Exploring patterns of identity usage in tweets: a new problem, solution and case study

Kenneth Joseph
Carnegie Mellon University
5000 Forbes Ave.
Pittsburgh, PA
kjoseph@cs.cmu.edu

Wei Wei
Carnegie Mellon University
5000 Forbes Ave.
Pittsburgh, PA
weiwei@cs.cmu.edu

Kathleen M. Carley
Carnegie Mellon University
5000 Forbes Ave.
Pittsburgh, PA
kathleen.carley@cs.cmu.edu

ABSTRACT

Sociologists have long been interested in the ways that *identities*, or labels for people, are created, used and applied across various social contexts. The present work makes two contributions to the study of identity, in particular the study of identity in text. We first consider the following novel NLP task: given a set of text data (here, from Twitter), label each word in the text as being representative of a (possibly multi-word) identity. To address this task, we develop a comprehensive feature set that leverages several avenues of recent NLP work on Twitter and use these features to train a supervised classifier. Our model outperforms a surprisingly strong rule-based baseline by 33%. We then use our model for a case study, applying it to a large corpora of Twitter data from users who actively discussed the Eric Garner and Michael Brown cases. Among other findings, we observe that the identities used by individuals differ in interesting ways based on social context measures derived from census data.

Categories and Subject Descriptors

J.4 [**Social And Behavioral Sciences**]: Sociology

Keywords

Computational Social Science, Twitter, Identity

1. INTRODUCTION

An *identity label*, or simply an *identity*, is a term that conveys a culturally-shared meaning of a person or group of people [35]. Identity labels exist for things like our physical characteristics (e.g. "tall person", "handsome", "man") and the social roles we take on in everyday life (e.g. "lawyer", "doctor"). Identities are thus central to how we communicate social information. For example, when reading the sentence, "Jim is a liar", we know that the identity label "liar" represents the fact that Jim is a person who often states things which are not true. In addition this denotive meaning, there is also an *affective*, or emotional, meaning of the identity "liar" - most English speakers would agree that a "liar" is bad [15].

Social scientists have long been interested in how and when identities are applied, used and created. Given the denotive and affective meaning identities convey, the specific one we choose to describe a person has a significant impact on the way others will act towards her [15]. For example, because liars are "bad", we are unlikely to seek out a friendship with someone that we know has been labeled a liar by others. As social beings, we are implicitly aware of the importance of our identity, and are, consciously or not, consistently managing it in order to appear "worthy" of desirable identities in particular contexts [13].

The study of identity is also prevalent in the natural language processing (NLP) community, though it tends to go by different names. Fine-Grained Named Entity Recognition (FG-NER) [42, 9, 12, 23, 19, 11, 33] is focused in part on the problem of determining from a large set of identities which are most appropriate for specific individuals. Concept-level sentiment mining techniques have been applied to identity labels to understand the affective meaning they carry [2, 21]. Finally, the authors of [5] extract general semantic characteristics of particular "personas", which are similar to identities, from movie reviews.

Surprisingly, however, there does not seem to exist work that considers an even more basic question than those posed above- how do we capture the set of all identities that exist in a particular corpora? While a variety of heuristics have been applied to bootstrap lists of identities for FG-NER (e.g. [42]), to the best of our knowledge, the following prediction problem has not yet been explored:

> *Given a set of text data, label each word in the text as being representative of a (possibly multi-word) identity*

From a sociological perspective, even a method for extracting where identities are used in a given corpora would provide a new way to study their semantic and affective properties. For example, MacKinnon and Heise [16], two prominent social psychologists, argue in their recent book on identity that understanding the structuring of identities into semantic clusters using text analysis can help us to understand how individuals navigate social life and how cultures as a whole create taxonomies of identities (e.g. into those related to occupations versus those related to family).

The first contribution of the present work is a supervised classifier that addresses the NLP task posed above for

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
WOODSTOCK '97 El Paso, Texas USA
Copyright 20XX ACM X-XXXXX-XX-X/XX/XX ...$15.00.

tweets. We first sample 1000 tweets from a large, domain specific corpora and annotate it with labels indicating which terms represent identities. As we are the first to approach this problem, we use an iterative coding scheme and consult with identity scholars to derive theoretically grounded rules for the labeling process. We then construct features derived from both standard lexical structures and from a variety of tools recently produced by the NLP community around Twitter. In particular, features are derived from the output of a Twitter-specific dependency parser [22] as well as from word-vectors trained on a large Twitter corpus using the GloVe algorithm [28]. Additionally, we make use of existing dictionaries of identity labels and also construct a bootstrapped dictionary of identities from unlabeled data via the use of high-precision lexical patterns.

Model performance is first tested on this set of 1000 tweets using cross-validation. As an additional step to assess the quality of our predictions, we also obtain and label an additional 368 tweets made public by other NLP researchers and use them as a validation set. On this validation set, model performance is compared to a rule-based, dictionary-based approach. F1 scores exceed .75 for the full proposed model and outperform this baseline by 33%.

The second contribution of this work is a case study that demonstrates one particular opportunity for sociological research that our method allows. We apply our trained classifier to over 750M tweets sent by 250K Twitter users who were actively engaged in discussion of the recent deaths of Michael Brown and Eric Garner. Our case study explores the following question: *how do the identities in our dataset cluster into semantically organized sets of identities?*.

We provide both quantitative and qualitative analyses of identity sets, or clusters, that result from applying latent Dirichlet allocation (LDA) [6] to a user by identity matrix extracted from our corpus. Encouragingly, the resulting clusters line up well with both prior work and intuition. We find several clusters of identities that match those found by Heise and MacKinnon [16] in their related work and also find clusters that align strongly to contemporary social issues (e.g. the Arab/Israeli conflict). We then briefly explore how these two "types" of identity sets can be differentiated based on their affective meanings. We also consider how differences in social contexts may compel individuals to utilize particular identity sets more often than others.

2. LITERATURE REVIEW

We divide our review of the literature into two parts, one focusing on the sociological literature on identity and the second on related research in the NLP community.

2.1 Sociological Literature

Smith-Lovin [35] defines identities as the ways in which an individual can label another person with whom she has had an interaction. Smith-Lovin continues to define three general types of identities. Role identities indicate positions in a social structure (e.g. occupations). Category identities come from identification with a social category, which are "inclusive [social] structures that require merely that all members share some feature" [8] (e.g. race, gender). Finally, social identities indicate membership in social groups, a collections of individuals who a) perceive they are in the same social category, b) share a common understanding of what this category represents and c) attach an emotional meaning to this category.

The broad definition of identity provided by Smith-Lovin foreshadows various difficulties in developing a methodology to extract them from text. While we provide a more nuanced discussion of practical difficulties in Section 4.1, the chief socio-theoretic issue relates to the difficulty of distinguishing between a "compound" identity, one that has multiple words, and a single identity that is modified. For example, the phrase "black woman" could be viewed as the identity "woman" modified by the adjective "black", or as a single identity, "black woman".

From a linguistic perspective, Recasens et al. [31] suggest that identity should be considered to be both relative and to be varying in granularity, and thus a determination of the granularity of interest should absolve us from these problems. A complementary sociological perspective can be drawn from the theory of intersectionality, which emphasizes the importance of understanding social categories as being social constructions [34] and thus the importance of defining which identities are or are not compound via social consensus. While much remains to be done along these lines theoretically, our labeling scheme attempted to utilize a small number of agreed-upon intersectional identities from the literature. At the same time, where intersectionality was non-obvious, we adopted a coarse-grained labeling approach, looking only for root-level identities and leaving identification of modifiers to future work.

Regardless of the definition of identity used, MacKinnon and Heise [17] perform the only attempt we are aware of to enumerate identity labels on any large scale. Their efforts come in two parts. First, they extract all terms from WordNet [26] that are lexical descendants (recursively) of the term "human being" and then perform a qualitative analysis to understand taxonomic structure in the resulting sets of identities. Their work suggests a set of twelve categories of identities that include, for example, occupation and religion. The authors then perform a semantic analysis of identities in an offline, professionally written dictionary, where they cluster a semantic network extracted from the dictionary to obtain a similar collection of identity sets. These structures are referred to as *institutions* - for example, one institution includes identities such as siblings and parents, representing the institution of family and marriage (pg. 79). Institutions thus consist largely of semantically coherent sets of identities that are applied together in specific social contexts.

The approaches taken by MacKinnon and Heise [17] to define identities on a large scale and to uncover semantically coherent sets of these identities has certain advantages. In particular, both WordNet and the professional dictionary are human curated and widely used, suggesting a high level of precision in both semantic relationships and identity labels used. However, the approach also has disadvantages. First, the datasets used are curated by a specific collection of individuals whose views may not be entirely reminiscent of social consensuses on identities or their meanings. Second, the datasets that they use base semantic relationships largely on denotive meanings of identities. As we will see, affective meanings can be equally important in our understanding of semantically coherent clusters of identities.

2.2 NLP Literature

The task of *Named Entity Recognition* (NER) is define by the goal of extracting entities from text and categorizing

them into a general typology, most often into the categories of People, Locations and Organizations. One of the earliest applications of NER to Twitter data is the work of Ritter et al. [33], who develop and test a semi-supervised model based on Labeled LDA [30]. Ritter et al.'s work moves beyond the simple Person, Organization, Location classification to finer-grained classifications of entities, and is thus one of, if not the, earliest application of FG-NER to Twitter.

Research on FG-NER uses large sets of entity labels and tries to apply them to, for example, people. Thus, as opposed to labeling "Michael Jackson" as a Person entity, an FG-NER system might label him as a "musician". These entity labels used to classify people are by definition identities. It is thus unsurprising that recent work in FG-NER [9, 11] uses WordNet to construct lists of entity types in a very similar fashion to the work of MacKinnon and Heise. Because of this connection between entity labels and identities, features used in FG-NER models are directly applicable to the present work. Of particular interest are the feature sets utilized by Hovy et al. [19] and del Corro et al. [9]. Features across these two works are derived from lexical patterns (e.g. typing Steve Young with the label quarterback given the text "Quarterback Steve Young"), dependency-parses (e.g. "Steve Young is a quarterback"), parts-of-speech and word-vectors, all of which are similarly utilized here.

Yao et al.'s [42] recent work in the FG-NER domain is perhaps most relevant to the work here. The authors develop an approach to type entities with labels from free text. In their work, the "type system", which is loosely equivalent to the set of identity labels we wish to construct, is generated from a pattern-based extraction method from text. After constructing this dictionary of entity types, a matrix factorization method is developed to apply these types to Named Entities.

While our work is thus in many ways related to FG-NER research, it is important to observe that the problem we are interested in has a fundamentally different goal. Whereas in FG-NER, one is attempting to find appropriate labels for Named Entities, here we attempt to find all labels that are used to describe any human or set of humans. This distinction is important for two reasons. First, NER systems assume that entities can be labeled with factual types. However, in highly emotional situations, like the case study considered here, it seems unlikely that such factual labels will be prevalent or even interesting, particularly in Twitter data. Rather, as Bamman has noted [4], what is interesting is how different identities are found in text based on the current social context of the individual writing the text. Second, of particular interest to us is how identity labels themselves are related to each other, not how they apply to entities. While it may be interesting to understand how, for example, different labels are applied to Michael Brown, the current work is focused largely on how people view generic groups of others (e.g. the police) that are rarely regarded as entities.

Because of our focus on semantic coherence of text in Twitter, our case study is similar to a variety of recent efforts to perform topic-modeling on Twitter [41, 43]. Additionally, our focus on affective meaning is relevant to more recent approaches that combine sentiment analysis with semantic connections between terms [18]. The efforts in the present work complement this line of research by considering a particular kind of topics- specifically, "topics"[1] of identities. While the present work uses straightforward methods to perform this clustering, we look forward to leveraging more complex models that account for, e.g., spatio-temporal properties of the data in the near future [1, 10, 40].

3. DATA

A variety of data sources were used in the present work. Here, we give a brief description of each. All code used to collect data, all dictionaries mentioned and all labels for the supervised problem, as well as all code to run the models to reproduce results, will be made available at `http://github.com/kennyjoseph/identity_extraction_pub`.

3.1 Twitter Corpus

On August 9th of 2014, Michael Brown, an unarmed 18-year old African American male, was shot to death by Darren Wilson, a member of the Ferguson, MI police department. Over the next few days, questions began to arise surrounding the circumstances of Brown's death. Over the next several months, two important series of events played out. First, a grand jury was organized to determine whether or not to indite Officer Wilson for any charges related to the death of Michael Brown. Second, a host of mostly peaceful protests were carried out on the streets of Ferguson and elsewhere, demanding justice for yet another young black male that they believed had been wrongly killed at the hands of a police officer.

On November 24th, the grand jury determined there was no probable cause to indite Darren Wilson for any crimes related to the death of Michael Brown. This decision was met harshly by critics both online and on the streets of cities around the United States. Less than two weeks later, another grand jury, this time in Staten Island, also chose not to indite a white police officer over the death of Eric Garner, another black male. Garner's death, which was notably caught on video, reignited flames from the protests in Ferguson, both online and in the streets and from those that both condemned and, unfortunately, those that celebrated the deaths of Garner and Brown.

The tweets used for the present work are a subset of a corpus of approximately two billion tweets from around one million Twitter users who we considered to have been an active participant in these discussions. From August, 2014 through December, 2014, we monitored the Twitter Streaming API with a variety of keywords that were relevant to events in Ferguson following the death of Michael Brown and events in New York City leading up to and following the trial resulting from the death of Eric Garner. For all users that sent more than five tweets in the sample collected from the Streaming API, we then collected their full tweet stream[2]. For the present work, we focus on a subset of users that we expect to be both human and to be active on the site. Specifically, we focus on users who have sent between 50 and 15K total tweets, have less than 25K followers and that have been on the site for 2 or more years. From this set, we consider only English language tweets[3] without URLs[4] that have five or more tokens.

[1] or preferably here, "clusters" or "sets"
[2] Up until their last 3200 tweets, as allowed by the API
[3] determined with the languid library [24]
[4] as determined by Twitter

For the purposes of developing our classifier, we extracted 1000 non-retweets from this set of filtered tweets. We ensure that this sample contains at most one tweet per unique user. Because we expected tweets with identities to be relatively rare, we used three methods to over-sample tweets with identities in them. First, we use Vader [20], a sentiment classifier, to extract only tweets with some form of sentiment. This is because affective relationships between identities tend to have an affective component [15]. Second, we ensure that 10% (100) of the tweets we sampled had one of twenty generic identities labels (e.g. bully, husband) drawn from one of the identity dictionaries described below. Finally, because we were specifically interested in views on the police, we ensure that 15% (150) of the tweets had the word "police" in them.

Due to the large extent to which we utilized sub-sampling, and the fact that our corpus selects on a very distinct dependent variable [38], it was necessary to obtain an outside dataset to validate the model. We chose to use the corpus discussed in [27]. Of the 547 tweets in the corpus, only 368 of them were still able to be extracted from the API (i.e. the rest had been deleted or sent by a user who had since closed their account). We hand labeled each of these 368 tweets and use this set as a validation set.

3.2 Dictionaries

A variety of dictionaries, or word lists, of identities already exist. We leverage several of these dictionaries here. First, Affect Control Theorists, who focus on culturally-shared affective meanings of identities and their behaviors, maintain an open-source listing of identities used in their survey studies [16]. As noted above, WordNet contains an implicit identity dictionary, which can be constructed by collecting all terms that derive from the "Person" term. In addition to these two lists of general identity terms, we also adopt dictionaries for specific types of identities we expect to be prevalent in our dataset. From the GATE [12] set of gazeteers, we utilize a listing of occupations (frequently used as role identities) and nationality-based identities. Finally, we obtain a set of racial slurs for a variety of races from the Racial Slur Database[5]. In total, we thus have five distinct lists of identities.

In addition to dictionaries for identities, we also utilized dictionaries for non-identity words. We drew these from the GATE set of gazeteers, as well as from the set of dictionaries from the Twitter NLP library[6], although several dictionaries from each were excluded based on manual inspection. We also use all terms in WordNet that do not derive from the Person entity as a non-identity dictionary. Finally, we include a generic stopword list.

3.3 Word Vectors

As opposed to allowing the model to learn from one-hot encodings of unigrams, we opt for the dense representations of words afforded by recent work in representation learning [25]. Specifically, we leverage a large set of 50-dimensional word vectors that have been trained on a large set of Twitter data using the GloVe algorithm[7] [28]. As opposed to a unigram-based approach, these dense word vectors allow the model to learn general classes of words that are identities rather than specific unigrams.

[5] `http://www.rsdb.org/`

[6] `https://github.com/aritter/twitter_nlp`

[7] available for download at `http://nlp.stanford.edu/projects/glove/`

4. METHODS

4.1 Labeling Process

For the present work, each term in each tweet was labeled as being either inside or outside of an identity label - an "IO" labeling scheme. Labeling of the data was completed in two steps. In the first, a set of thirteen annotators, most unfamiliar with the project, were each asked to annotate around 150 tweets, giving us two labels for all 1000 tweets. Annotation was performed using the brat rapid annotation system [36]. Guidelines gave annotators an expanded definition of identity as compared to the one presented in this article, as well as a variety of examples. We also asked annotators not to label individuals or pronouns as identities (including modified forms, e.g., "people I know"), not to include organizations themselves as identities (e.g. "Congress announced today" would not have any identities, while "A member of Congress" is an identity), and to only label full words as identities (e.g. "#blacklivesmatter" would *not* be labeled as an identity).

While the guidelines given left us with a very limited definition of what constituted an identity, such a limited definition was necessary for the task to be completed with any amount of agreement. As identity labels were sparse, we only chose to evaluate inter-annotator agreement on tokens where at least one annotator claimed that the span of words was a part of an identity label. On such tokens, agreement was 67%, that is, if one annotator labeled a particular span of tokens as being an identity, there was a 67% chance that the annotation matched exactly with the second annotator.

In reviewing these annotations, we found three main sources of disagreement. First, a few annotators, although told that they were being ignored, still labeled pronouns as identities. Second, annotators varied in the extent to which they included modifier terms in their annotations. This was particularly the case where identities served to modify individual persons (e.g. in the phrase "Mayor de Blasio", "mayor" is a modifier and thus not of interest as an identity in the present work). Finally, identity words that were used as generic, emphatic statements in the text ("man" in "Come on, man!") were differentially labeled by annotators. We eventually chose to ignore these, as they serve largely as general pronouns rather than statements about identity.

After resolving these general themes of disagreements between annotators and fixing annotation errors, we consulted with identity theorists to finalize any additional rules and guidelines for annotations. Once firmly established, we reviewed all annotations to confirm their adherence to these guidelines and also applied them to the additional 368 tweets from outside the corpus.

4.2 Creation of Bootstrapped Dictionary

Existing identity dictionaries did not include many of the identity terms in our labeled data- in fact, even though they contained over 11K entries, they captured only 64% of the identities in our labeled tweets. While, of course, a statistical model allows us to generalize beyond these dictionary terms, another important source of information to leverage was the set of unlabeled tweets in our corpora. A common usage of unlabeled data in FG-NER studies is to extract

"I Am" Rule	"Person" Rule	Not in Dicts
girl	black person	mess
man	wrong person	human
bitch	young person	legend
kid	favorite person	joke
guy	old person	pussy
idiot	nice person	thot
asshole	beautiful person	blessing
woman	amazing person	nightmare
boy	bad person	disgrace
h*e	real person	cutie
friend	innocent person	texter
baby	stupid person	goddess
keeper	homeless person	g
$$$ga	random person	old

Table 1: Three lists of terms from the bootstrapped dictionary, sorted by frequency of occurrence. On the left, top terms from the "I am a", "he is a" etc. ruleset. In the middle, top terms from the "[Identity_Label] person" rule. The final column gives the 15 most frequent phrases extracted from the "I am" ruleset that were not in any existing identity dictionary we used

possible entity types for individuals by *bootstrapping* a dictionary using high-precision lexical patterns [42, 9]. Generally, these bootstrapped dictionaries are created by first performing coarse-grained NER on the data, and then using lexical patterns like "[Entity_Type] such as [Entity]" to extract entity types.

Unfortunately, most of the prior work is focused on news or web data, and many of the patterns that were used in these sources of data, such as appositional phrases ("Joe, the dentist"), almost never occurred in our Twitter corpora. Further, NER on Twitter data is still a very difficult and time-intensive problem [33]. Consequently, we adopted a pair of slightly modified lexical patterns from the existing literature to build our bootstrapped dictionary. First, as opposed to using NER as a precursor for label extraction, we instead use pronouns as the base of our patterns. We thus extract sets of tokens starting with "he is", "she is", "I am" or "you are" that were followed by the word "a" or "an" and consider the first noun that follows to be an identity (e.g. "liar" is extracted from "he is a liar"). Second, we found that in almost all cases, terms proceeding the words "person" or "people" (and variants of these words) were identities (e.g. "annoying person"). We thus extract all terms of the form "[X] people" or "[X] person" and add these to our bootstrapped dictionary as well.

We used our unlabeled corpus to extract all phrases matching these two sets of patterns, and kept a count of the number of times we capture each unique phrase using each pattern. After obtaining these counts from our full, 2B tweet dataset, we remove all phrases that occurred fewer than 10 times within one of these patterns. In total, we were left with 30.5K unique identity terms. Table 1 displays three columns that help to describe the resulting dictionary. The first column shows the fifteen most frequently captured terms from the "I am", "he is", etc. ruleset[8]. The second column shows the fifteen most frequent terms collected using the "person" ruleset. The final column shows the fifteen most frequent terms from the "I am" rule set that were not in any of the obtained dictionaries. As is clear, true identities (e.g. "cutie", "g") are mixed with noise words, like "blessing".

However, the resulting dictionary is nonetheless useful. In the 1000 tweets used for development and testing, 91% of all terms labeled as identities are found in this bootstrapped dictionary, and 96% are captured when we combine the bootstrapped dictionary with the existing dictionaries. This high level of coverage from our dictionaries allows us to focus heavily on precision.

4.3 Model Description

The prediction problem we address - determining whether or not each word in our text is an identity label or part of an identity label - is highly imbalanced. Only around 4% of the words in our labeled data are identities. Consequently, our modeling approach and our evaluation are geared towards techniques for imbalanced prediction problems. One such technique is to run a filter through the data to remove uninteresting words and thus reduce the imbalance. Consequently, we develop a two-stage model to predict, for each token in each labelled tweet, whether or not it was (or was part of) an identity label. The first stage of the model is a rule-based classifier that labels all stopwords and words not in any of our identity dictionaries as negative (i.e. as not containing an identity).

The second step of the model applies a straightforward, L1-regularized logistic regression model on each term in each tweet individually. While we expected that a sequential model (e.g. a CRF) would perform better on the task at hand, we found that a per-term regression approach with a strong set of features tended to perform as well or better than the sequential models we tested. We expect that this may occur due to the fact that the majority of identities (85%) were only one word long.

The features for each word, W_i, we used in our model are given in Table 4.3. The table displays three columns. The first column provides feature names. The second column gives the words for which the features are created. For example, for each word W_i, we use the Penn Treebank POS tag for the word W_i itself, the previous word W_{i-1} and the following word, W_{i+1}. The final column provides additional information, if necessary.

Features fall into one of three general categories. First, we use standard lexical features. These features include coarse-grained part-of-speech (POS) tags using the tagset described in [27], as well as finer-grained, Penn Treebank style tags. Lexical features also include various traditional word form properties (e.g. is the word capitalized?) and the Brown clusters utilized by [27] in their POS tagging model. The second set of features includes vector representations of the word W_i itself, its head word in the dependency parse and the word that exists at the end of W_i's "chunk". Importanty, the dependency parser provided by [22] for Twitter sacrifices the semantics of the Stanford Dependencies in order to provide reliable accuracy, and thus only provides rough dependency connections between words. Further, chunks are quoted as they are determined heuristically, largely by con-

[8]Note that $$$ will uniquely stand for the letters "nig" throughout the article due to the relevance of this set of identities. All other words we do not wish to print will be edited with *s

Lexical Features		
Penn Treebank POS tags	W_{i-1}, W_i, W_{i+1}	e.g. NNP, VBP
Coarse-grained POS tag	$W_{i-2}, W_{i-1}, W_i, W_{i+1}, W_{i+2}$	e.g. V, N
Prefix/Suffix, length 1 and length 3	W_i	From "liar", the set l, lia, iar, r
First letter Captialized, All Capitalized	W_i	e.g., "ALL_CAP"
Has digit, Is a Single Digit, Has a dash	W_i	e.g., "SINGLE_DIGIT"
Brown Cluster using data from [27]	W	"One-hot" vector encoding
Word Vector Features		
50-dimensional word vector	W_i	All zeros if word not in vocabulary.
50-dimensional word vector	Head of W_i in dependency parse	All zeros if word not in vocabulary.
50-dimensional word vector	Last word in chunk for W_i	All zeros if word not in vocabulary.
Dictionary Features		
Is in any existing identity or non-identity dictionary	W_{i-1}, W_i, W_{i+1}	Feature for name of each dictionary the word or its bi- or trigram is found in; e.g. *in_dict_wordnet_identities*
In bootstrapped dictionary at a particular cutoff	W_{i-1}, W_i, W_{i+1}	Cutoffs of 1000, 10000 and 100000 are used ; e.g. *in_bootstrap_dict_*1000
Is in stopword list	W_{i-1}, W_i, W_{i+1}	Generic stopword list for Twitter

Table 2: Features used in our statistical model

necting consecutive noun phrases together. More advanced chunking approaches [33] were too time consuming for our full dataset.

Finally, we incorporate features that use the afore mentioned dictionaries. More specifically, if a word or any bigram or trigram the word is in is found in a particular existing dictionary, we add a feature to the model to indicate this. So, for example, if the word "liar" were to be found in both the Affect Control Theory list of identities and the WordNet list of identities, it would have both the binary features *in_dict_wordnet_identities* and *in_dict_ACT_identities*. We use this approach because various dictionaries showed various levels of noise, and using features differentiated by dictionary name thus improved the model.

We take a similar approach in our utilization of the bootstrapped dictionary, assuming that the more frequently a word is captured by our patterns, the more likely it is to be an identity in any given tweet. We thus create various frequency "cutoffs" and use each as a feature. Consequently, if the word "liar" were to be extracted 10000 times by our lexical patterns, it would have both the binary features *in_bootstrapped_dict_*1000 and *in_bootstrapped_dict_*10000. This use of coarse-grained cutoffs worked better for this problem than using the actual frequency value itself (or any transformation of it we tried). Note that we also include dictionary features for the words before and after W_i, and that we consider all unigrams and bigrams a word is in when looking in the bootstrapped dictionary (identities in the bootstrapped dictionary are at most two words long).

4.4 Model Evaluation

We evaluate model performance in two ways. First, we use cross validation on the 1000 labeled tweets from our corpus, where we tune the regularization parameter for the logistic regression and analyze model performance with various feature sets. Note that five-fold cross validation is performed with an 85/15 train/test split instead of an 80/20 split because we ensure that only the 750 tweets selected at random from our corpora (and not the 250 selected via a keyword search) are used in the test set. Doing a full cross validation would only served to artificially inflate model performance.

We then analyze performance of the best model on the validation set. In both cases, the outcome metric of interest is the F1 score[9] on the positive (identity) class. We use F1 as opposed to a simple accuracy score due to the imbalance between the classes.

4.5 Baseline Model

In order to provide a useful comparison of model performance, we develop a simple but effective dictionary and rule-based classifier as a baseline. The classifier works in a similar fashion to the filter described above, with two differences. First, the baseline model is tested with various subsets of the identity dictionaries, as opposed to simply using all words in all identity dictionaries. Second, it uses POS tags, only labeling nouns as identities.

In sum, the baseline model classifies any noun in the list of identities it is given as an identity and labels all other terms as non-identities. We run this baseline model with all possible combinations of dictionaries (e.g. all dictionaries by themselves, all pairs of dictionaries, etc) to find the strongest baseline model to compare to, "optimizing" for F1 score on the 1000 tweets not in the validation set.

5. RESULTS

5.1 Model Performance

Figure 1 shows model performance on the five-fold cross-validation task. On the vertical axis of Figure 1 are the different feature combinations we tested, on the horizontal axis, the mean F1 score for the model with the optimal regularization parameter for that feature set. Error bars show one standard deviation, and results are given with and without the filtering step for each feature set combination. Figure 1

[9]The F1 score is the harmonic mean of precision and recall

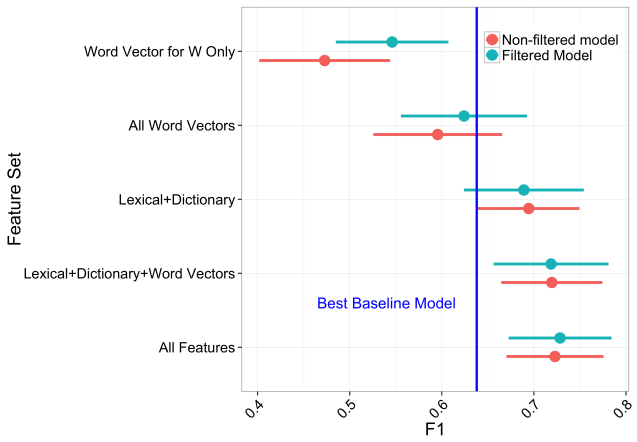

Figure 1: Cross-validation results. Different feature sets are given on the vertical axis, F1 score on the horizontal axis. Error bars are 1 standard deviation, different colors represent with/without the filtering step. The blue line represents the best dictionary, rule-based baseline

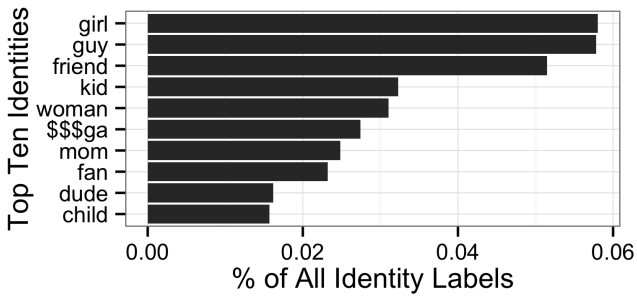

Figure 2: On the vertical axis, the top ten identities uncovered by the model. The horizontal axis shows the number of times this label was used as a percentage of the total number of identity labels in the corpus

also displays a vertical blue line that depicts performance of the best rule-based baseline. As the rule-based model did not need to be trained, error bars are not shown- we simply ran it once on the entire dataset.

Figure 1 shows that lexical, dictionary-based and word vector features are all predictive on their own. Combining all of these features into a single model results in improved performance, though word vectors for the head term and final chunk word add only a small amount of additional information. The best performing model overall is the full model using the two-step filtering, with an average F1 score of .74. Taking the best full model and applying it to the validation set, we find it performs slightly better than during cross-validation, with an F1-score of .76, an improvement over the baseline method (F1=.57) by 33%.

5.2 Error Analysis

Errors made by the model during both cross validation and on the validation set can be roughly categorized into four types. First, errors were made when terms typically used for identities were applied to non-human entities. For example the word "member" in the phrase "Big East member" refers to a university, not a person, that is a member of the Big East athletic conference. Second, much like our original annotators, the model had difficulties distinguishing modifiers from identities, particularly when these modifiers were applied to people. For example, using our definition of identity, the term "quarterback" in "quarterback Steve Young" is *not* an identity. Third, the model occasionally mis-classified organization names as identity labels (e.g. "Citizen" in "Citizens United". Adding NER labels from the classifier developed in [33] helped slightly, but substantially increased the amount of time it took to run the model. Finally, the model struggled with misspellings (e.g. "team-mate").

The errors we observed signify two possible avenues of future work. First, continued discussions with identity theorists and linguists may help to better understand how to address certain labeling issues, in particular those related to modifiers/compound identities. It is likely that our labeling process still had inconsistencies, which detracted from our ability to learn patterns in the data. Second, our work will benefit from leveraging additional types of features, in particular ones that leverage verb-based patterns (e.g. [14]).

6. CASE STUDY

6.1 Overview

For our case study, we first trained our classifier on the full set of labeled tweets. We then ran it on tweets from 250K users in our dataset that fit the qualifications provided in Section 3.1 (User had 50-15K tweets and <25K followers, no retweets, tweets with URLs or non-English tweets). As a heuristic to identify compound identities, we combine all sequential identity terms in to a single identity phrase. Such labels were sufficiently sparse that we do not expect this decision to have a strong influence on the general results we present here. Of the 750M tweets sent by users in our case study, 6% (45.4M) both fit our requirements and contained at least one term our model identified as an identity. Visual inspection of the model's predictions on tweets from a handful of users suggests that precision and recall are roughly the same as in our validation set, around 75% for both precision and recall.

In total, our model extracts 145K unique identity labels from the text, around 14K of which occur more than ten times. This number is considerably higher than the number MacKinnon and Heise drawn from WordNet (5.5K), suggesting, unsurprisingly, that a significantly higher level of both social complexity and noise is fostered via our approach. Figure 2 shows the top ten identities discovered along with their frequency of use, as represented by the proportion of all identities that they account for. Nearly one in three times a user expressed an identity, it was one of the ten terms listed in Figure 2. The identities shown generally fit intuitions - they capture two of our most obvious physical traits, sex ("girl", "guy", "woman", "dude") and age ("child", "kid"), as well as our most important forms of relationship - friendship ("friend") and kinship ("mom", noting that "dad" is the 11th most popular identity). Further, as we know that sports are a popular topic of discussion on Twitter [7], we were

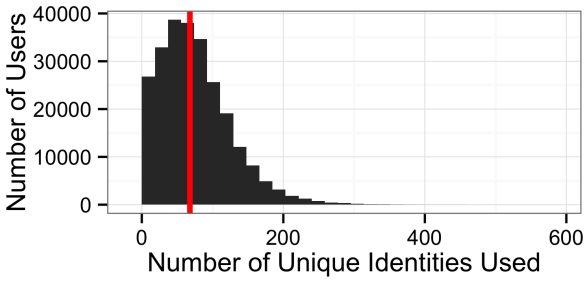

Figure 3: A histogram of the number of unique identity tweeted by each Twitter user. A red line has been drawn at the median of 68 unique identity labels

unsurprised to see frequent use of the identity "fan". Finally, given that we selected our sample based on a racially charged issue, we were also unsurprised that tweets in our dataset frequently contained the term "$$$ga".

Figure 3 plots a histogram of the number of unique identity labels expressed in tweets by each user in our dataset. The median user had 68 *unique* identity labels in their tweets. As people are well-known to discuss only a handful of different topics on Twitter [7], this result suggests that within each topic of conversation, individuals perceive a rich typology of identities.

6.2 Semantic Clusters of Identities

Our primary question for the case study was to understand how identities clustered into semantically coherent, "institutionalized" sets of identities. In order to extract these sets, or clusters, of identities, we make two assumptions. First, we assume that the use of each identity is a "mixture" over a finite set of latent identity clusters. This assumption is generally supported by social theory - recall, for example, that MacKinnon and Heise emphasize the differential association of identities to latent institutions they extract from a semantic network analysis [16]. Similarly, we assume that individuals are a mixture over institutions. This assumption is based on literature which suggests that social processes, like segregation, create disparities in the social contexts that we frequent. The social contexts in which we reside in turn influence our perceptions of the identities of those around us [37, 39].

In defining both people and identities as mixtures over latent, institutionalized sets of identities, a natural algorithmic fit to extract these latent identity sets is LDA [6]. We apply LDA to the user by identity matrix M, where each cell of the matrix $M_{u,i}$ represents the number of times a particular user u mentioned the identity i. To avoid issues with shorter documents, we only run the LDA on users with more than 50 unique identity labels. To avoid reliance on universally common terms, we drop identities tweeted by more than 50% of our users, and to address sparsity we drop terms tweeted by fewer than 100 users. After cleaning, we are left with 161K users (65% of the original set) and 4293 identities (approximately 5% of the full set captured).

The number of topics for LDA were set based on the domain knowledge. In particular, as Heise and MacKinnon observed only twelve taxonomic collections of identities in their qualitative analysis of WordNet, we kept the number of topics k to a lower number (30) than is traditional in the topic modeling literature. We use the version of LDA implemented in gensim [32] and allow the concentration parameter α to be estimated from the data.

Figure 4 shows one plot for each of the 17 identity clusters we were able to interpret from the LDA,along with one example uninterpretable topic that was typical of the 13 topics we could not provide a coherent label for. Labels for each cluster, provided by us, are given in the grey headers for each subplot. Within each subplot, we show the top ten identities for that identity set, along with the identities' associations to the topic as defined by the posterior from the LDA model.

Of the twelve identity taxonomies that MacKinnon and Heise identified from their qualitative study of WordNet, we are easily able to observe seven of them from the clusters extracted by our model. At least one and sometimes two topics were observed that closely identified with the political, kinship, religion, race/ethnicity, leisure/sporting, occupation and sexuality classifications they provided. This connection to prior work gives us confidence that our approach can reproduce traditional, denotive, taxonomic clusterings of identities. In addition to those taxonomic clusters found by MacKinnon and Heise in WordNet, we also find strong evidence for two addition clusters that fit this description, namely the school/college and military taxonomies of identities. These additional categories show the value of exploring semantic patterns in larger datasets using unsupervised approaches.

One advantage of Twitter in particular as a data source is that a small portion of tweets contain geospatial information. This allows us the opportunity to observe how spatial indicators of social context might influence the use of particular identities clusters by particular individuals. As an initial exploration of this, we consider how use of the "Race" identity cluster changes based on the racial make-up in the area a user tweets from most often.

From our data, we extract 71K who have at least one geotagged tweet from within the United States. For each user, we determine the county within which they tweet most frequently, and then retrieve information from the 2013 American Community Survey on the percentage of that county's residents who are African American. As the posterior distributions of user to identity sets was heavily bimodal, we discretize user associations to each identity cluster into a binary variable. Each geo-tagged user is thus represented here by the percentage of African Americans in their county and a set of binary variables representing whether or not she used is associated with each of the 18 identity clusters in Figure 4.

In terms of the expected relationship between context and use of identities in the Race cluster, two competing hypotheses are relevant. First, the *contact hypothesis* [29], perhaps the most well-tested social psychological theory, states that the more frequently we come in contact with someone of a particular race, the more favorable our view of that race will be. Given the negative connotation of several of the identities in this cluster (e.g. "$$$ger","slave", "negro"), we might thus expect that use of this cluster of identities is associated with a lower African American population in the users' county. In contrast, more general models of associative cognition (e.g. [3]) suggests that regardless of affect, the more

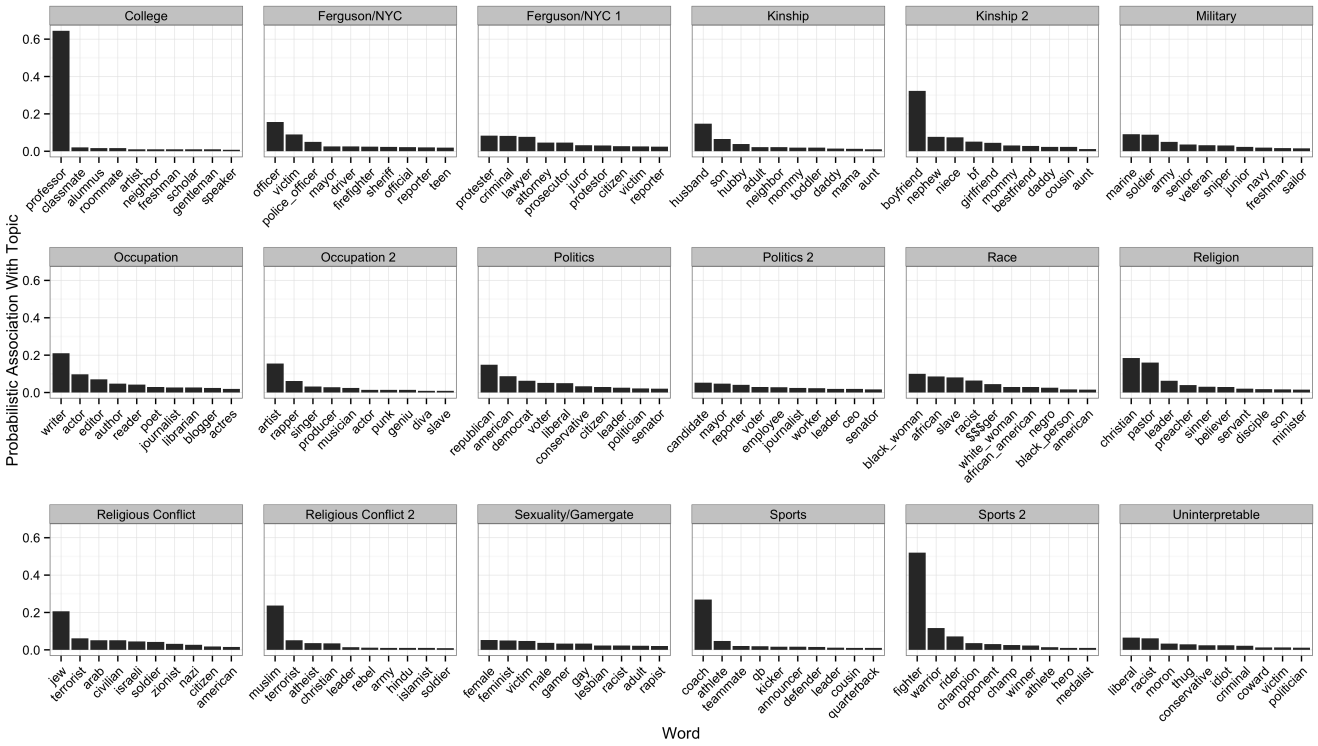

Figure 4: Results of the LDA. Each sub-plot is an interpretable topic. Within each plot we show the top 10 words associated with the topic. Bar height represents the probabilistic association of the identity to the identity cluster based on the posterior of the model

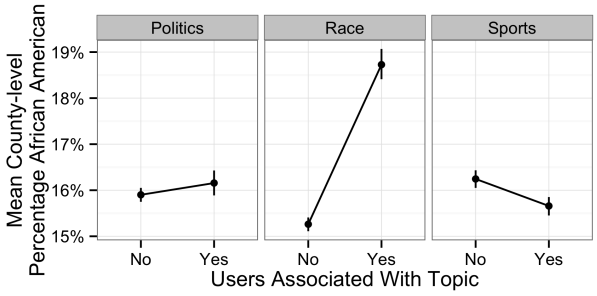

Figure 5: Differences in racial make up of geotagged users' counties for three identity clusters. The x-axis differentiates users who were associated and not associated with each cluster. The y-axis shows the percentage of the users' county that was African American. Error bars are 99% bootstrapped CIs.

we come in contact with someone having a particular identity, the more likely it is we will think about and talk about that identity and identities semantically associated related to it.

Figure 5 shows 99% bootstrapped confidence intervals for the average percentage of African Americans in a users' county for users that were and were not associated with three identity clusters - race, politics and sports. The latter two are included simply as points of comparison. We find that in places where race enters the social context to a greater degree-that is, in places where the percentage of African Americans is greater, people are more likely to use racialized identities. In contrast, we see in Figure 5 that there is little, if any, practical difference between the use of identity labels in the "Politics" and "Sports" identity sets according to race.

This finding by no means attempts to discredit contact theory, rather, it simply suggests that the denotive, purely semantic coherence of the Race identity cluster may be stronger than its affective coherence. Consequently, general theories of associative cognition are more applicable than theories which focus on affective relations across racial lines. Semantic coherence also seems to be important for the several identity clusters in our data that are specific to particular social issues. For example, two clusters are relevant to religious conflict, containing terms relating to both military and religious identities. The Sexuality cluster also contains the term "gamer", a connection to the Gamergate controversy that has led to frank discussions about sexism in the video gaming culture. Finally, as we would expect, we find two clusters relating to the events that occurred in Ferguson and New York City.

As these identity clusters cover emotionally charged and ever-evolving social issues, we would expect them to have a very different set of *affective* meanings than the more static and less emotional institutionalized topics found by MacKinnon and Heise in WordNet and replicated in our analysis. This assumption can be tested empirically by comparing affective meanings across the different identity clusters. In order to test this assumption, we take a coarse-grained mea-

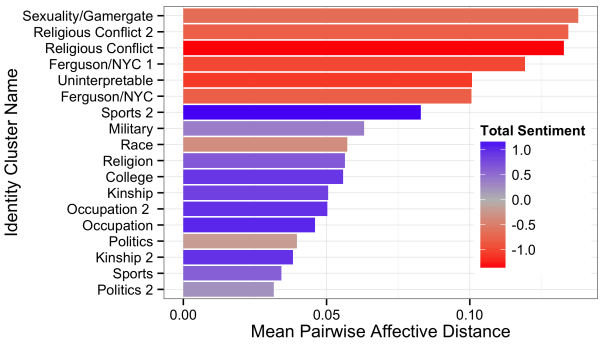

Figure 6: Affective meanings of the different identity clusters. Color indicates the sum of the affective meanings for each cluster, the size of the bar represents the mean pairwise distance between affective meanings of identities within the cluster

sure of affective meaning for each identity in Figure 4. We use Vader [20], a lexicon-based, tweet-level sentiment analyzer and apply it to each tweet containing an identity. The "affect score" for each identity is the average sentiment of each tweet that it is seen in. Importantly, before running Vader we remove from the sentiment lexicon all identity terms.

Figure 6 presents two pieces of information about the affective nature of each identity cluster. First, along the horizontal axis, we plot the mean pairwise distance of the affect scores for the identities that represent each cluster. The higher this value, the less similar the affective meanings of the identities within the cluster are. Second, the color of each bar represents the sum of the affective scores for the identities within the cluster. Here, the darker the red, the more negative the identities were in total, and the darker the blue, the more positive they were.

Patterns in Figure 6 support the qualitative suggestion above that identity clusters which develop around social issues have stronger affective contrasts. As we would expect, they carried significantly more negative emotion as well. These clusters are thus interesting in that they are semantically coherent but affectively disparate. While work remains to be done, our ability to uncover identity clusters fitting this description may help to learn more about how affective and semantic meanings co-evolve during complex social events, like those that occurred in Ferguson and New York City last year.

7. CONCLUSION

The present work makes two major contributions to the literature on identities and their use in text. First, we pose the novel problem of extracting all identities from a given corpora and develop a straightforward machine learning approach to address this problem. Our model outperforms a strong rule-based baseline by 33% on a validation set. Along the way, we develop new standards for determining what constitutes an identity label in text and produce a new, public, labeled corpora for other researchers to utilize.

Second, we perform a case study using our classifier on a large corpora of Twitter data collected around the Eric Garner and Michael Brown tragedies. We analyze semantically coherent clusters of identities and find they have important connections with a previous study on such structures [17]. We also observe identity clusters that are affectively disparate and highly negative but that are still semantically cohesive. A closer evaluation of the temporal and spatial dynamics of identities in these clusters, particularly those relating to the Eric Garner and Michael Brown tragedies, may provide unique theoretical insights into how semantic relationships between identities coevolve with their affective meanings. Finally, we consider how social contexts of Twitter users effect their use of particular identity labels. Specifically, we observe a positive association between the percentage of people in an individuals' home county that are African American and her use of racial identities on Twitter.

While these findings are fairly general and are supported by existing literature, conclusions should be taken with care for several reasons. First, we focus on a particular domain using a particular source of data. Second, the classifier we develop could be improved in several ways. In addition to issues stated in our error analysis, it is likely that we can make better use of our unlabeled data than by simply constructing a dictionary. We also should be able to leverage information in knowledge bases to improve classification and to perform Word Sense Disambiguation.

Regardless of these drawbacks, our case study findings present several interesting avenues of future work that can already be addressed with the tools developed here. For example, we have not yet considered how identities cluster on purely affective meaning as opposed to on semantic connections. Further, we have not perform any sort of temporal analysis. Finally, as opposed to affective analysis at the tweet-level, concept level analysis of sentiment would likely prove interesting in understanding how different people have different feelings towards the same identities.

8. ACKNOWLEDGEMENTS

We would like to thank the Data Science for Social Good fellows who volunteered their time for this article. We would also especially like to thank Jonathan Morgan for all of his time spent in discussion annotations with us, and to David Bamman and Waleed Ammar for discussions on the classification task. Support was provided, in part, by the Office of Naval Research (ONR) through a MURI N00014081186 on adversarial reasoning and the Office of Naval Research through a on State Stability under the auspices of the Office of Naval Research. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Department of Defense, the Office of Naval Research or the U.S. government.

9. REFERENCES

[1] A. Ahmed, L. Hong, and A. J. Smola. Hierarchical geographical modeling of user locations from social media posts. In *Proceedings of the 22nd international conference on World Wide Web*, pages 25–36, 2013.

[2] A. Ahothali and J. Hoey. Good News or Bad News: Using Affect Control Theory to Analyze Readers Reaction Towards News Articles. ACL, 2015.

[3] J. R. Anderson. *How can the human mind occur in the physical universe?* Oxford University Press, Oxford [etc.], 2007.

[4] D. Bamman. *People-Centric Natural Language Processing*. PhD thesis, Carnegie Mellon University, 2015.

[5] D. Bamman, T. Underwood, and N. A. Smith. A bayesian mixed effects model of literary character. In *Proceedings of the 52st Annual Meeting of the Association for Computational Linguistics (ACL'14)*, 2014.

[6] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, Mar. 2003.

[7] R. Bosagh Zadeh, A. Goel, K. Munagala, and A. Sharma. On the precision of social and information networks. In *Proceedings of the first ACM conference on Online social networks*, pages 63–74, 2013.

[8] M. Cikara and J. J. Van Bavel. The Neuroscience of Intergroup Relations An Integrative Review. *Perspectives on Psychological Science*, 9(3):245–274, 2014.

[9] L. Del Corro, A. Abujabal, R. Gemulla, and G. Weikum. FINET: Context-Aware Fine-Grained Named Entity Typing. *EMNLP'15*, 2015.

[10] J. Eisenstein, B. O'Connor, N. A. Smith, and E. P. Xing. A latent variable model for geographic lexical variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 1277–1287, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[11] A. Ekbal, E. Sourjikova, A. Frank, and S. P. Ponzetto. Assessing the Challenge of Fine-grained Named Entity Recognition and Classification. In *Proceedings of the 2010 Named Entities Workshop*, NEWS '10, pages 93–101, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[12] M. Fleischman and E. Hovy. Fine grained classification of named entities. In *Proceedings of the 19th international conference on Computational linguistics-Volume 1*, pages 1–7. Association for Computational Linguistics, 2002.

[13] E. Goffman. *The Presentation of Self in Everyday Life*. Anchor, June 1959.

[14] A. Grycner, G. Weikum, J. Pujara, J. Foulds, and L. Getoor. RELLY: Inferring Hypernym Relationships Between Relational Phrases. 2015.

[15] D. R. Heise. *Expressive Order*. Springer, 2007.

[16] D. R. Heise. INTERACT: Introduction and Software, Aug. 2010.

[17] D. R. Heise and N. J. MacKinnon. *Self, identity, and social institutions*. Palgrave Macmillan, 2010.

[18] T.-A. Hoang, W. W. Cohen, and E.-P. Lim. On modeling community behaviors and sentiments in microblogging. SIAM, 2014.

[19] D. Hovy, C. Zhang, E. Hovy, and A. PeÃśas. Unsupervised Discovery of Domain-specific Knowledge from Text. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 1466–1475, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[20] C. J. Hutto and E. Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth International AAAI Conference on Weblogs and Social Media*, 2014.

[21] K. Joseph, W. Wei, M. Benigni, and K. M. Carley. Inferring affective meaning from text using Affect Control Theory and a probabilistic graphical model, to appear. *Journal of Mathematical Sociology*, 2016.

[22] L. Kong, N. Schneider, S. Swayamdipta, A. Bhatia, C. Dyer, and N. A. Smith. A dependency parser for tweets. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing, Doha, Qatar*, 2014.

[23] T. Lin, O. Etzioni, and others. No noun phrase left behind: detecting and typing unlinkable entities. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 893–903. Association for Computational Linguistics, 2012.

[24] M. Lui and T. Baldwin. langid.py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 system demonstrations*, pages 25–30. Association for Computational Linguistics, 2012.

[25] T. Mikolov, W.-t. Yih, and G. Zweig. Linguistic Regularities in Continuous Space Word Representations. In *HLT-NAACL*, pages 746–751. Citeseer, 2013.

[26] G. A. Miller. WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41, 1995.

[27] O. Owoputi, B. O'Connor, C. Dyer, K. Gimpel, N. Schneider, and N. A. Smith. Improved part-of-speech tagging for online conversational text with word clusters. In *Proceedings of NAACL*, 2013.

[28] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. *Proceedings of the Empiricial Methods in Natural Language Processing (EMNLP 2014)*, 12:1532–1543, 2014.

[29] T. F. Pettigrew and L. R. Tropp. How does intergroup contact reduce prejudice? Meta-analytic tests of three mediators. *European Journal of Social Psychology*, 38(6):922–934, 2008.

[30] D. Ramage, D. Hall, R. Nallapati, and C. D. Manning. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, pages 248–256. Association for Computational Linguistics, 2009.

[31] M. Recasens, E. Hovy, and M. A. Marti. Identity, non-identity, and near-identity: Addressing the complexity of coreference. *Lingua*, 121(6):1138–1152, May 2011.

[32] R. Rehurek and P. Sojka. GensimâĂŞPython Framework for Vector Space Mo delling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, 2011.

[33] A. Ritter, S. Clark, and O. Etzioni. Named entity recognition in tweets: an experimental study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1524–1534. Association for Computational Linguistics, 2011.

[34] A. Saperstein, A. M. Penner, and R. Light. Racial Formation in Perspective: Connecting Individuals,

Institutions, and Power Relations. *Annual Review of Sociology*, 39(1):359–378, 2013.

[35] L. Smith-Lovin. The Strength of Weak Identities: Social Structural Sources of Self, Situation and Emotional Experience. *Social Psychology Quarterly*, 70(2):106–124, June 2007.

[36] P. Stenetorp, S. Pyysalo, G. Topic, T. Ohta, S. Ananiadou, and J. Tsujii. BRAT: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107. Association for Computational Linguistics, 2012.

[37] H. Tajfel and J. C. Turner. An integrative theory of intergroup conflict. In *The social psychology of intergroup relations*, pages 33–47. Brooks/Cole, Monterey, CA, w austin & s. worche edition, 1979.

[38] Z. Tufekci. Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls. In *ICWSM âĂŹ14: Proceedings of the 8th International AAAI Conference on Weblogs and Social Media.*, 2014.

[39] J. C. Turner, M. A. Hogg, P. J. Oakes, S. D. Reicher, and M. S. Wetherell. *Rediscovering the social group: A self-categorization theory.* Basil Blackwell, Cambridge, MA, 1987.

[40] W. Wei, K. Joseph, W. Lo, and K. M. Carley. A Bayesian Graphical Model to Discover Latent Events from Twitter. In *Ninth International AAAI Conference on Web and Social Media*, 2015.

[41] S.-H. Yang, A. Kolcz, A. Schlaikjer, and P. Gupta. Large-scale high-precision topic modeling on twitter. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1907–1916. ACM, 2014.

[42] L. Yao, S. Riedel, and A. McCallum. Universal schema for entity type prediction. In *Proceedings of the 2013 workshop on Automated knowledge base construction*, pages 79–84. ACM, 2013.

[43] W. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan, and X. Li. Comparing Twitter and Traditional Media Using Topic Models. In P. Clough, C. Foley, C. Gurrin, G. Jones, W. Kraaij, H. Lee, and V. Mudoch, editors, *Advances in Information Retrieval*, volume 6611 of *Lecture Notes in Computer Science*, pages 338–349. Springer Berlin / Heidelberg, 2011.