can be interpreted as initials depending on which grammar rules match.

```
Name ⇒ FirstName NameInitial LastName |
       FirstName NameInitial NameInitial |
       LastName NameInitial NameInitial
       NameInitial NameInitial LastName |
       FirstName LastName |
       LastName FirstName |
       FirstName NameInitial |
       LastName NameInitial |
       NameInitial FirstName |
       NameInitial LastName|
```

If multiple matches are found, the longest one is taken into consideration. If more than one match of the same length are found, strings are preferred in the order as they are written in the above grammar.

Inherent Challenges in Name Identification As names do not adhere to any strict convention or rules—a person's name can literally be anything—there are some inherent challenges that come hand in hand with name identification. Too liberal an approach will result in a lot of false positive cases, for example: *On* is a common surname in Hong Kong, but if the token *on* is found in a handle of a person who is from another part of the world, the token most likely does not represent a name. On the other hand, too conservative an approach leaves a majority of the less-common names remain unidentified. Since our approach relies solely on the handle as its source of information, identifying a token as a name with 100% confidence is nearly impossible.

3.4 Interpreting Locations

The location interpretation module tries to identify if a token represents a location. To accomplish that, the module makes use of lists of place names drawn from geographical datasets. Locations are resolved in hierarchical levels. The module first compares each token against the regions and subregions of the world, then all the countries of the world along with their alternative name and demonyms, and then all the major cities and states to determine if the token represents a location.

Geographical Datasets: GeoNames The geographical datasets for the location interpretation module are taken from GeoNames (2023) [4]. GeoNames is a free geographical database that covers the whole world and contains more than eleven million place names. In our approach, we use only place names for continents, countries, U.S. states, and cities with a population greater than 15k. Geonames has many additional places in its database, and these can be easily added to the interpreter (for example, when analyzing a French social media dataset, we added the French département).

Challenges in Interpreting Locations There are some significant challenges involved in interpreting locations. Firstly, place names often coincide with common human names, such as Washington, Troy, or Charlotte. It is difficult to determine whether a given token refers to a place or a person without additional information or context. We found in our experiments that in most cases where a token can be interpreted as both a name and a place, the place interpretation is a false positive. Therefore, we discard location interpretations for tokens that are also interpretable as names, unless the location is a continent or a country. This is driven by the intuition that a token which is interpretable as both a name and a location will be more likely to be perceived as a name unless the location is widely known (such as *France* or *Asia*). This approach could easily be modified to discard location interpretations on the basis of populations, or other information external to the handle.

[4]Licensed under Creative Commons Attribution 4.0 License.

A second challenge in location interpretation is that many locations share names. For example, there are 88 cities and towns in the United States with the name *Washington*, and 67 with the name *Springfield*. Thus identifying the precise location in such cases is difficult; we assign by default the place with the largest population, but the approach could easily be modified to incorporate information from outside the handle to narrow the number of possibilities.

3.5 Interpreting Numbers

Numbers are common in handles. They may denote an age (`Alexis21`), a specific year (`Fillon_2017`, `jennifer1994`), an index that separates the user from others with similar handles (`GrahamWolfe1` might mean there are other users with the handle `GrahamWolfe`), or have no easily identifiable meaning at all (`harry77564`, where the number 77564 may only mean something to the user, but have no obvious significance). The number interpretation module is a simple pattern matching module that tries to identify these different types of numerals in a handle using regular expressions. The following mutually exclusive rules are used to interpret numeric tokens:

- **Date**: If there is a numeric token of length 4, 6 or 8 within a handle that matches a valid date of the patterns MMDD, MMYY(YY), MMDDYY(YY) or DDMMYY(YY) it is given an interpretation as a date with confidence (0.99). (e.g., `james08251994`)
- **Year**: If a handle contains a numeric token of length 4 that represents a number from 1700 to 2099, it is given an interpretation as a year with high confidence (specifically, 0.99). If the token represents a number from 1300 to 1699, it is also interpreted as a year, but with lower confidence (0.80). If a numeric token of length 2 within a handle represents a number from 80 to 99, it is interpreted as a year with slightly higher confidence (0.90). (e.g., `fillon2017`)
- **Age**: According to Iqbal (2023), around 80% of Twitter users are of 18–44 years of age. Therefore, if a numeric token of length 2 within the handle matches with a number between 18 and 44, it is labeled as an age with high confidence (specifically, 0.99). If a numeric token of length 2 within the handle matches with a number between 13 and 17 or 45 and 69, it is labeled as an age with lower confidence (specifically, 0.80) (`LizaMaria16`)
- **Index**: If a numeric token is of length 1 to 3, and is not labeled under any of the three previous categories, it is labeled as an index with confidence 0.99. `mikestrutter123`)

3.6 Interpreting Sentiment

Positive or negative connotation of words used in a handle can provide clues how a user wishes to be perceived. For example, handles such as `AmazingClara` or `CrazyLorenzo` contains connotative words that can evoke positive or negative emotional response in an audience. In our work, we make use of the VADER (Valence

Aware Dictionary for Sentiment Reasoning) sentiment analysis tool by Hutto and Gilbert (2014)[5] to identify sentiments expressed through meaningful tokens within a handle. This is accomplished in two simple steps:

1. Filter out tokens that have been interpreted as names, locations or acronyms. For example, in `AmazingClara`, the token `Clara` is interpreted as a name and therefore is ignored.
2. Remaining tokens are then analyzed by VADER to determine their polarity. The result generated by VADER contains a *compound score* that is normalized between -1 and $+1$ and is indicative of the degree of the sentiment. We assume that a compound score $>= 0.5$ implies the token has positive sentiment, while a compound score $<= -0.5$ implies the token has negative sentiment. It is worth noting that tokens that are not identified as dictionary words as assessed by the VADER model are given a compound score of 0.0 and considered neutral by default. For tokens with positive or negative sentiment, we report the absolute value of the compound score as the confidence score of the interpretation.

3.7 Interpreting Acronyms

As mentioned above, three or more consecutive capital letters often indicate an acronym or abbreviation. This module merely proposes that any sequence of three or more capital letters is an acronym; it does not attempt to identify the acronym. It would be relatively easy to connect the system to a relevant acronym dictionary to provide more detail or enhanced evaluation of the interpretation.

4 Evaluation Methods

4.1 Evaluation by Manual Inspection

There are no datasets, as far as we are aware, that provide complete semantic interpretations of handles. Therefore, our main evaluation method is a manual inspection of the handle interpretations. We constructed an evaluation set by selecting 500 random user handles from two Twitter datasets: the *US Election 2020 Tweet* (USET) dataset and the *Twitter User Gender Classification* (TUGC) dataset, both from Kaggle (Kaggle 2016, 2020).

The Gold Standard: The first author and two research assistants (the *coders*) carefully analyzed the evaluation set to generate a "gold standard" interpretation, which includes a tokenization for each of the handles as well as interpretations for each token within that tokenization. Initially, the first author and one research assistant individually produced two separate set of interpretations. They agreed 69.6% of the time on what the correct tokenization was, and had an overall agreement of 0.85 κ (Fleiss Kappa) for interpreting tokens into one of the 11 different semantic categories. The other research assistant acted as an adjudicator (tie-breaker) for instances where disagreements occurred, thus producing a single gold standard evaluation set. We observed that disagreements often stemmed from differences in the coders' level of familiarity to certain pop culture references, foreign naming conventions etc.. To help with this, coders were encouraged to use the internet to look for meanings of potential tokens during the coding process.

[5]Licensed under the MIT License.

Category Type	Count	%
Name (FirstName, LastName)	457	37.7%
Numeric (Date, Index, Year, Age)	101	8.3%
Name Initials	54	4.5%
Pos/Neg Sentiment	41	3.4%
Location	23	1.9%
Acronym	22	1.8%
Uninterpreted	515	42.5%
Total	1,213	100.0%

Table 1: Summary of appearance of different types of interpretation categories in gold standard data.

Table 1 shows the count and percentage of tokens from each category in the gold standard interpretations, with *Name* being by far the most common category. 42.5% of the tokens remained uninterpreted by the human annotators, which shows the difficulty of the problem.

Two coders then compared the gold standard interpretation to the top three ranked tokenization interpretations produced by the system for each of these handles. For each tokenization interpretation, the coders evaluated the tokenization step itself as well as each interpretation category (names, locations, numbers, sentiment, and acronyms) according to the rules laid out below. As discussed before, there may not be one single objectively correct tokenization interpretation for many handles. For this reason, we also explore the cases when a tokenization interpretation does not match with the gold standard, but is reasonable. We note that "reasonability" is a subjective metric, but because the system's goal is to interpret how a handle is perceived by an audience rather than the absolute meaning intended by the author of the handle, we depend on a human annotator's best judgement as to what can be considered as an acceptable interpretation of the handle.

For the tokenization step:

- If a tokenization produced by the system matches with the gold standard, it is marked as ***correct*** tokenization.
- If a tokenization produced by the system does not match with the gold standard, but is deemed reasonable by the coder, it is marked as a ***reasonable*** tokenization.
- If a tokenization is neither correct nor reasonable, it is marked as an ***incorrect*** tokenization.

For each token interpretation:

- If a token interpretation within a correct tokenization matches with the gold standard interpretation for that token, it is marked as a ***correct*** token interpretation.
- If a token interpretation within a correct tokenization does not match with the gold standard interpretation for that token, but is deemed reasonable by the coder, it is marked as a ***reasonable*** token interpretation.

- If a token interpretation within a *reasonable* or *incorrect* tokenization is deemed reasonable by the coder, it is marked as a ***reasonable*** token interpretation.
- If a token interpretation is neither correct nor reasonable, it is marked as an ***incorrect*** token interpretation.

The coders had an overall agreement of 0.78 κ (Cohen's Kappa) for evaluating tokenizations, and 0.64 κ (Cohen's Kappa) for evaluating token interpretations as *reasonable* or *incorrect*. Again, the third coder acted as a tie-breaker to settle disagreements. Note that according to our rules, the *correctness* of a tokenization or token interpretation is an objective decision.

Given these evaluations, we computed a variety of evaluation scores, as described below. For the categories X below, the possible values are {`FirstName, NameInitial, LastName, Location, Date, Index, Year, Age, PosSentiment, NegSentiment, and Acronym`}.[6]

- **Tokenization Accuracy (Strict)** fraction of times the correct tokenization is produced. This can be computed for the top ranked tokenization (*Top-1*), the top two tokenizations (*Top-2*), or all three of the top tokenizations (*Top-3*).
- **Tokenization Accuracy (Optimistic)** fraction of times the correct or a reasonable tokenization is produced, again computed for Top-1, Top-2, and Top-3.
- **Overall Interpretation Accuracy (Strict)** is defined as the fraction of tokens in the *correct* tokenizations (if it appears in the Top-3) that have *correct* token interpretations.
- **Overall Interpretation Accuracy (Optimistic)** is defined as the fraction of tokens in the *correct* tokenization (if it appears in the Top-3) that have *correct* or *reasonable* token interpretations.
- **For Category = X (Strict)** is defined as an F_1 measure over all correct tokenizations (if it appears in the Top-3), where:
  - **TP**: the number of true positives is the number of times a token is marked as both X and *correct*.
  - **FP**: the number of false positives is the number of times a token is marked as a category X and *reasonable* or *incorrect*.
  - **FN**: the number of false negatives is the number of times a token is marked as a category X in the gold standard, but not in the interpretation.
- **For Category = X (Optimistic)** is defined as an F_1 measure over all *correct* tokenizations (if it appears in the Top-3), where:
  - **TP**: the number of true positives is the number of times a token is marked as X and either *correct* or *reasonable*.
  - **FP**: the number of false positives is the number of times a token is marked as a category X and *incorrect*.
  - **FN**: the number of false negatives is the number of times a token is marked as a category X in the gold standard, but not in the interpretation.
- **Overall Interpretation F_1 (Strict)** is defined by the micro-average of the Interpretation F_1 (Strict) scores of all interpretation categories.
- **Overall Interpretation F_1 (Optimistic)** is defined by the micro-average of the Interpretation F_1 (Optimistic) scores of all interpretation categories.
- **Overall Interpretation Accuracy in Reasonable Tokenizations** is defined by the fraction of tokens in *reasonable* tokenizations that appear in the Top-3 that have *reasonable* token interpretation.
- **Interpretation Accuracy for Category = X in Reasonable Tokenizations** is defined by the fraction of tokens in *reasonable* tokenizations that appear in the Top-3 that are interpreted as X and are marked *reasonable*.

[6]Note that the Name interpretation module assigns the categories `FirstName`, `NameInitial`, `LastName`, but also `Name`, which indicates a token that statistically looks like a name but cannot be identified as specifically a first or last name. When matching categories, `Name` is considered equal to `FirstName` or `LastName`.

4.2 Evaluation of Gender Interpretation via TUGC

Datasets of handles labeled with demographic attributes are quite rare. Gender-labeled Twitter datasets previously developed by Burger et al. (2011), Liu and Ruths (2013), Volkova, Wilson, and Yarowsky (2013) are no longer publicly available. To evaluate gender identification capabilities of our system, we used the Twitter User Gender Classification (TUGC) dataset (Kaggle 2016) which contains approximately 20k Twitter profiles that include handles that are labeled with a gender. Possible values for gender labels are *male*, *female*, *brand*, and *unknown*. The gender label also includes a confidence score between 0 and 1. In our experiment, we first removed all user handles with gender labels *brand* or *unknown*, this resulted in 12,894 samples. Then, we discarded all user handles that had a confidence score of less than 0.95, resulting in 10,023 gender-labeled handles. This gives us a reasonably reliable dataset with minimal loss in data volume. We applied our method to these handles, inferring the gender using the techniques (based on first name identification) described above. If more than one token was interpreted as first names, the name with the highest gender confidence score was selected. If the system failed to detect a name within a handle or if it failed to infer a gender from the identified names, it was considered an automatic failure for gender identification.

It is worth noting that the TUGC dataset also contains the location and timezone information of the user, which potentially could have been used to evaluate the location interpretation capabilities of our system. However, the location information in this dataset contains many missing values, in addition to being highly irregular in terms of format and granularity. Therefore, we were unable to easily use the location information in this dataset in any meaningful way.

Metric	Tokenization Accuracy Top-1	Top-2	Top-3
Tokenization Accuracy (Strict)	53.1	79.3	85.0
Tokenization Accuracy (Optimistic)	59.2	90.9	97.1

Table 2: Overall system tokenization accuracy scores.

4.3 Evaluation of Name Interpretation via TUA

For evaluating how well our system can detect a user's name from their handle, we used the *Twitter Username Alias* (TUA) Dataset provided by McKelvey et al. (2017). The dataset includes 113k Twitter handles correctly aligned with their corresponding profile names. All the handles in this dataset were taken from public Twitter pages and aligned with the names entered on their profiles. We ran these handles through our system, and compared tokens assigned a name interpretation against the profile name provided in the dataset. If token string was a sub-string (ignoring case) of the profile name, this was considered a match.

For both gender and name interpretation, we considered name tokens from the Top-3 tokenizations.

5 Results

5.1 Tokenization Capability

The system shows strong tokenization capability, shown in Table 2. When limited to only the top tokenization *(Top-1)*, accuracy is a modest 53.1%. When considering the top two *Top-2* and three *Top-3* tokenizations, the strict accuracy measure improves quite a bit to 79.3% and 85.0% respectively, while the optimistic measure produces the highest accuracy score of 97.1% *Top-3*. It is interesting to note that while the improvement in accuracy from *Top-1* to *Top-2* is a large 26.2%, percentage points, accuracy only improves 5.7% percentage points from *Top-2* to *Top-3*. For the optimistic measure, these numbers are 31.7% and 6.2% respectively, indicating that most correct tokenizations are captured within the first two tokenizations. When doing error analysis, we found that when there are multiple ways to tokenize a handle in a reasonable fashion, sometimes the "correct" tokenization does not float to the top. This is a failure of the confidence score computation, which, as we have explained elsewhere, we did not spend very much time optimizing.

5.2 Interpretation Capability

Table 3 depicts the overall performance of the system. The system produces accuracy scores of 67.97% and 90.13%, and F_1 scores of 0.74 and 0.89 when using the strict and optimistic measures respectively. These are very promising numbers considering the difficulty of the task. Note that these results are based on a total of 1008 tokens from *correct* tokenizations that appear in the Top-3, as described in Section 4.

To put the tokenization and interpretation results in context, we asked ChatGPT (OpenAI 2021)—an LLM (GPT-3.5) based AI chatbot—to perform the same task as our system on our gold standard dataset of 500 user handles. The prompts used are given in Appendix A, and the results are shown in Table 4. While ChatGPT performed the tokenization task almost as well as our rule based system, our system largely outperformed it in interpretation. ChatGPT also generated different results on different runs (all being similar in performance), which is another reason why a rule based system might be preferred to produced reproducible results in such a tasks. We note that our analysis of ChatGPT (and LLM's more generally) here is purely illustrative. In future work it might be quite useful to evaluate this deeply and systematically, but this is beyond the scope of the work reported here.

Metric	Strict	Optimistic
Overall Accuracy	68.0%	90.1%
Overall F_1 Score	0.74	0.89

Table 3: Summary of overall interpretation performance in correct tokenizations

System	Strict Tokenization	Strict Overall Intepretation
Our System	53.1%	65.3%
ChatGPT	51.1%	33.4%

Table 4: Comparison with ChatGPT in *Top-1* tokenization and overall interpretation (strict).

Tables 5 and 6 depict the individual interpretation F_1 scores (strict and optimistic) in correct tokenizations and the interpretation accuracy scores in reasonable tokenizations, respectively. In our error analysis, we observed that the *Location* semantic category in particular lags behind than the other categories due to a large number of false negatives when tokens were interpreted as names instead of locations.

5.3 Gender Identification

Running our system on the TUGC dataset produces an accuracy of 69.6% in gender identification: the system failed to infer any gender in 5.9% of the handles, and mislabeled the gender in the other 24.5% cases. In our error analysis, we observed that many of these mislabelings resulted from a handle having two or more tokens interpretable as a first name. As noted above, in these cases we used the gender of the highest confidence interpretation, which is a fairly noisy estimation, resulting in many errors.

In the same task, Jaech and Ostendorf achieved an accuracy score of up to 72.2% using a semi-supervised learning algorithm while and Burger et al. achieved an accuracy of 77% using a supervised linear classifier framework (*Winnow*). However, the comparability of these results is suspect due to the vast difference in both amount and quality of data that were used.

5.4 Name Identification

When run on the TUA dataset, our system could correctly identify the first names and surnames of 58.3% of the user

Metric	Strict	Optimistic
FirstName	0.74	0.89
NameInitial	0.74	0.94
LastName	0.75	0.90
Location	0.42	0.70
Date	1.00	1.00
Index	0.84	0.98
Year	0.73	0.76
Age	0.76	0.98
PosSentiment	0.68	0.84
NegSentiment	0.55	0.71
Acronym	0.59	0.70

Table 5: Summary of interpretation F_1 score by category in correct tokenizations. Note that only 3 samples in the evaluation set contained the "Date" category, all of which were interpreted correctly, which is why *Date* is 1.0.

Metric	Accuracy (%)
FirstName	88.7
NameInitial	97.2
LastName	89.0
Location	69.2
Index	100.0
Year	100.0
Age	100.0
PosSentiment	67.7
NegSentiment	80.0
Acronym	100.0
Overall Accuracy	91.3

Table 6: Summary of interpretation accuracy in reasonable tokenizations. Note that the category "Date" does not appear in reasonable tokenizations.

profiles: the system could not find any names in 3.1% of the handles, and found names that do not match with the profile names in the other 38.6% cases). A large number of profiles in this dataset have handles that do not have any resemblance with the names reported by users, which affected the performance of the tool. McKelvey et al. (2017) reported a 80.1% accuracy in predicting a correct alignment between profile name and user handle, using a variety of complex procedures. We also performed this task using the output of our system, counting as a positive match when a token with a name interpretation appears as a substring of the profile name, and achieved 77.7% accuracy.

5.5 Error Analysis for Other Categories

The final two categories with non-ideal performance were Positive and Negative sentiment. We observed that errors in these categories resulted from standard problems with single-word sentiment analysis, such as lack of broader context (which is not available for the handles), conflict with name tokens, and polysemous meanings for individual words such that they have both positive and negative connotations depending on their use.

6 Discussion

Information derived from user handles can be useful for many purposes, including marketing and advertisement, social media analytics, content personalization, moderation, etc. In our own work we are particularly interested in how handles can be crafted to project a certain identity to affect influence or disinformation campaigns. As an illustrative example, consider how the tweet "*The sanctions on Russia are going to prove ineffective*" would be received depending on whether it was sent from the handle `iLoveMoscow` versus `AlexBurnsNYT`. The difference in influence of the tweet in these two cases can only be attributed to the semantic difference of the two user handles. The construction of a user handle can dramatically affect how the information presented is perceived among an audience.

Our system shows promising results in both tokenizing handles and extracting demographic clues from tokens. However, as a rule based system, it understandably struggles when said clues are intentionally obscured in the handle by the user or are simply not present. As a result, the system is not likely to perform well in platforms where users actively seek anonymity.

Our work relies solely on the handles themselves and external data sources, and is limited by the quality of available data, but also by the inherent difficulty of the task. Even with high quality datasets, making sense of a user handle will always be a challenging task due to the very nature of how handles are constructed.

6.1 Ethical Considerations

A system for extracting information from handles raises privacy concerns. First, there is some debate as to whether a handle by itself is personally identifiable information (PII). According to U.S. General Services Administration (2019), *"The definition of PII is not anchored to any single category of information or technology. Rather, it requires a case-by-case assessment of the specific risk that an individual can be identified."* Some researchers have considered handles to be PII when it is accompanied with other information (such as a password, e.g., Ren et al. 2016; Huang et al. 2019). Subahi and Theodorakopoulos (2012) classified handles as *"non-sensitive PII"*. We note that many handles, in isolation, cannot identify an individual with certainty.

Our system allows the extraction of information of a potentially sensitive nature such as possible names, possible gender, possible race or ethnicity, possible ages, and relevant locations. While this information cannot be extracted from the handle unless it was included there by the user themselves, it does present the potential for abuse, as it enables extraction of information that was hinted at but not explicitly provided by the user. Furthermore, it is definitely the case that any demographic information extracted from handles, used in conjunction with information drawn from other sources, can facilitate racial profiling, gender discrimination, or other unethical activities.

Given the above considerations, such a system should be used with caution and appropriate measures taken to protect user privacy. Such measure might include voluntarily restricting allowed uses of the technology or discarding any sensitive information (e.g., specific names) after processing. In the release of our own code and data, we will vet requests for the supplementary material and also require researchers to sign an ethics statement.

Despite the above risks, we do believe this work can be put to positive use. We are particularly interested, for example, in the assumption of particular identities that are used to malicious effect in online disinformation campaigns. To effectively combat this kind of false identity presentation we need to be able to parse those identities. Another possible positive use of the system is to provide a middle ground between including or not including handles for those generating datasets of social media data. Our system allows the extraction of useful, general demographic clues (supposed country of origin, inferred gender, age group, etc.) that can be included in a dataset along with a hash of a handle, thus preserving demographic hints while eliminating potential PII. This would allow researchers to leverage demographic information in the handle without exposing to information that can potentially identify individuals.

7 Contributions

Our contribution in this paper are two fold, First, we presented a complete semantic interpreter for social media user handles which, to the best of our knowledge, is the first of its kind. We showed that such a system can be used to produce relevant results that are useful in many use cases with reasonable confidence. We showed that our rule based parsing technique can outperform more modern, generative large language models. Secondly, we have developed a dataset of social media user handles that have been manually annotated with their tokenization and interpretation with near perfect inter-annotator agreement. This dataset is also a first of its kind and can be a valuable resource for future research work. Finally, we release our code and evaluation data via our institutional repository, using standard code and data formats, to allow validation, reproduction, and extension of the work[7].

Acknowledgements

We gratefully acknowledge the valuable annotation work carried out by Ana Oliveira and Tisa Islam Erana. This work was supported by DARPA via INCAS Program Prime contract HR001121C0186. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the DARPA.

References

Back, M. D.; Schmukle, S. C.; and Egloff, B. 2008. How extraverted is honey.bunny77@hotmail.de? Inferring personality from e-mail addresses. *Journal of Research in Personality*, 42(4): 1116–1122. https://doi.org/10.1016/j.jrp.2008.02.001.

Baldwin, T.; and Kim, S. N. 2010. Multiword expressions. *Handbook of Natural Language Processing*, 2: 267–292.

Bechar-Israeli, H. 1995. From <Bonehead> to <cLoNehEad>: Nicknames, Play, and Identity on Internet Relay Chat1. *Journal of Computer-Mediated Communication*, 1(2): JCMC127. https://doi.org/10.1111/j.1083-6101.1995.tb00325.x.

Bird, S.; Klein, E.; and Loper, E. 2019. *Natural Language Processing with Python, Version 3.0*. Sebastopol, CA: O'Reilly Media, Inc.

Brants, T.; and Franz, A. 2006. LDC Corpus LDC2006T13: Web 1T 5-gram Version 1. https://catalog.ldc.upenn.edu/LDC2006T13. Accessed: 2024-04-01.

Burger, J. D.; Henderson, J.; Kim, G.; and Zarrella, G. 2011. Discriminating Gender on Twitter. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, 1301–1309. Edinburgh, Scotland, UK. https://aclanthology.org/D11-1120.

Creutz, M.; and Lagus, K. 2007. Unsupervised Models for Morpheme Segmentation and Morphology Learning. *ACM Transactions on Speech and Language Processing*, 4: 3:1–3:34. https://doi.org/10.1145/1187415.1187418.

Gatson, S. 2011. Self-Naming Practices on the Internet: Identity, Authenticity, and Community. *Cultural Studies â Critical Methodologies*, 11: 224–235. https://doi.org/10.1177/1532708611409531.

GeoNames. 2023. GeoNames Geographical Database. http://geonames.org/. Accessed on July 10, 2023.

Heisler, J. M.; and Crabill, S. L. 2006. Who are "stinkybug" and "Packerfan4"? Email Pseudonyms and Participants' Perceptions of Demography, Productivity, and Personality. *Journal of Computer-Mediated Communication*, 12(1): 114–135. https://doi.org/10.1111/j.1083-6101.2006.00317.x.

Hogan, B. 2013. Pseudonyms and the Rise of the Real-Name Web. In *A Companion to New Media Dynamics*, chapter 18, 290–307. West Sussex, UK: John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118321607.ch18.

Huang, J.; Klee, B.; Schuckers, D.; Hou, D.; and Schuckers, S. 2019. Removing Personally Identifiable Information from Shared Dataset for Keystroke Authentication Research. In *Proceedings of the 2019 IEEE Fifth International Conference on Identity, Security, and Behavior Analysis (ISBA)*, 1–7. Hyderabad, India. https://doi.org/10.1109/ISBA.2019.8778628.

Hutto, C.; and Gilbert, E. 2014. VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. In *Proceedings of the Eighth International AAAI Conference on Web and Social Media*, 216–225. Ann Arbor, MI. https://ojs.aaai.org/index.php/ICWSM/article/view/14550.

Hämäläinen, L. 2022. From Bonehead to @realDonaldTrump: A Review of Studies on Online Usernames. *Names: A Journal of Onomastics*, 70(2): 36–53. https://doi.org/10.5195/names.2022.2364.

[7]Code and data may be found at https://doi.org/10.34703/gzx1-9v95/FIY3KZ.

Iqbal, M. 2023. Twitter Revenue and Usage Statistics (2023). https://www.businessofapps.com/data/twitter-statistics/. Accessed on July 10, 2023.

Jaech, A.; and Ostendorf, M. 2015. What Your Username Says About You. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2032–2037. Lisbon, Portugal. https://aclanthology.org/D15-1240.

Kaggle. 2016. Twitter User Gender Classification. https://www.kaggle.com/datasets/crowdflower/twitter-user-gender-classification. Accessed on July 17, 2023.

Kaggle. 2020. US Election 2020 Tweets. https://www.kaggle.com/datasets/manchunhui/us-election-2020-tweets. Accessed on July 17, 2023.

Knowles, R.; Carroll, J.; and Dredze, M. 2016. Demographer: Extremely Simple Name Demographics. In *Proceedings of the First Workshop on NLP and Computational Social Science*, 108–113. Austin, Texas. https://aclanthology.org/W16-5614.

Lidstone, G. J. 1920. Note on the general case of the Bayes-Laplace formula for inductive or a posteriori probabilities. *Transactions of the Faculty of Actuaries*, 8(182-192): 13.

Liu, W.; and Ruths, D. 2013. What's in a name? Using first names as features for gender inference in twitter. In *Proceedings of the AAAI Spring Symposium on Analyzing Microtext*, 10–16. Stanford, CA. https://cdn.aaai.org/ocs/5744/5744-24477-1-PB.pdf.

McKelvey, K.; Goutzounis, P.; da Cruz, S.; and Chambers, N. 2017. Aligning Entity Names with Online Aliases on Twitter. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, 25–35. Valencia, Spain. https://aclanthology.org/W17-1104.

Nguyen, D.; Gravel, R.; Trieschnigg, D.; and Meder, T. 2013. " How old do you think I am?" A study of language and age in Twitter. In *Proceedings of the Seventh International AAAI Conference on Web and Social Media*, 439–448. Cambridge, MA. https://doi.org/10.1609/icwsm.v7i1.14381.

Norvig, P. 2009. Natural Language Corpus Data. In Segaran, T.; and Hammerbacher, J., eds., *Beautiful data: the stories behind elegant data solutions*, chapter 14. Beijing: O'Reilly. ISBN 978-0-596-15711-1.

OpenAI. 2021. ChatGPT: A Large-Scale Generative Language Model. https://chat.openai.com/. Accessed in August, 2023.

Pennacchiotti, M.; and Popescu, A.-M. 2011. Democrats, Republicans and Starbucks Afficionados: User Classification in Twitter. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 430–438. New York, NY. https://doi.org/10.1145/2020408.2020477.

Remy, P. 2021. Names Dataset. https://github.com/philipperemy/name-dataset. Accessed: 2024-04-01.

Ren, J.; Rao, A.; Lindorfer, M.; Legout, A.; and Choffnes, D. 2016. Recon: Revealing and controlling PII leaks in mobile network traffic. In *Proceedings of the Fourteenth Annual International Conference on Mobile Systems, Applications, and Services*, 361–374. Singapore. https://doi.org/10.1145/2906388.2906392.

Rosenman, E.; Olivella, S.; and Imai, K. 2022. Race and ethnicity data for first, middle, and last names, Version 9. Cambridge, MA: Harvard Dataverse, https://doi.org/10.7910/DVN/SGKW0K. Accessed on July 10, 2023.

Silva, R.; and Topolinski, S. 2018. My username is IN!: The influence of inward versus outward wandering usernames on judgments of online seller trustworthiness. *Psychology & Marketing*, 35: 307–319. https://doi.org/10.1002/mar.21088.

Social Security Administration. 2023. Popular Baby Names. https://www.ssa.gov/oact/babynames/limits.html. Accessed on August 23, 2023.

Subahi, A.; and Theodorakopoulos, G. 2012. Automated Approach to Analyze IoT Privacy Policies. In Cagáňová, D.; and Horňáková, N., eds., *Industry 4.0 Challenges in Smart Cities*, 163–186. Cham, Switzerland: Springer International Publishing. https://doi.org/10.1007/978-3-030-92968-8_12.

U.S. General Services Administration. 2019. Rules and Policies Protecting PII - Privacy Act. https://www.gsa.gov/reference/gsa-privacy-program/rules-and-policies-protecting-pii-privacy-act. Accessed on September 9, 2023.

van der Nagel, E. 2017. From usernames to profiles: the development of pseudonymity in Internet communication. *Internet Histories*, 1: 312–331. https://doi.org/10.1080/24701475.2017.1389548.

Virpioja, S.; Smit, P.; Grönroos, S.-A.; and Kurimo, M. 2013. Morfessor 2.0: Python Implementation and Extensions for Morfessor Baseline. Technical report, Aalto University, Aalto, Finland. http://urn.fi/URN:ISBN:978-952-60-5501-5.

Volkova, S.; Wilson, T.; and Yarowsky, D. 2013. Exploring Demographic Language Variations to Improve Multilingual Sentiment Analysis in Social Media. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 1815–1827. Seattle, WA. https://aclanthology.org/D13-1187.

Wood-Doughty, Z.; Andrews, N.; Marvin, R.; and Dredze, M. 2018. Predicting Twitter User Demographics from Names Alone. In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, 105–111. New Orleans, LA. https://aclanthology.org/W18-1114.

Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? There is some debate as to whether handles are personally identifiable information (PII) or not (see discussion in text). Aside from that debate, our system allows the extraction of information of a potentially sensitive nature such as possible names, possible gender, possible race or ethnicity, possible ages, and relevant locations, although this information cannot be extracted from the handle unless it was included there by the user themselves. So there is potential for abuse, as it enables extraction of information that was hinted at but not explicitly provided by the user. Therefore, such a system should be used with caution and appropriate measures taken to protect user privacy. This includes discarding any sensitive information (e.g., specific names) after processing. Despite these risks, we do believe this work is of use, as the assumption of particular identities is used to malicious effect in online disinformation campaigns, and to effectively combat these we need to be able to track those identities. We discuss these issues in Section 6.1.
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes, see Section 4.
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes, see Section 3.
   (e) Did you describe the limitations of your work? Yes, see Section 6.
   (f) Did you discuss any potential negative societal impacts of your work? Yes, see Section 6.1.
   (g) Did you discuss any potential misuse of your work? Yes, see Section 6.1.
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? We carefully document our data and code in the supplementary materials at https://doi.org/10.34703/gzx1-9v95/FIY3KZ. We will also implement access control on the supplementary materials, which will allow us to keep control over who has access to the original code and data. We will only release of code and data after vetting researchers and asking them to sign a ethical use statement, which states that they will not use the work to enable profiling, violating privacy, or releasing PII.
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? N/A
   (b) Have you provided justifications for all theoretical results? N/A
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? N/A
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? N/A
   (e) Did you address potential biases or limitations in your theoretical framework? N/A
   (f) Have you related your theoretical results to the existing literature in social science? Yes, see Section 2.
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? N/A
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? N/A
   (b) Did you include complete proofs of all theoretical results? N/A
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? Yes, all of this information is included in the supplemental material located at https://doi.org/10.34703/gzx1-9v95/FIY3KZ.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? For the name detection statistical model we provide a description of how positive and negative examples were obtained. Training splits were randomly generated and are not provided.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? For the name detection statistical model, we reported error bars. See Section 3.3.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? Yes, see Section 3.
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes, see Sections 4 and 5.
   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? Yes, we discuss the potential errors in extraction of names and locations in Sections 3.3 and 3.4, although we generally *DO NOT* discuss the overall cost of misclassification, as that depends heavily on the end use-case.
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...

(a) If your work uses existing assets, did you cite the creators? Yes, see citations and footnotes throughout.

(b) Did you mention the license of the assets? Yes, licenses of the datasets and python modules are listed in footnotes.

(c) Did you include any new assets in the supplemental material or as a URL? We provide the code for our work as well as a small annotated gold-standard evaluation set.

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? No, because all data that we used was drawn from publicly available datasets.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes, See Section 6.1. While there is debate on whether handles are PII, our system allows the extraction from handles of information which may be considered PII. Thus, care must be taken in the handling of the output of the system. For that reason we will require an ethics agreement before providing the code and data to other researchers.

(f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR? Yes, see discussions in Section 6.1 and 7. The dataset will be released via a publicly accessible permanent institutional repository which assigns a DOI to every artifact. The data and code will be provided in common text-based formats.

(g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset? Yes, this is included in the code/data bundle released at https://doi.org/10.34703/gzx1-9v95/FIY3KZ.

6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...

(a) Did you include the full text of instructions given to participants and screenshots? N/A

(b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? N/A

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? N/A

(d) Did you discuss how data is stored, shared, and deidentified? N/A

A ChatGPT Prompts

We used the following prompt to generate tokenization interpretations from ChatGPT (OpenAI 2021) for our evaluation set.

```
 Separate the following list of user
handles into their component tokens,
labeling each token with its appropriate
interpretation category. All tokens
must make up the whole username and each
token must be assigned an interpretation
category. If there is no meaningful
interpretation for the token, use the
category "U".
For possible interpretation categories, use
the following mapping:
{
"FN": "FirstName",
"IN": "NameInitial",
"SN": "Surname",
"N": "Name",
"L": "Location",
"D": "Date",
"A": "Age",
"Y": "Year",
"I": "Index",
"AC": "Acronym",
"PS": "PosSentiment",
"NS": "NegativeSentiment",
"U": "No Interpretation"
}

For each handle, only output the
parse itself as given in the example:
JotaHLozano=jota_FN,h_IN,lozano_LN. Note
that the user handle has upper and lower
cases characters but the tokens on the
right, only contains lower cases. Maintain
this format.
The list of user handles is:
[*List of handles*]
```