ANTI-SOCIAL MEDIA

Jamie Bartlett
Jeremy Reffin
Noelle Rumball
Sarah Williamson

February 2014

DEMOS

PARTNERS CREDITS

In collaboration with Open Society Foundations

Published by Demos 2014

Third Floor
Magdalen House
136 Tooley Street
London SE1 2TU

T 0845 458 5949
F 020 7367 4201

hello@demos.co.uk
www.demos.co.uk

Published by Demos 2014

Third Floor
Magdalen House
136 Tooley Street
London SE1 2TU

T 0845 458 5949
F 020 7367 4201

hello@demos.co.uk
www.demos.co.uk

Disclaimers:

(a) We have made the decision to focus this work on 'racial, religious, and ethnic slurs', and excluded homophobic, misogynist or other types of identity based slurs. This is not because they are insignificant or unimportant, but rather time and resource constraints.

(b) This is a study of slurs. It is not a study of 'hate speech', which is a separate subject, with its own legal, ethical and philosophical traditions. While the two do overlap at times, they are more helpfully understood as separate subjects.

(c) Following the convention in linguistics studies, we have decided to repeat highly offensive words, terms and phrases in full, as it is considered essential for rigorous detailed study of the subject. This also allows the reader to make a more informed judgement relating to the analysis undertaken.

EXECUTIVE SUMMARY

How to define the limits of free speech is a central debate in most modern democracies. This is particularly difficult in relation to hateful, abusive and racist speech. The pattern of hate speech is complex.[i] But there is increasing focus on the volume and nature of hateful or racist speech taking place online; and new modes of communication mean it is easier than ever to find and capture this type of language.

How and whether to respond to certain types of language use without curbing freedom of expression in this online space is a significant question for policy makers, civil society groups, law enforcement agencies and others. This short study aims to inform these difficult decisions by examining specifically the way *racial and ethnic slurs* (henceforth, 'slurs') are used on the popular micro-blogging site, Twitter.

Slurs relate specifically to a set of words, terms, or nicknames which are used to refer to groups in a society in a derogatory, pejorative or insulting manner. Slurs can be used in a hateful way, but that is not always the case. Therefore, this research is not about hate speech *per se*, but about epistemology and linguistics: word use and meaning.

In this study, we aim to answer two following questions:

(a) In what ways are slurs being used on Twitter, and in what volume?

(b) What is the potential for automated machine learning techniques to accurately identify and classify slurs?

Method

To collect the data, we scraped the publically available live Twitter feed (via its stream application programme interface) for all tweets

containing one or more candidate slurs over a nine day period (19 November – 27 November 2012). The list of terms judged candidate slurs was crowd sourced from Wikipedia.[ii] The tweets were then filtered to ensure that the slurs were contained in the body of the tweet and were not part of a user's account name and then passed through an English language filter to exclude non-English tweets. In total 126,975 tweets were collected: an average of 14,100 tweets per day. All of the tweets in our samples were publically available to any user of Twitter as a live comment (i.e. at the time the tweet is published by the sender).

Using this data set, we ran two types of analysis.

In study 1, we used automated machine classifiers to categorise the data sets. This involved human analysis of a sample to identify categories, followed by training a natural language processing technique to recognise and apply those categories to the whole of the data set automatically.

In study 2, we used human analysts to categorise subset samples of the data. This involved in-depth, iterative, analysis by researchers of small and then larger random samples of the data to reveal a stable set of categories.

Results (a): volume, nature and type of ways racial, religious and ethnic slurs are being used on Twitter

- We estimate that there are approximately 10,000 uses per day of racist and ethnic slur terms in English (about 1 in every 15,000 tweets). The ten most common terms found in our data set were (in order of prevalence) "white boy", "paki", "whitey", "pikey", "nigga", "spic", "crow", "squinty" and "wigga". The distribution was uneven across the terms, with "white boy" appearing in 49 per cent of tweets, and of the rest, only "paki" and "whitey" comprised more than five per cent of the total (12 and eight per cent respectively).

- **Slurs are used in a very wide variety of ways – both offensive and non-offensive.** We identified six distinct ways in which slurs are employed on Twitter: negative stereotype;

casual use of slurs; targeted abuse; appropriated; non-derogatory; and offline action / ideologically driven.

- **Slurs are most commonly used in a non-offensive, non-abusive manner: to express in-group solidarity or non-derogatory description.** Both human and machine analysis identified non-derogatory use as the largest category of tweets (estimated at 47.5-70 per cent of tweets, respectively). If casual use of slur terms is included in the human analysis (as in "pikey" being interchangeable with 'West Ham supporter'), the proportion rises to about 50 per cent. Both analyses also showed that relatively few tweets – from 500 to 2,000 per day – were directed at an individual and clearly abusive.

- **There were very few cases that presented an imminent threat of violence, or where individuals directly or indirectly incited offline violent action.** We estimate that, at the very most, fewer than 100 tweets are sent each day which might be interpreted as threatening any kind of violence or offline action. (This does not mean there are no other threats taking place which do not include the use of a slur).

- **Casual use of racial slurs account for between 5-10 per cent of use.** A significant proportion of use cases are what we have termed 'casual use of racial slurs', which means the term is being used in a way that might not be employed to intentionally cause offense (or where the user is unaware of the connotations of the term) but may be deemed so by others. The way in which racist language can seep into broader language use and reflect underlying social attitudes is potentially an area of concern.

- **Different slurs are used in very different ways.** Different slurs are used very differently. One of the most common terms, "whitey" is more often used in a non-derogatory, descriptive way compared to other terms, such as "coon" or "spic". There

are some indications that “paki” is becoming an appropriated term – a significant proportion of its use was by users identifying themselves as of Pakistani descent, despite it remaining in regular use as an ethnic slur.

Even though racist, religious and ethnic slurs tend to be used in a non-derogatory way on Twitter, this does not mean that hate speech is not being used on this platform. Language does not require the use of slurs in order to be hateful. We therefore do not make any broader claims about the prevalence of hate speech on this platform, an issue that warrants further study.

Results (b): what is the potential of automated machine learning techniques to accurately identify and classify racial and ethnic slurs?

Overall, the medium of Twitter provides an unprecedented source of data for studying slurs, and language use more generally. However, context is extremely important in determining the underlying significance and meaning of language, especially in contentious areas. For example, standing literature (and indeed our research) suggests that racial slurs can be appropriated by the targets of these slurs, and used in non-derogatory ways, defined as: ‘used without displaying contempt or causing hurt’. There are many versions of this, including humour, satire, and assumed in group norms (appropriation).[iii] This means that the relationship of a speaker to the group concerned is vital, but not always clear in the short text form tweets.[iv] This can make purely automated techniques quite difficult to apply.

- **Machine classifiers were extremely useful to identify and filter data sets into more manageable data sets.** The automated classifiers performed well in initially distinguishing between relevant and irrelevant tweets i.e. tweets where the terms were being used in racial or ethnic senses rather than unrelated senses.

- **On more nuanced categories – such as distinguishing between the casual use of slurs and targeted abuse, they performed less well.** Some of the categories created for the different types of slur usage were quite nuanced distinctions. The automated classifiers performed reasonably well at correctly identifying certain cases, although the smaller and more nuanced the category, the less well they performed.

- **Qualitative analysis was useful to determine nuanced categories.** Qualitative analysis of the data sets allowed the analysts to find more nuanced categories, such as appropriated use. Careful analysis of individual tweets also revealed how significant context is in determining meaning and intent – and how often it is lacking in the short form text of Twitter.

- **Even following detailed discussion, human analysts would often still disagree on meaning, intent and purpose.** Even where analysts discussed disagreements over how to classify specific texts (for example, whether it was 'non-derogatory' or 'casual use of slur') there remained continued disagreement. Two analysts working on the same data set were able to agree 69 per cent of the time when classifying the data. There were several reasons for this, including cultural biases.

Implications

We limited our data collection to a very short time segment on one social media platform. Even based on this short study, however, the use of social media data collection to understand trends and changes in language use is an excellent new resource for researchers – and especially linguists and those interested in the relationship between language and belief. We recommend that consideration be given to apply similar techniques for other language use: for

example conversations about certain groups and communities. However, it is vital that subject matter specialists be involved.

Twitter sampling works on the basis of key word matches. This type of analysis automatically creates systemic bias into the research method. Our use of a crowd-sourced word cloud was a simple way around this problem, but there are no doubt several other, fast changing, terms that we may have missed. It is certainly the case that automated key word matches are of limited power in respect of finding genuine cases of serious ethnic slurs or hate speech. Each case is highly contextual; and often will depend on approximations about the individuals involved.

Any conclusions drawn from these or similar data sets need to bear in mind these limitations. In general, therefore, language classifiers are extremely useful as tools to filter and manage large data sets. When combined with careful and detailed qualitative efforts, their use is magnified.

Overall, perhaps the key finding of this paper was the significant proportion of Twitter slurs that were found to be superficially non-derogatory. One working hypothesis is that the language used in tweets with such cases – and the sentiments expressed – reflect social norms within the sender's personal community. Notwithstanding the absolutes of legislation, these social norms are negotiable, contestable, and contested. For example, it has been argued that, in Britain, it is socially acceptable to use the term "chav" in a prejudicial (and typically insulting) context even though such a term arguably refers to a distinct and identifiable ethnic sub-group of the wider population. Other prejudicial terms were historically deemed acceptable, and are no longer seen as acceptable by society-at-large, but arguably remain acceptable to significant sub-communities within society. The ways in which slurs of this type may encourage or enable certain behaviours, or reflect certain underlying beliefs in society, deserves further consideration.

BACKGROUND

Hate speech

How to define the limits of free speech is a central debate in most modern democracies. This is particularly true in respect of speech that might be deemed hateful, abusive, or racist. Defining and legislating against this type of speech is extremely difficult, and has spawned a large philosophical, linguistic, theoretical, and legal literature.

There is no single definition of hate speech, although most definitions tend to cover utterances that demean, abuse or disparage a group of individuals based on characteristics such as race, sexual preference, gender or ethnicity. Other definitions include speech directed at individuals in the form of insults or threats.[v]

The pattern of hate speech is complex. In some places (or for some groups) it appears to be on the increase, while in others it is stable or falling.[vi] There is increasing focus on the volume and nature of hateful or racist speech taking place online. Although it is hard to find research that has statistics measuring the volume of hate speech online, it does appear to be increasing dramatically.[vii] This might reflect a change in the way we communicate rather than an increase in the amount of hateful speech taking place: communicating online makes it easier to find and capture instances of hate speech, because the data is often widely available and stored. Either way, there has been a large increase in the number of convictions against people using hate speech online, using both electronic communications legislation[viii] and broader hate speech legislation.[ix] (Whether this is a result of increases in volume or a keener interest in the subject from the relevant authorities is not certain.)

This may be especially marked following major events. Recent reports suggest there was a growth in anti-Islamic abuse on social media following the murder of Drummer Lee Rigby in May 2013.[x]

Importantly, some research has found that racist or hateful language often seeps into broader mainstream use: and the Internet may be facilitating that trend.[xi]

There are frequent public debates how to deal with racial slurs. In October 2013, UK football fans were warned that use of the word "Yid" – a slur used to describe Jewish people and sometimes used at football matches – could result in arrest and prosecution.

Racial, religious and ethnic slurs

Hate speech sometimes includes the use of racial or ethnic slurs. Slurs are similar to hate speech, although they relate specifically to a set of words, terms, or nicknames that are used to refer to groups in a society in a derogatory, pejorative or insulting manner. They usually "convey negative, emotional content beyond the truth conditional content they are normally taken to encode" and are generally understood to "convey contempt and hatred towards their targets".[xii] Part of a slur's derogatory nature is that it is broadly recognised and understood by the user as having those derogatory implications; and that there are non-derogatory alternatives that are intentionally *not* used.[xiii]

Slurs are distinguishable from other kinds of terms, such as descriptive and expressive (emotionally charged, often hurtful) words because they are taken to target certain groups on the basis of a descriptive gesture such as race or sex.[xiv] Ethnic slurs are specifically slurs on the basis of race, ethnicity, nationality or religion.[xv]

Slurs of this type are thought to have played – and continue to play – an instrumental role in the perpetuation of race-based discrimination because they offer a linguistic resource with which to dehumanise or diminish targets. Although some academics consider slurs in and of themselves hate speech (this is sometimes called a 'semantic' approach), most consider that they are not always by definition necessarily offensive, and can be appropriated

and used in a variety of ways, not all of which are hateful (called the 'pragmatic' approach).[xvi]

Certainly, the use of these slurs is clearly not always hateful: a number of studies have examined how the slur "nigger" has been appropriated by African Americans as a way of actively rejecting the connotations it carries: for comedic purposes, a status symbol, a shorthand term expressing familiarity among friends, or even forgetting what the term ever denoted in the first place.[xvii] Equally, much hate speech does not need to include any specific racial slur at all: rather aggressive and hateful use of non-offensive words.

Slurs have recently attracted growing interest from linguists and philosophers of language. Most studies of slurs have been limited to relatively small data sets from ethnographic research, or larger studies of word use in mainstream media. Given individuals often use slurs in closed settings – at home or with friends – data has been hard to collect.

However, as a new space where racial slurs are employed, social media has also become a new space where study of the phenomenon is possible. Indeed, as we have argued elsewhere, Twitter has become an increasingly useful source of data for understanding several social phenomenon.[xviii] There have been a small number of studies of language use on Twitter, where machine learning and natural language processing techniques have been employed to detect use patterns (see below for a technical description). For example, in 2012, Floatingsheep ran an analysis of all instances of geo-located Tweets between June 2012 and April 2013 (only around 1-4 per cent of Tweets are geo-located) and found around 150,000 examples of hateful slurs being used over the period.[xix] However, the use of automated tools to collect and classify data sets in this way is an emergent field, with several limitations at present.[xx]

STUDY 1: AUTOMATED ANALYSIS

Methodology

We scraped the publically available live Twitter feed (via its 'stream' application programming interface) for tweets containing one or more racial slurs. We used the crowd sourced Wikipedia as a starting point,[xxi] but our final list was shorter, excluding a number of highly ambiguous terms (for a full list, see annex [2]). We then filtered the tweets to take out any instances where the only slur term(s) were part of an account name. The tweets were then passed through an English language filter to exclude non-English tweets.

The resulting data set collected with this filtering process over a 9 day period contained 126,975 tweets , on average 14,100 per day. (Roughly 160 million tweets are sent in the English language per day at the time of the research being undertaken).

All of the messages in our samples were publically available to any Twitter user as a live comment (i.e. at the time the tweet was published) if the user was either a follower of the sender, or if the user was scraping Twitter using keywords and the tweet contained one of those keywords (this is the route we used). Typically, a tweet can be accessed by a Twitter user for up to 14 days after the time of publication, provided that neither Twitter nor the original sender has deleted it.

Social media datasets, including those gathered in this report, are often too large to be manually analysed, and require automated ('machine') analysis. Machine analyses are those conducted by a computer, and are capable of processing social media data at great scale and speed.

Our study makes use of a software platform called the Agile Analysis Framework (AAF), which is designed to help the researcher isolate tweets of interest and then identify and quantify the different ways in which language is used in those tweets ('patterns of usage' analysis).[xxii] AAF allows the researcher to construct standard and bespoke filters (what we call classifiers). Each classifier automatically places tweets into certain (human-defined) categories, allowing the researcher to

iteratively sort very large data sets into separate categories for further study.

The work was conducted in four phases:

- Phase I: The analyst took a small random sample of the data set (20 – 100 tweets) and examined these data and tried to identify the most frequent pattern of usage. This process follows grounded theory methodology, so in the default approach, the data are supposed to suggest the usage pattern, rather than the analyst imposing his/her own pre-conceived expectations.
- Phase II: The analyst manually annotated a set of tweets (typically 100-200 tweets). The classification was either two-way ('identified pattern', 'retain for further investigation') or three-way ('identified pattern', 'retain for further investigation', 'irrelevant'). This allowed the analyst to subject the classification idea to a practical test, and provides the system with a 'gold standard' set of "correct" answers.
- Phase III: The analyst trained the machine-learning classifier. Classifiers attempt to replicate the classification decisions of the analyst. This is achieved by the manual annotation of a further set of tweets (typically 200-500 tweets). The computer then uses machine-learning algorithms to find correlations between human classification decisions and linguistic features in the tweet. Having learned these associations, the computer can then apply inferred generalized pattern correlations to un-classified tweets and make its own classification decisions.
- Phase IV: The analyst reviewed the performance of each trained classifier against the (previously unseen) gold-standard data set.

Using this process, the analyst built a series of classifiers to categorise the data set. Together, these classifiers formed a 'classification cascade' through which tweets could be passed in order to assign them to one of our established categories. These classifiers were:

- Tier 1: Relevance Classifier. This classifier takes the scraped data as input and places tweets into two classes: 'relevant' or

'irrelevant'. The intended criteria were: classify as relevant tweets that are using the words as racial or ethnic terms. Classify all other tweets as irrelevant.

- Tier 2: GroupID Classifier. The input for this classifier is the 'relevant' output from Tier 1. It divides these tweets into three classes: 'further investigation', 'group ID', and 'irrelevant'. The intended criteria were: classify as group ID all tweets using the terms in a manner that indicates demarcation of group boundaries, but the use of the term is otherwise basically non-prejudicial in the context of the tweet. Classify as further investigation all tweets otherwise using the words as racial or ethnic terms. Classify all other tweets as irrelevant.
- Tier 3: Casual Use of Slur Classifier. The input is the Tier 2 further investigation class, placing the output into three classes: 'further investigation', 'casual use of slur, or 'irrelevant'. The intended criteria were: classify as 'casual use of slurs' all tweets using the terms as racial or ethnic slurs but in an offhand or casual fashion. Classify as further investigation all tweets otherwise using the words as racial or ethnic terms. Classify all other tweets as irrelevant.
- Tier 4: Ideological Classifier. The input (again) is the Tier 3 further investigation class, placing the output into three classes: 'personal attack', 'ideological', or 'irrelevant'. The intended criteria were: classify as personal attack all tweets that direct racist abuse at an individual or small group (apparently) known personally to the sender. Classify as ideological all tweets that are making a political statement or call to action in the real world while employing racial or ethnic terms. Classify all other tweets as irrelevant.

RESULTS

Prevalence of terms

In total, roughly 14,000 English language tweets per day contained at least one of the slurs searched for; and around 10,000 of them used the term in their racial, ethnic or religious sense (rather than, for example, "crow" being used to describe the bird).

The most prevalent term in our data set was "white boy". Whether this ought to be classified as a racial slur is controversial (see below). Given the prevalence of the terms – and the fact that many more tweets are sent from the US than other English language speaking countries – it is not surprising that several of the terms are American in orientation. The top ten most common terms are listed below (note, the average daily use does not reflect that there were occasional spikes in certain term frequency).

Table 1: Frequency of slurs

Expression	% of tweets	Average Daily Use[xxiii]	Cumulative %
white boy	48.9	4,890	48.9
Paki	11.7	1,170	60.5
Whitey	7.9	790	68.5
Pikey	4.1	410	72.5
Coon	3.2	320	75.7
nigga	3.2	320	78.9
Spic	3.0	300	81.9

Crow	2.1	210	84.0
squinty	1.8	180	85.9
Wigga	1.7	170	87.6

Pattern of usage

On the basis of the analysis and our classifiers we identified four categories of patterns of usage. These were as follows:

Group ID. Racial/ethnic tags are used consciously as racial or ethnic terms. However, the terms are here used to demarcate group boundaries (i.e. whether someone is inside or outside a social group) in a manner that is broadly non-prejudicial in the context of the tweet. The tweets are typically not at all heated (i.e. they are not highly invective or highly emotional in tone).

> Example: @^^^: *Whew Brady Quinn is one sexy white boy!*

Directed attack. These tweets consciously use racial/ethnic slurs to direct abuse at an individual or group (apparently) known personally to the sender. They are invariably very heated. In many cases, single conversations may lead to several tweets of this nature back and forth, so these figures may reflect a few hundreds of conversations worldwide each day.

> Example: *fUcK yOu In ThE aSs PuNk AsS wHiTe BoY*

Casual use of slurs. These tweets contain a racial/ethnic slur, but are using the slurs in an offhand or casual fashion. The message is typically not particularly heated. It perhaps suggests underlying racial/ethnic prejudice, but conveyance of that viewpoint is usually not the point of the message.

> Example: @^^^: *Fucking paki shops that charge you for cash withdrawals or paying by card need nuked*

Ideological. These tweets consciously use racial or ethnic slurs within a political statement or a call to action in the real world. The message is typically not particularly heated and may make a broad claim about the state of the world.

Example: *The raghead / muslims will subjugate us FROM WITHIN... Obama is on their team. If you voted in Obama.. stand up... http://###*

The average number of tweets per day reflects the estimates made by the classifiers about the frequency of these use types. As discussed below, in some categories, the classifier performance was not uniform across use type: therefore the average number of tweets per day should be taken as an approximation.

Table 2: Use Types

Classifier	*Invective*	*Racial awareness*	*Other features*	*Average number of tweets per day*
Group ID	Very low	High	Used to mark people as in or out group members of the writer's group	c.7,000
Targeted abuse	Very high	High	Directed at someone known to the writer	c.2,000
Casual use of racial slurs	Low-medium	Low-medium	Main point of the message may not be race or ethnicity	c.1,000
Ideological	Low – medium	High	Proposes of justifies action in the real world	<100

Classifier performance

Measuring the classifiers against a gold standard data set allows you to determine how well the machine classifies tweets compared to a human analyst. Machine learning technology is still at an early stage of development, so a failure to model the human decisions may just reflect the weakness of its learning procedure rather than an invalid categorical distinction between usage patterns.

Table 3: Classifier Performance

Classifier	*Recall score*	*Precision score*	*Overall performance (F1)*
Relevancy	0.97	0.84	0.90
Group ID	0.79	0.75	0.79
Targeted abuse	0.54	0.57	0.56
Casual use of slur	0.23	0.30	0.26
Ideological	Too few cases	Too few cases	Too few cases

Annex 2 provides a detailed analysis of the performance of the classifiers.

Overall, the classifiers did a good job of distinguishing between relevant and irrelevant tweets, i.e. tweets where the terms were being used in racial or ethnic senses rather than unrelated senses (e.g. “whitey” as a colloquialism for “unwell”). The classifier had a precision score of 0.84 for relevant tweets (i.e. 84 per cent of the tweets it classified as relevant were actually relevant), and a recall score of 0.97 (i.e. 97 per cent of truly relevant tweets were classified as such). The ‘F1 score’ is a rating combining precision and recall. It was 0.90, which is high (random guessing would lead to a score of 0.50).

However, as the classifications became narrower, the classifiers were increasingly less effective.

In the next stage, the classifier attempted to identify tweets from the 'group ID' pattern of usage. Performance was not unreasonable. In a 3-way classification, the precision score was 0.75 (75 per cent of gold standard tweets classified by the machine as group ID had been manually classified as such) and recall was 0.79 (79 percent of gold standard tweets manually classified as group ID were classified as such by the machine). The F1 score was 0.79. This good level of performance (random guessing would give a score of 0.33) suggests that the machine was able to lock on to objective linguistic features in the text that corresponded to a distinct pattern of usage.

Support for the other proposed usage patterns is somewhat weaker. In the next tier, the classifier attempted to learn the distinction between tweets with casual use of racial slurs and more inflammatory tweets (personal attacks and ideological messages / calls to action). In a 3-way classification, precision for the classification of the inflammatory usage patterns was 0.57 and recall 0.54, for an F1 of 0.56. Whilst better than chance (0.33) this performance is not particularly strong. More interestingly, performance at detecting casual use of slurs was extremely weak. On the same 3-way classification, precision was 0.30 and recall 0.23 for an F1 score of 0.26. In other words, the classifier could not learn any general objective linguistic features that allowed it to recognize members of the casual use of slurs usage pattern. There were too few examples of the proposed 'ideological' pattern to be able to train a classifier. This failure to "lock on" to the casual use pattern might reflect a weakness in either the proposed classification scheme itself or a weakness in the machine learning algorithms to capture the patterns on which human judgment relies (or both).

Observations by the analyst, however, also suggest potential questions with the classification itself. On manual annotation, it was difficult to maintain a stable boundary between the 'casual use' and more inflammatory (personal attack) categories from session-to-session. Furthermore, it appeared that some terms were consistently placed in one usage pattern compared to another; the questions arose whether (i) these distinctions in usage pattern were also term-specific; and (ii) whether the analyst's responses were specific to his own cultural

background and experiences; for example, the researcher responded to otherwise identical tweets differently based on the racial term itself.

STUDY 2: MANUAL ANALYSIS

In addition to the automated analysis above, we then undertook a detailed manual analysis of a random selection of tweets drawn from the same data.

Methodology

We started with a draft list of categories into which the tweets could be meaningfully divided, based on a short literature review. A group of 50 tweets were randomly selected from our data set using a random number generator. Two analysts – separately and without discussion – then categorised each of the 50 tweets into the categories (creating new categories if necessary), and then discussed and compared their results. To reduce inter-annotator disagreement, and in response to specific themes found within the group of 50 tweets, the categories were clarified and expanded.

Using methods borrowed from grounded theory the analysts continued the process iteratively until there were no additional categories required, at which point we felt that a 'saturation point' had been broadly reached – i.e. further analysis was no longer having an effect on the results. This process of analysis, discussion, and amendment of categories was repeated a total of five times across three different sets of 50 tweets, where the final amendment of categories was a reduction of the number of categories. We then moved onto larger groups of tweets – one of 250, and one of 500. Interestingly, during the process, the percentage split across categories never entirely stabilised; the reasons for this are discussed below.

Results

Following our manual mark-up process, we arrived at eight main categories of use. These were the following:

Negative Stereotypical Attitude. A derogatory stereotypical attitude, ascribing physical or behavioural attributes to an individual or a group, directly or indirectly.

> Example: *Ain't nothin worse than a corny nigga. Wait yes there is. A corny loud nigga smh. Coon*
>
> *Cannot believe some pikey shit stole the seat from my bike today! #thecheek #stillfuming*

Casual use of slurs. A term associated with a particular group is being used in a derogatory way, but there are no physical or behavioural attributes being ascribed, and the term could be swapped out for a term without slur connotations without affecting the meaning of the tweet.

> Example: *having an emosh breakdown because i cant play harry potter :@@@ fucking pikey controllers not working why ps1&2 fukin pikey why :(((H E L P*

Targeted Abuse. Slur words are being tweeted directly at a specific person with the intent to cause harm or distress. This can include casual use of slurs or derogatory stereotypes, so long as it is specific and is some kind of personal attack at "you". Tangential references to specific people (i.e. including @^^^ signs within the tweet) are not enough on their own to be considered 'targeted abuse', nor is abuse towards specific third parties (e.g. Obama-bashing).

> Example: @^^^ *Hahahahahah thts alll u got fucken bitch like go fucken suck a cunt like I said bitttch ass nigga u fucken spic wetback Beaner*
>
> @^^^ *you dirty little spick!*

Appropriated. A group is using (or reclaiming) for itself a term normally considered negative or derogatory. This can be done sarcastically or straight.

> Example: *Omg I love being a spic :D*
>
> *I'm a dirty paki and I'll blow up London in 5 years #ha*

Non-derogatory. A term is being used in a descriptive or otherwise neutral way, or where a stereotype is applied, but it is not hurtful or derogatory.

Example: *You know the world is gonna end when a white boy drops 138 PTs in a college basketball game #CRAZY*

@^^^ yeah u shud tbh. Coz not eating doesn't help. Just don't eat paki food. It's not the best when ur Ill

Offline action. An explicit incitement to do something "in the real world", whether that's going on a march or killing someone.

Example: *Attention all white boys, come and holla at some real niggas. RT @^^^: Me and @^^^ on our white boy search.*

*#101HispanicWaysToDie Come across the border Spic bastard *loads shotgun**

Impossible to Determine. The expectation here is that with specific additional information (e.g. the rest of the conversation or the ethnicity/group membership of the sender/receiver), it will clearly fall into one of the other categories, but without that crucial piece of information, it's not clear. Primary use was where a tweet would be non-derogatory if sent by a member of the group, but derogatory if sent by someone outside the group.

Example: *@^^^ that's what you get , Ayye coon did you tell that girl what I told you*

@^^^ YO WHATS UP MA NIGGA TELL MA NIGGA WHITEY GOT AT ME

Error. Tweets that should not form part of the analysis. They either do not contain a slur word, or if they do, it is accidental, or is a wholly non-racial use of the word, or a person is describing their experience / perception of a slur being used ("meta" conversations about language use) to a third person. Anything incomprehensible was also included here.

Example: *@onetrey_thereal I Would Neva Fuck Rikko Or Nate...So Fall Bak Goofy Dnt Yhu Dink If i Fuck'd Them Dha Block Would've Heard About It*

Grand Forks is warmer than Coon Rapids right now #winning [Coon Rapids is a suburb of Minnesota]

*#WoG #Gratitude Week heads to the homestretch w/ Wizened Merry Fools, Insane Logic and Soul ;) @^^^ @^^^ @^^^*

My boyfriend just told me to \shut it paki.\ Should this be classed as racism!?

The table below sets out the total daily tweets reflects an estimate range based on extrapolating the analysts' different results. They should be viewed as estimates, rather than comprehensive.

Table 4: Use Pattern

Categories	*250 tweets (iteration 6)*		*500 tweets (iteration 7)*		*Total Daily Tweets*[xxiv]
	Analyst 1	*Analyst 2*	*Analyst 1*	*Analyst 2*	*Range*
Negative, Stereotypical Attitude	21%	24%	15%	7%	1,100 - 2,250
Casual Use of Slurs	16%	16%	4%	5%	450 - 1,600
Targeted Abuse	4%	15%	4%	4%	400 – 950
Appropriated	6%	3%	5%	5%	450 – 500
Non- derogatory	28%	30%	53%	59%	2,900 – 5,600
Offline action	0%	0%	1%	0%	0 – 50
Impossible to Determine	19%	6%	7%	15%	1,100 - 1,250
Error	7%	6%	11%	5%	650 – 800

Manual coding and categorizing performance

In manually coding, analysts had considerable difficulty achieving inter-annotator agreement (the extent to which both annotators agreed on the meaning of the same tweet). Once we had settled on an agreed coding category and then increased the sample size, there was a gradual improvement. The table below shows the evolution of the agreement scores.

Table 5: Inter-annotator agreement

Iteration	Number of categories	Two-Way Agreement
1 (50 tweets)	11	65%
2 (50 tweets)	12	38%
3 (50 tweets)	11	39%
4 (50 tweets)	13	65%
5 (50 tweets)	8	59%
6 (250 tweets)	8	48%
7 (500 tweets)	8	69%

We suggest that there were four major reasons for this, which illustrate some of the difficulties with the data set itself.

- **Unanticipated themes not catered for in definitions**. At each discussion stage, unexpected themes came up that had not yet been explicitly dealt with, and different analysts chose different categories. For example, inter-annotator agreement for the group of 250 tweets (iteration 6) was particularly

impacted by this, where 7 per cent of tweets were conversations quoting slur terms:

> *@^^^ Yes it is and I agree but there's a real difference between calling someone a pikey or cunt & singing songs about a plane crash*

These were systematically categorised differently by different analysts: one as non-derogatory because they were meta-discussions, and another as derogatory according to the nature of the slur being discussed. Having agreed on how to deal with a particular theme, inter-annotator agreement is improved, but only for that theme, which might only have been prevalent within a particular sample of tweets.

- **Multiple patterns of usage in a single tweet**. This is best explained with specific examples:

> *@^^^ I have a job, you immigrant cunt. Soldiers have died but I hope more muzzies are killed in the coming days #MuzDead*

In this case, there are several types used: 'targeted abuse'; 'offline action'; and the statement taken as a whole could be interpreted as 'negative stereotype'. Each analyst can legitimately defend their position in discussions afterwards, and so the level of inter-annotator agreement is not improved.

- **Terms whose usage straddles the divide of 'ethnic slur'**. Two extremely common terms, "white boy" and "whitey", generally were difficult to categorise. "White boy" is frequently used to identify individuals, but sometimes in the context of expressing a negative attitude:

> *Dear waiter: about the dirty rag on your head to hide your dreds: YOU ARE WHITE. Dreds on a white boy just makes you look homeless.*

Both terms created inter-annotator disagreement on the boundaries of derogatory and non-derogatory, and whether it should be an error due to the complex ways in which the terms are used.

- **The presence of analyst-specific cultural bias**. Many of the tweets depended on considerable contextual knowledge that analysts may or may not have (and may or may not be trying to avoid affecting their judgment). For example, there were a number of tweets referring to a particular American college basketball game in which a white player scored 138 points, with a combination of "whitey", "white boy", and "nigga(s)" within each tweet. For one analyst, anything referring to this event was obviously non-derogatory, because of the genuine rarity of professional-ability white basketball players. Other analysts believed this was projecting preconceived notions about how people think about basketball, and categorised some tweets as non-derogatory, but others as casual use of slurs or negative stereotyping. It is very difficult to adjust for cultural bias when reading and analysing speech of this kind. Discussing personal contexts in relation to categorisation made some impact on subsequent processes, but sometimes in a way that created inter-annotator disagreement – for example, after one process the two analysts agreed that one was under-using a category, while the other was over-using it. In the next iteration, the positions were reversed as the analysts tried to overcome their bias.
- Nevertheless, the analysts showed high levels of agreement for the non-derogatory category. This reflects the experience in study 1 where the automated classifier was able to imitate quite well the analyst's classification decisions for the near-equivalent 'group ID' usage pattern. The analysts in study 2 were also able to demonstrate good agreement for tweets with casual use of slurs, in contrast to study 1 where the classifier was unable to imitate the human decisions. This may be attributable to the decision to create a new category 'negative stereotypical attitude' which appears to have overlapped heavily with the broader 'casual use of slurs' category in study 1. Dividing up the usage patterns in this way may have produced cleaner definitions and greater clarity. Furthermore, this may also reflect the weakness of the machine learning system. Humans were able to agree on a definition and rely on features of the

tweets that the machine learning system was unable to recognise. Nonetheless, continued variation in the scoring for negative stereotypical attitude on iteration points to some remaining instability of definition.

The 'impossible to determine' pseudo-category varied both between and within analysts. Whilst allowing this category is useful in highlighting the lack of required contextual knowledge in some situations, the amount of this category observed is largely a function of the analysts' degree of caution, which can vary both across analysts and between experiments for the same analyst. A "forced choice" paradigm – as used in study 1 – avoids this problem, but fails to highlight the fact that in some cases (5-10 per cent of tweets in these experiments) there is insufficient data available to make anything beyond a random assignment of usage pattern.

There was a large variation in prevalence by usage pattern between iterations 6 and 7, even for categories where there was high inter-annotator agreement. The analysts postulated that the reason the percentage split across categories was so varied was due to the fact that different slur terms made up very different proportions of each sample, and that each slur term had a very different type of use pattern. For example, tweets with "white boy" were rarely categorised as negative stereotypical attitudes or casual use of slurs, but were frequently categorised as non-derogatory. Tweets containing "white boy" accounted for 29 per cent of iteration 6 and 65 per cent of iteration 7. Similarly, "pikey" was frequently categorised as a casual use of slurs, and accounted for 23 per cent of iteration 6, but only 4 per cent of iteration 7. The analysts were confident that, with sample sizes with stable percentages of terms used, the categories would stabilise.

This issue of patterns of usage by specific term was therefore separately studied in a second experiment, below.

Specific term use

Because the 250 tweet (iteration 6) and 500 tweet (iteration 7) groups are random samples from the data set, certain terms were more prominent than others. Forty nine per cent of the data set contained

tweets with the term “white boy”, which researchers were concerned might be skewing the results, and/or obscuring the ways in which different terms were employed. (The contrast in this methodology to study 1, where the most common classified usage pattern is usually (largely) automatically removed from the data set between iterations, allowing other less common patterns of usage to become apparent.)

Therefore, 400 tweets were randomly selected, comprised of approximately 50 tweets for each of the following terms: “coon”, “crow”, “nigga”, “paki”, “pikey”, “spic”, “whitey” and “wigga”. These were selected as the most commonly used terms within the data set after “white boy”. Fifty tweets is only a very small data set, and was selected to provide a general sense of major differences in the data, so should not be considered as comprehensive.

The following table shows how annotators assigned the most common slurs across the different categories. It also shows (in brackets) category weightings if we restrict ourselves to those for which we got a definite assignment, by ignoring those tweets that were assigned either to the “error” category or the “impossible to determine” category.

Table 6: Annotator Ascriptions of Categories to 400 Tweets, in percentage

	coon	crow	nigga	paki	pikey	spic	whitey	wigga	All
1. Negative Stereotypical Attitude	20 (40)	9 (90)	16 (20)	32 (43)	21 (22)	22 (31)	11 (19)	10 (15)	18 (27)
2. Lazy use of racial slurs	6 (12)	0	3.0 (4)	8 (10)	64 (69)	3 (4)	13 (23)	1 (1)	13 (20)
3. Targeted Abuse	14 (28)	0	4 (5)	5 (7)	4 (4)	14 (20)	5 (9)	10 (15)	7 (11)

4. Appropriated	6 (12)	0	3 (4)	8 (10)	1 (1)	13 (19)	2 (3)	6 (10	5 (8)
5. Non-derogatory	4	0	54 (67)	23 (30)	4 (4)	18 (26)	26 (46)	39 (59)	22 (34)
6. Offline action	0	0	1 (1)	0	0	0 (0)	0	0	0.1 (0.1)
7. Impossible to Determine	27	9	12	14	1	14	18	31	16
8. Error	23	81	7	10	6	16	25	3	19

Judging by these results, different slurs tend to be used in quite different ways. For example, over half the cases of "nigga" were used in a non-derogatory way. The terms "paki", "spic" and "coon" were less likely to be used in a non-derogatory way, and more likely suggestive of stereotyping.

Not all racial slurs are equally offensive. The term "whitey" is worth further consideration, because of its prominence in our data set (7.9 per cent). Surprisingly, there is very little written about it, but it was likely a common slur term during the American civil rights movement in the 1960s and 70s. (For example, Gil Scott-Heron's 1970 song "Whitey on the Moon".) It certainly remains a charged word. In 2008, the Obama campaign was concerned enough about rumours that Michelle Obama had used "whitey" to post an official online rebuttal. [xxv] Maruse Heath, head of the Philadelphia chapter of the New Black Panther Party, has "Kill Whitey" tattooed on his face. [xxvi] There does

not seem to be any societal agreement on whether or not the term "whitey" is a racial slur.[xxvii]

Some of the other results are perhaps not surprising, like "pikey" being associated with 'casual use of slurs'. For example, Tottenham Hotspur FC fans (infamously) adopted the term "pikey" to describe West Ham United FC fans, in a connotation that has (arguably) no links with its original meaning. (This argument is less convincing when applied to West Ham fans' use of Jewish slurs for Spurs supporters.) However, there are some conclusions that are perhaps less obvious. "Spic" has an unusually broad distribution compared to other terms, including the highest level of (clear) appropriation as well as high levels of negative stereotyping and non-derogatory uses. "Paki" also has both high levels of negative stereotyping and non-derogatory uses. The term was originally used as a derogative slur against immigrants from Southern Asia in the UK, however there appears to be a trend among younger Asians in reclaiming the term. For example:

> *@^^^ I'm a paki innit. Got the patter. Standard*

Despite the rising appropriation of the word it still has the power to anger and enrage and is used for negative stereotyping of Southern Asian communities, causing great offence. However there are also non-derogatory uses of the word, in tweets that refer to "paki shop". Where the term itself is considered a derogatory description (and we decided to categorise it as either negative stereotypical or lazy use of racial slur, depending on the case), the way in which the term is used in tweets were usually for the purpose of literally distinguishing between stores,

> *@^^^ paki shop opposite clock tower Â£20.*

There were also a significant number of non-relevant tweets across all terms – in particular, the use of "crow", which, despite accounting for 2.1 per cent of machine-classified relevant tweets, was found to be irrelevant in over 80 per cent of cases in the manual analysis, pointing to a failure of the machine relevance classifier to weed out the non-ethnic uses of the term.

The analysts also found 50 per cent of the occurrences of the term "coon" to be either impossible to categorise or deemed an error. This may reflect our ignorance of how the term is being employed by the

social groups responsible rather than a correct analysis of the usage pattern.

Discussion: comparing human analysts and automated classifiers

Direct comparison across the manual analysis method and the use of automated machine learning classifiers is difficult because, despite certain similarities in results, they are actually performing quite different tasks. Study 1 meaningfully categorised a very large data set of tweets in order to facilitate the analysis of the Twittersphere. Study 2 executed a detailed qualitative analysis of a very small data set to reveal more detailed insights about (some) language use on Twitter.

Both have advantages depending on the ultimate purpose of the study. Machine learning classification analysis is much quicker and less resource-intensive: study 2 required multiple person weeks just to produce the data (the iterative annotations). Machine analysis can also be more "reliable" in counteracting identified bias, because it can apply rules consistently. Certainly it is difficult for human analysts to overcome a bias so cleanly: when two analysts identified their respective under- and over-use of a category in study 2, they reversed their positions in the next iteration. In other cases, analysts had to agree to disagree on interpretations. (However, it is worth noting, that the machine learning classifiers are seeking to mimic decisions based on the training data it has received: and so are also liable to replicate any biases).

On the other hand, human analysis can pick out subtleties of language and nuances of meaning – with a reasonably high degree of reliability – that would be impossible for a machine classifier. It is of note, however, that despite these differences in approach, the headlines for the two studies were broadly similar: we see a high use of non-derogatory language, and a low use of directly threatening language. Reports that uncritically cite a high incidence of racial or ethnic slurs are in danger of providing a misleading impression of how people employ these terms in their everyday online conversations.

ANNEX

Slur terms used

The following table shows the list of ethnic slurs used to scrape Twitter. A tweet matched these criteria if it contained one or more of these letter streams.

Table 7: Slurs used for sampling

bohunk	boong	bounty bar	buddhahead
buffie	burrhead	ching chong	chink
chonky	coolie	coon	coonass
crow	cunt-eyed	cushi	dago
darkie	darky	dink	dogan
dune coon	gable	ghjji	gin jockey
gipp	golliwog	gook	gringo
groid	gubba	guido	gyppie
gyppo	gyppy	hairyback	half-breed
hambaya	hebe	heeb	house negro
house nigger	hymie	ikey	jap
jigaboo	jigarooni	jigga	jiggabo
jigger	kaffir	kala	kraut
kyke	limey	malaun	moulie
mussie	muzzie	muzzies	neche
nichi	nichiwa	nidge	nig-nog
niger	nigette	nigga	nigger

niggress	niglet	nip	nitchee
nitchie	nitchy	ocker	paki
pancake face	pickaninny	pikey	piky
polack	quashie	raghead	razakars
schvartse	seppo	sideways cooter	sideways pussy
slant-eye	slopehead	spearchucker	spic
spick	spig	spigotty	spik
squarehead	squaw	squinty	thicklips
towel head	whigger	white boy	whitey
wigga	wigger	wog	

Detailed classifier results

Tier 1: Filtering for Relevance

Our initial scraping and filtering gave a sample of 126,975 tweets each of which contained at least one of the ethnic slurs listed in attachment 1. The meaning of many of these terms is ambiguous, however, so we developed a classifier for the tweets in an attempt to identify only those tweets using the words consciously as racial terms in a manner that might be interpreted as prejudicial, abusive, or insulting.

Of the sample of 126,975 tweets, the relevance classifier identified 106,691 as relevant. We tested the accuracy of the classifier against a gold standard of 474 manually classified tweets. This analysis suggested that the classifier was capturing nearly all of the tweets that were actually relevant: 97 per cent of relevant tweets in the gold standard were identified as such (relevant recall = 0.97). The analysis suggests that the classifier is, however, also classifying as relevant 52 per cent of all irrelevant tweets in the gold standard (irrelevant recall =

0.48). This reduces the precision of the relevance category to 84 per cent (relevant precision = 0.84) i.e. we estimate 84 per cent of tweets in the relevant category really are relevant.

The gold-standard manual analysis identified 74 per cent of the sample tweets as relevant. The classifier's accuracy of 0.84 compares to a random classifier, which for two-way classification would have had an accuracy of 0.5, and a classifier that assigned all tweets to the most common category (relevant), which would have had an accuracy of 0.74.

Adjusting for these classifier characteristics by calculating the inferred number of misclassifications from the gold-standard analysis, and also by direct estimation from the split of the gold-standard data (with which the analysis closely agreed), we estimate that there were 93,894 true relevant tweets in the sample. If we assume that there are 400 million tweets per day on average, of which 39 per cent are in the English language, then this implies that a racial slur from the list is used in approximately 1 in 15,000 tweets (0.0067 per cent of tweets).

Table 8: Evaluation of relevancy classifier

Evaluation	Relevant	Irrelevant	Overall
F1	0.90	0.61	
Precision	0.84	0.83	
Recall	0.97	0.48	
Accuracy			0.84

Naturally, the figures for the rate at which slurs are employed will vary significantly with inclusion or exclusion of certain terms from such a list. The distribution of words (terms) in messages follows approximately a 'Zipfian' or power-law relationship: a small number of terms will typically be responsible for a large proportion of a sample gathered in this manner. The most common term "white boy" occurs

in almost 50 per cent of relevant tweets. The top 5 slurs (“white boy”, “paki”, “whitey”, “pikey”, and “coon”) account for over 75 per cent of relevant tweets. Figure 5 provides examples of relevant (green) and irrelevant (red) tweets at this stage of analysis.

Table 9: Examples of relevant and irrelevant tweets at this stage of the analysis

@^^^ Aye!! Lmao. You my nigga too white boy! We gone have to hang out one time for the Hood.	Relevant Irrelevant X
RT @^^^: “@^^^: @^^^: ^^^ is a wigga who sucks ass at rapping. #^^^ true but debra goes ham” @^^^	Relevant Irrelevant X
@^^^ dis white boy Clarke is stupid sick with black nigga handles.	Relevant Irrelevant X
@^^^ @^^^ @^^^ what the hell is white boy swag	Relevant Irrelevant X
White boy can play	Relevant Irrelevant X
The after effect of being a wigger. http^^^	Relevant Irrelevant X
Yes my last three tweets are a white boy rap	Relevant Irrelevant X
RT @^^^ “@^^^: Im finding me a white boy.” YES!!!!! Me too Im too done with black love, onto interracial love	Relevant Irrelevant X
@^^^ i told yo racist ass to stop callin me a nigga, white boy ! But okay :p	

Relevant Irrelevant X

@^^^ here you ya towel head cunt!

Relevant Irrelevant X

@^^^ did that pikey give em back!

Relevant Irrelevant X

“@^^^: Everyone go to the #WOG game tonight!”

Relevant Irrelevant X

I meant Captain Charlie Chunk not Chink lol!

Relevant Irrelevant X

I have a good friend from Paki and I am happy!

Relevant Irrelevant X

@^^^ but I call the crows nest!!! #causeimacoolkid

Relevant Irrelevant X

So I can see iphone emojis now, soooo coolie!!!!

Relevant Irrelevant X

@^^^ ur teeth is spick and span!!

Relevant Irrelevant X

Teleton Hebe no SBT, acho chic!! #Teleton

Relevant Irrelevant X

everyones like I want a bf I want a gf and Im like I just want a new car and some old gringo boots.. and vodka

Relevant Irrelevant X

Tier 2: Racial Slurs as Markers of Group Identity

Looking at the data, we believe we can pick out a number of different patterns of usage. A large proportion of the tweets appear to be using these terms in a racially aware fashion, but for the purpose of demarcating group boundaries; the use of the term is otherwise basically non-prejudicial in the context of the tweet.

This observation closely ties in with recent publicised work that found the use of certain distinctive terms in social media (including “nigga”) as characteristic of certain “tribes” or social groupings. In other words, social groups use language in certain distinctive ways to define themselves, and in certain groups this includes the use of racial terms. Our working hypothesis is that these social groupings (and their source tweets) will usually be geographically isolatable (recalling that Twitter is a global platform). This is because online expressions reflect distinct real-world social communities.

We developed a classifier for the tweets that splits the data into three groups: (i) ‘racial’ tweets, slur terms used in a manner that could be interpreted as racially prejudiced; (ii) ‘group identification’, slur terms used to demarcate group boundaries, but the use of the term is otherwise basically non-prejudicial in the context of the tweet; and (iii) ‘irrelevant’, tweets that have (incorrectly) passed through previous filters.

106,755 tweets identified as relevant in Tier 1 were taken as the input sample. Of these, 22 per cent were classified as ‘racial’, 56 per cent as ‘group identification’, and 23 per cent as ‘irrelevant’. Excluding the tweets classed as irrelevant, the estimated split on 87,923 tweets was 28 per cent and 72 per cent for racial and group identification categories, respectively. We also estimated the category splits manually by taking a random sample of 500 tweets from the input sample, which gave similar split estimates for the data (25 per cent, 52 per cent, and 23 per cent). Taking these as independent estimates, we arrive at prevalence estimates of 1 in 56,000 for racial tweets (2,800 tweets per day).

These estimates imply that ethnic slurs are used as markers of group identity over twice as often than they are used as (apparently) prejudicial terms. Initial analysis suggests that “white boy” (the most

common term), “nigga”, and “wigga” are used particularly frequently in this way.

Table 9 summarises the evaluation of the classifier. Classifier accuracy has dropped (overall accuracy = 0.67). For three-way classification, a random classifier would have had an accuracy of 0.33 and a classifier that assigned all tweets to the most common category (GroupID) would have had an accuracy of about 0.5. Figure 7 provides examples of tweets categorized as racial and GroupID.

If we accept the classification of very many of these tweets as reflecting or signalling racially-aware group identity rather than racist belief, then we are left with a much smaller kernel of racial messages that appear to be using such terms in a racially prejudiced manner.

Table 10: Evaluation of group identification classifier

Evaluation	Racial	GroupID	Irrelevant	Overall
F1	0.53	0.77	0.59	
Precision	0.55	0.75	0.60	
Recall	0.51	0.79	0.57	
Accuracy				0.67

Table 11: Examples of racial (green) and groupID (red) tweets

White boy in auto parts just said I get purp by the pound ??? I laughed so hard

Racist GroupID Irrelevant X

RT @^^^: But why did that white boy drop 138 points.. ?

Racist GroupID Irrelevant X

RT @^^^: Whitey had a feast thanking the Native Americans for teaching them how to survive on American land, then they kill em off & steal that land.

Racist GroupID Irrelevant X

@^^^ @^^^ @^^^ I agree with Steph and Tom, Iceland is like ultimate pikey!

Racist GroupID Irrelevant X

@^^^ pretty sure theres a more obvious one on the same show? Silly pikey, probably bad signal in her caravan.

Racist GroupID Irrelevant X

@^^^ I KNOW IM LIKE WHAT?! AND I SEARCHED SLOTHMAS AND NOTHING WAS THERE AND IM LIKE WTF R U TALKIN BOUT WIGGA

Racist GroupID Irrelevant X

@^^^ white boy got soul http://^^^

Racist GroupID Irrelevant X

Will sheehey will always be that white boy

Racist GroupID Irrelevant X

Normal white pussy is much, much better than twitter white wigger bitch pussy.

Racist GroupID Irrelevant X

LOOL fuckin hell. thats what its like every Eid you know I swear. Police jeans, Carlotti sweatshirt and some Voi pumps for these paki boys

Racist GroupID Irrelevant X

Rick Ross a coon nigga said he got that Justin Bieber please believe me -_- FOH

Racist GroupID Irrelevant X

@^^^ Youre Paki Suicide Bombers, so we sung Well be running round Tottenham with our willies hanging out. Awful banter.

Racist GroupID Irrelevant X

Mums acts like such a fucking kid not going if your dads going oh fuck off you raghead

Racist GroupID Irrelevant X

@^^^ yeah was even worse coz he was drenched. Some horrible muzzies out there...

Racist GroupID Irrelevant X

FUCK ALL THIS SHIT,AND ESPECAILLY SPICK MUSIC!!!! FUCK EVERYONE AND EVERYHING

Racist GroupID Irrelevant X

@^^^ @^^^ stop being such a gook and enjoy the game with us nigger

Racist GroupID Irrelevant X

Tier 3: "Casual" use of racial slurs

The racially-prejudiced tweets appear not to be uniform in nature; we suggest that a proportion of such tweets might be described as showing a 'casual use' of slurs. They contain phraseology that might be deemed insulting, abusive, or threatening, and they use terms in a prejudicial fashion, but the terms are used in an off-hand or casual manner. The message is typically not particularly heated. It suggests underlying racism, but conveyance of that racist viewpoint is often not the point of the message.

A working hypothesis is that the language used in tweets with such 'casual use' slurs – and the sentiments expressed – reflect social norms within the sender's personal community. Notwithstanding the absolutes of legislation, these social norms are negotiable, contestable, and contested. For example, in British society it is arguably currently socially acceptable to use the term "chav" in a prejudicial (and typically insulting) context even though such a term arguably refers to a distinct and identifiable ethnic sub-group of the wider population. National television programmes such as "The Only Way is Essex" are arguably constructed around feeding that prejudice. Other prejudicial terms were historically deemed acceptable, and are no longer seen as

acceptable by society-at-large, but arguably remain acceptable to significant sub-communities within society.

The movement over time of what constitutes “acceptable” and “unacceptable” expressions of prejudice is of course well known and reflected not least in changes in legislation over time. Social media is perhaps just laying bare the fact that such boundaries are not uniform across societies, not only on the nation state level, but also within communities. We have all always been aware of such differences, but social media displays these differences with greater prominence. This observation raises interesting questions regarding the degree to which “authorities” in a society might wish to impose or encourage more uniformity (of expression or perhaps even belief) – a deeply contestable notion that lies at the heart of much discourse in this arena today.

Tweets showing ‘casual use’ often make use of ethnic slurs in an adjectival form (“paki shop”, “pikey clothes”). Some expressions of this form have crossed over into standard urban slang and appear to have lost explicit racist connotations (e.g. ‘white boy wasted’), while for others their status is contested. Figure 8 provides an evaluation of the classifier, and figure 9 some examples of tweets assigned to the inflammatory and casual use of slurs categories, respectively. It is notable that the classifier does not very reliably distinguish between inflammatory and casual use tweets. For human annotators, the categorical distinction is also labile and it is hard to gain agreement. Manual classification of a random sample of 495 tweets suggested that (excluding any irrelevant tweets) 28 per cent of racial tweets demonstrated the ‘casual use of slurs’ and 72 per cent were more directly inflammatory.

If we accept this classification tweets with ‘casual use of slurs’, there remain a cohort of messages that appear not only racially-prejudiced but also contain a strongly invective element.

Table 11: Evaluation of Inflammatory classifier

Evaluation	Inflammatory	Casual	Irrelevant	Overall
F1	0.56	0.26	0.61	
Precision	0.57	0.30	0.58	
Recall	0.54	0.23	0.65	
Accuracy				0.54

Table 12: Examples of inflammatory (green) and casual use (red) tweets

@^^^ youre a damn coon so shut yo trap	Inflammatory Casual Irrelevant X
West Ham are pikeys.	Inflammatory Casual Irrelevant X
@^^^: @^^^ lol u got a shit day then and the worst thing is ur with pikey to pikeys cool!	Inflammatory Casual Irrelevant X
He dope RT @^^^: <---still a John Mayer fan, even if he dont like black women. White boy can write his ass off.	Inflammatory Casual Irrelevant X
RT @^^^: Call me a fucking indian its okay, call me a dirty paki ill muay thai kick you on your head soo bad your mother will feel it. #NotAPaki	Inflammatory Casual Irrelevant X
Where do you shop most often? — charity shops cos im pikey xxx http://^^^	

Inflammatory Casual Irrelevant X

“@^^^: Paki man made me break my bb charger look what I’ve resorted 2 #dedication lol! http://^^^

Inflammatory Casual Irrelevant X

First tym in my lyf I was called a fukin paki n tld 2 go back 2 my country. People these days

Inflammatory Casual Irrelevant X

@^^^ mate do you take off your turban everynight after a long dqy of being a long haired terrorist cunt? #alladine #paki

Inflammatory Casual Irrelevant X

Stupd fuckin chink bitchfucked up my test cock suckin bitch gook FUCK YOU SWAMP RAT

Inflammatory Casual Irrelevant X

Its this #Mangu that you tasting, got chu crazy for the licks, my ride or die know how to chop it, dumbest loudest fuckin spik

Inflammatory Casual Irrelevant X

I hate getting on a busy bus and the person sitting directly in front of me stinks of BO #haveawash #stinks #pikey

Inflammatory Casual Irrelevant X

New all time low.. Stealing cardboard boxes from the SU ready for initiation #pikey #goodbyedignity @^^^ http://^^^

Inflammatory Casual Irrelevant X

@^^^ @^^^ Fuck that shovin shit, specially from whitey

Inflammatory Casual Irrelevant X

I liked a @YouTube video http://t^^^ Towel Head Roommate Freestyle Diss

Inflammatory Casual Irrelevant X

Just heard some guido at the tanning salon refer to himself as the kid @^^^ #notcool

Inflammatory Casual Irrelevant X

Tier 4: Inflammatory racist remarks: insults and ideology

We judged the remaining tweets in our sample as insulting, abusive, or threatening (based on considerations of race or ethnicity), and written in a manner apparently intended to give offense.

By our standards, this is unpleasant speech. We tentatively describe a tweet of this kind as an 'inflammatory racist remark' which would meet a "common sense" criterion of being racially hateful. But is it anything else? Are there any contextual circumstances that help explain the utterance and mitigate their gravity? Are such utterances of sufficient gravity that they might meet legal definitions of speech that is unlawful?

Microblogs such as Twitter are a novel form of communication. Whilst they are globally available public published works, they are typically composed in a minute or so, in the "heat of the moment" so-to-speak. They often therefore reflect an individual's visceral reactions to a situation rather than a considered published commentary. Twitter is an open form – anybody can tweet, regardless of age, qualification, or level of awareness of the niceties of legislation covering freedom of expression across multiple legal jurisdictions. In addition, analysis of usage suggests that many people use Twitter for one-to-one (or "one-to-very few") communication, notwithstanding the fact that the communication is technically viewable by all. In this respect, Twitter usage in part echoes the days of telephone "party lines".

One upshot of all this novelty is that many essentially private conversations or arguments are enacted over Twitter. Our working hypothesis is that many strongly invective messages containing racially insulting phraseology are "snapshots" from essentially private arguments. Whilst this is not to condone such language, this distinction between, on the one hand, public discourse and, on the other hand, publically apprehended but essentially private discourse, has historically been recognised in legislation (for the UK, see e.g. Part I, Section 4A of the Public Order Act (1986) for one attempt to address this distinction).

Forming a judgment over whether a message reflects an insult in the heat of a personal argument or a more "considered" racist attack –

perhaps intended for wide public apprehension - requires more context than is typically available from a single tweet. Microblogs such as Twitter can potentially provide such context; it is often possible, for example, to reconstruct from meta-data surrounding tweets what constituted the broader conversation. Tweets can also be traced back to the accounts of senders and recipients, providing another avenue for discovering context. These approaches can raise ethical issues, however, and are beyond the scope of our current work.

If we accept this classification of "personal attack" tweets (that might or might not reflect an essentially private argument overheard in a public context) then there may remain a small number (number as yet un-quantified) of strongly invective, racially-prejudiced tweets that are directed at a community as a whole.

Estimate of Prevalence

This large-scale quantitative analysis provides an opportunity to estimate the prevalence of these patterns of usage. Prevalence depends on the length of the list of racial / ethnic slurs that are at the search's core. Given the list used here at the beginning of this annex, we found a prevalence of relevant tweets of roughly 1 in 15,000 tweets (in the region of 10,000 tweets per day). Note that this figure halves if we remove the most common and contested term, "white boy".

Of these relevant tweets, we estimate here that over 70 per cent can be classed as using these words as markers of group identity, as opposed to indicating racial / ethnic prejudice *per se*. In other words, of the 10,000 tweets employing racial / ethnic slurs every day, 7,000 are employing them in a non-derogatory fashion.

As a result, we estimate a prevalence of about 1 in 55,000 tweets in the English language that are indicative of racial/ethnic prejudice on the part of the sender (in the region of 3,000 tweets per day).

Of these apparently racially or ethnically prejudiced tweets, manual classification of 500 tweets sampled randomly from this group suggested that around 30 per cent show casual use of slur terms (as

defined above), with the balance of tweets making comments that are more directly racially or ethnically prejudicial.

This suggests a prevalence of directed racially or ethnically prejudicial tweets of about 1 in 75,000 tweets in the English language (in the region of 2,000 tweets per day). More context would be required to determine the precise pattern of usage for these tweets e.g. humour between friends, a private argument being played out in a public space, response to goading, and so forth. In many cases, single conservations may lead to several tweets of this nature back and forth, so these figures may reflect a few hundreds of conversations worldwide each day.

NOTES

[i] Bartlett, J (2013, forthcoming) *Hate speech in Europe*, Demos

[ii] Wikipedia List of ethnic slurs http://en.wikipedia.org/wiki/List_of_ethnic_slurs. Note that white boy no longer appears in this list.

[iii] Anderson, L & Lefore, E (2013) 'Slurring Words' *Nous 47:1* p. *41*

[iv] This is known as 'contextualism' and was set out in Kennedy, R (2002) *Nigger: The strange careers of a troublesome word*, p.54

[v] Bastiaan Vanacker, *Global Medium - Local Laws*, El Paso: LFB Scholarly Pub., 2009; Warner ,W, Hirschberg, J, 2012, Detecting hate speech on the world wide web , Columbia university, Workshop on language in social media, 19-26 http://aclweb.org/anthology-new/W/W12/W12-2103.pdf

[vi] Bartlett, J (2013, forthcoming) *Hate speech in Europe*, Demos

[vii] INACH, Intervention by Ronald Eissens on behalf of the International Network against Cyberhate', 4 October 2011, www.osce.org/odihr/83488

[viii] www.legislation.gov.uk/ukpga/2003/21/section/127

[ix] 'Tweeter jailed for disgusting racist posts about Fabrice Muamba', South Wales Evening Post, 28 March 2012, www.thisissouthwales.co.uk/Tweeter-jailed-disgusting-racist-posts-Fabrice/story-15644497-detail/story.html (accessed 20 September 2012); http://www.legislation.gov.uk/ukpga/1998/37/contents

[x] http://tellmamauk.org/

[xi] A Roversi, *Hate on the Net – Extremist Sites, Neo-fascism On-line, Electronic Jihad*, Ashgate Publishing Limited: Aldershot, 2008; P M Meddaugh, J Kay, *Hate speech or "Reasonable Racism?" The Other in Stormfront*, Journal of Mass Media Ethics: Exploring Questions of Media Morality, Vol 24, Issue 4, 2009; http://www.journalismethics.info/JMMEper cent20-per cent20Hateper cent20Speechper cent20orper cent20Reasonableper cent20Racism.pdf; Adam G. Klein, *A Space For Hate: The White Power Movement's Adaption into Cyberspace*, Litwin Books LLC, 2010.

[xii] Hom 'The semantics of racial epitephs' Hom, (2010) 'Pejoratives', *Philosophy Compass 5/2*

[xiii] Croom, A (2013) 'How to do things with slurs: studies in the way of derogatory word', *Language & Communication*, p49

[xiv] Croom, A (2013) 'How to do things with slurs: studies in the way of derogatory word', *Language & Communication*

[xv] Croom, S (2010) 'Slurs'

[xvi] See Hornsby, J (2001) for a semantic approach; and Hom (2012) 'The Semantics of Racial Epiteths' for a good overview of both approaches.

[xvii] (Stephens-Davidowitz, 2011) & Croom, S (2010) 'Slurs'; Anderson, L & Lefore, E (2013) 'Slurring Words' *Nous 47:1* p. *39*

[xviii] Bartlett, et al (2013, forthcoming) Vox Digitas

[xix] Floatingsheep (2013) *Geography of Hate*

[xx] See Bartlett et al (2013, forthcoming) Vox Digital

[xxi] Wikipedia List of ethnic slurs http://en.wikipedia.org/wiki/List_of_ethnic_slurs. Note that white boy no longer appears in this list.

[xxii] AAF allows the researcher to scrape tweets from the Twitter live stream using the standard Boolean search terms provided for in the Twitter Filter Application Programming Interface (API). (The API is a portal that acts as a technical gatekeeper of the data held by the social media platform).

[xxiii] This calculation is based on 10,000 relevant tweets per day, multiplied by the percentage distribution in the previous column (% of tweets).

[xxiv] This was calculated as follows. First, the average of the two analysts' percentages was calculated for each data set and category. Whichever of these was highest was multiplied by 10,000 (the total number of relevant tweets per day according to study 1) to give the maximum value in the range; whichever of these was the lowest was multiplied by 10,000 to give the minimum value. *The maximum and minimum were then rounded to the nearest 500.*

[xxv] https://my.barackobama.com/page/share/notape, accessed 23rd July 2013.

[xxvi] "Black Panther leader who has the words 'Kill Whitey' tattooed on his FACE is arrested after cops catch him carrying an unlicensed loaded weapon on the streets of New York and wearing a bullet proof vest", *Daily Mail*, 20 June 2013, accessed 23 July 2013 at http://www.dailymail.co.uk/news/article-2346713/Leader-Black-Panthers-sports-tattoo-reading-Kill-Whitey-cheek-arrested-carrying-unlicensed-loaded-weapon-wearing-bullet-proof-vest.html

[xxvii] One – admittedly unreliable – online poll about the term on "Sodahead.com" asked participants to vote for "Yes, it is racist and should not be used!" (50 per cent), "No, it is just a word and is not racist." (23 per cent), and "Undecided" (27 per cent. Sodahead forum discussion, http://www.sodahead.com/united-states/do-you-find-the-word-whitey-a-racist-word-to-use-towards-caucasians/question-101650/, accessed 23 July 2013