THE USE OF MISOGYNISTIC TERMS ON TWITTER

Overview

To coincide with the *Recl@im the Internet* cross-party campaign, at the Centre for the Analysis of Social Media (a collaboration between Demos and the University of Sussex TAG laboratory) researchers are conducting an ongoing investigation into the use of misogynistic terms on social media.[i]

As part of this investigation, produced to support the launch of the campaign, we conducted a small scale study examining the use of two popularly used misogynistic terms ('slut' and 'whore') on the social media platform Twitter. The results were presented at House of Commons launch of *Recl@im* on 26 May 2016.

The objective was to provide a general overview of the volume and nature of how these two terms are being used. Specifically:

- How many times a day are these gendered, misogynistic terms used on Twitter?
- How many of these uses might be classed as aggressive?
- Are there other classes of use?
- How many users does this represent?
- How many of these users can be located to the UK?
- How many people are mentioned in tweets containing either one?
- What gender were the users?

The study is strictly limited in scope. It does not claim to be a comprehensive analysis of all misogynistic words being used on Twitter; the two terms analysed, 'slut' and 'whore', represent a small fraction of misogynistic terms being used on Twitter. Further, misogynistic language is deployed on Twitter in a wide variety of additional ways that are not investigated here. Rather, the work is intended to stimulate further research and raise questions for further research in this area as part of our investigation.

This paper presents the result and methods of this study.

Methodology

Collection

Between 23rd April and 15th May 2016, tweets containing 'slut' and/or 'whore' were collected through Twitter's Stream Application Programming Interface (API) using a text-analysis software platform (Method52).

Twitter's Stream API allows researchers to collect all instances in real time whenever the term is used by a user with a public Twitter account. These two terms were chosen based on findings from our 2014 study 'Misogyny on Twitter', which had identified both as the most frequently-used, gendered insults used on Twitter.[ii]

In total, we collected 1.46 million tweets from around the world over the 23-day period.

Analysis

A collection of 1.46 million tweets covers a range of patterns of use (i.e. themes or topics). We sought to gain a more coherent picture of the collection by breaking it down into a small set of broad categories that reflected these different use patterns, and we used Method52's classification tools to assist in this task.

This was an iterative process. First an analyst manually reviewed a random sample of the dataset and identified the single most prominent pattern of use (category). In this instance, we found that a significant proportion of the tweets were referencing or linking to pornography. The analyst then trained a machine-learning classifier to recognise patterns in language, using example-driven training data that reflected this pattern of use. This bespoke classifier was then applied to the entire data set, splitting away all those tweets that matched the classifier's criteria for that category. This left a set of tweets with the pornographic content largely removed.

Over three iterations, using this methodology each time, the analyst identified the following broad categories, and repeated the process:

1) Pornographic content: A very large proportion of tweets were links to pornography sites. These would typically include additional sexualised terms, and a link. *Example:* '*Busty slut and keana moire licking each other https://t.co/…'.*

2) 'Aggressive': A significant volume of tweets appeared to be overtly aggressive in language used. In particular, we looked for three attributes. First, use of additional expletives (*'What a fucking bitch slut !!')*; second, commands such as 'get out' or 'shut up' (*'stfu slut')*; and third, tweets aimed at 'you', the recipient (*'you are a fucking whore')*. Example: *'@xxx corporate whore', '@xxx bitch when you dying? satan is waiting. bye pack your bags ugly white whore'.*

3) 'Self-identity': A smaller proportion appeared to use the terms as a means of self-identification, re-appropriation or group identity. Example: '*I'm a slut for beautiful sunsets and the stars*' or '*happy birthday little slut I guess I love you*').

4) Other: all instances of tweets which did not obviously fit into one of the categories above was classified as 'other'. This included cases of people sharing news articles where the term was quotes, and discussion of 'slut shaming'.

The machine-learning classifiers support (but do not replace) the analyst's work in assessing the document set. Using classifiers this way offers two broad benefits. First, by peeling away tweets that match an identified category, the structure of the rest of the data is progressively revealed to the analyst - highlighting less frequent but important sub-strata of categories. Second, the iterative analysis itself creates a 'cascade' of classifiers which can be used to help the analyst divide up the documents into the categories that they have selected (see diagram in results section, below).

There are many ways in which any given data set can be divided up using this approach. These choices are made by the analyst, based on their assessment at each iteration of a random sample of the underlying data. This same data set could be used, for instance, to examine in more detail patterns within the sub-set of tweets which discussed women's rights movements such as 'slut shaming'.

Construction of classifiers

Classifiers are built by training an algorithm to spot patterns in the language used in the body of the tweet by providing examples. An analyst 'marks up' which category he or she considers a tweet to fall into, and this 'teaches' the algorithm to spot patterns in the language use associated with each category chosen. The algorithm looks for statistical correlations between the language used and the categories assigned to determine the extent to which words and bigrams are indicative of the pre-agreed categories. (For further reading on these methods, see the methodology annex (p.85) in *Vox Digitas (2014)*[iii]).

For this study, an analyst built three classifiers, each one making a binary decision. The first classifier was whether a tweet was 'pornographic' or not. This classifier was coded on 362 tweets. Once pornographic tweets were removed from the dataset, a second classifier was constructed using the remaining data. This was whether a tweet was 'aggressive' or not. This classifier was coded on 822 tweets. Finally, a third classifier was constructed using the remaining data (i.e., not including those considered to be 'aggressive') to determine if a tweet was 'self-identification' or not. This was coded on 271 tweets.

The reason there are different numbers of tweets coded for each classifier is that the analyst continued coding until the classifiers performed at what we judged to be a satisfactory level of accuracy, as discussed below.

Accuracy of classifiers

In order to estimate the accuracy of these algorithms at classifying data into the chosen categories we used a 'gold standard' approach. 100 tweets were randomly selected from the relevant dataset to form a gold standard test set for each classifier (a total of 300). These were manually coded into the categories defined above. These tweets were then removed from the main dataset and so were not used to train the classifier.

As the analyst trained the classifier, the software reported back on how accurate the classifier was at categorising the gold standard, as compared to the analyst's decisions. On the basis of this comparison, classifier performance statistics – 'recall', 'precision', and 'overall' are created and appraised by a human analyst. Each measures the ability of the classifier to make the same decisions as a human in a different way:

Recall: The number of correct selections that the classifier makes as a proportion of the total correct selections it could have made. If there were 10 relevant tweets in a dataset, and a relevancy classifier successfully picks 8 of them, it has a recall score of 80 per cent.

Precision: This is the number of correct selections the classifiers makes as a proportion of all the selections it has made. If a relevancy classifier selects 10 tweets as relevant, and 8 of them actually are indeed relevant, it has a precision score of 80 per cent.

Overall: All classifiers are a trade-off between recall and precision. Classifiers with a high recall score tend to be less precise, and vice versa. The 'overall' score reconciles precision and recall to create one, overall measurement of performance for the classifier.

(Note: no algorithms work perfectly, and a vital new coalface in this kind of research is to understand how well any given algorithm performs, and the implications of this performance for the research results.)

The overall accuracies of each classifier is given below.

Classifier 1: Pornography

	Precision	Recall	F-Score	Accuracy
Pornography	0.897	0.929	0.912	
Other	0.905	0.864	0.884	
				0.9

Classifier 2: Aggressive Tweets

	Precision	Recall	F-Score	Accuracy
Aggressive	0.655	0.704	0.679	
Other	0.887	0.863	0.875	
				0.82

Classifier 3: Self-Identification

	Precision	Recall	F-Score	Accuracy
Self-Identification	0.757	0.824	0.789	
Other	0.905	0.864	0.884	
				0.85

These figures are only indicative, however, as there are two important caveats. First, because these accuracy scores were used in part to guide when to stop training (see above), a more precise estimate would require testing against a new unseen gold standard test set. Second, the 'accuracy' or otherwise of the classifier tells us nothing about the validity or appropriateness of the analyst's choice of categories into which they wanted the data to be split, other than to say that the classifier was able to find features in the data that correlated with that categorisation scheme.

Gender annotator

In order to estimate the gender of people posting tweets, we used a pre-existing standard algorithm which is incorporated in Method52. Using a forced-choice approach, it classifies each tweet into one of three categories: 'male', 'female' or 'institution', based on information in the user name and user description fields. When tested in 2015 against a sample of 2,500 users, whose gender was known via traditional survey questioning, this algorithm had an accuracy of approximately 85 per cent. In order to re-test the accuracy of this algorithm on our data set, an analyst took a random sample of 250 users who had contributed to the dataset, and manually marked them up as ‘male’, female’ or ‘other/unknown’. (See results section, below).

Measurement of user volume

We measured the volume of unique user names in the body of a tweet, to estimate how many users received these messages. This was done by taking the ‘mentions’ metadata available from the Twitter API, then ensuring that no retweet data was contained within that tweet (as mentions also includes the original screenname if a tweet is retweeted).

Estimation of location

We used a location annotator to identify the location from where a tweet was sent. We used a pre-existing standard algorithm which is incorporated in Method52. The intention was to identify the country of origin of the sender of the tweet (or in the small proportion of cases with GPS data, the location from which the tweet was sent). The annotator uses an evaluation cascade that looks for GPS coordinates, recognisable descriptions of locations, and time zone descriptors in the tweet metadata. Using this method, we were able to locate 67 per cent of users to a country level.

Results

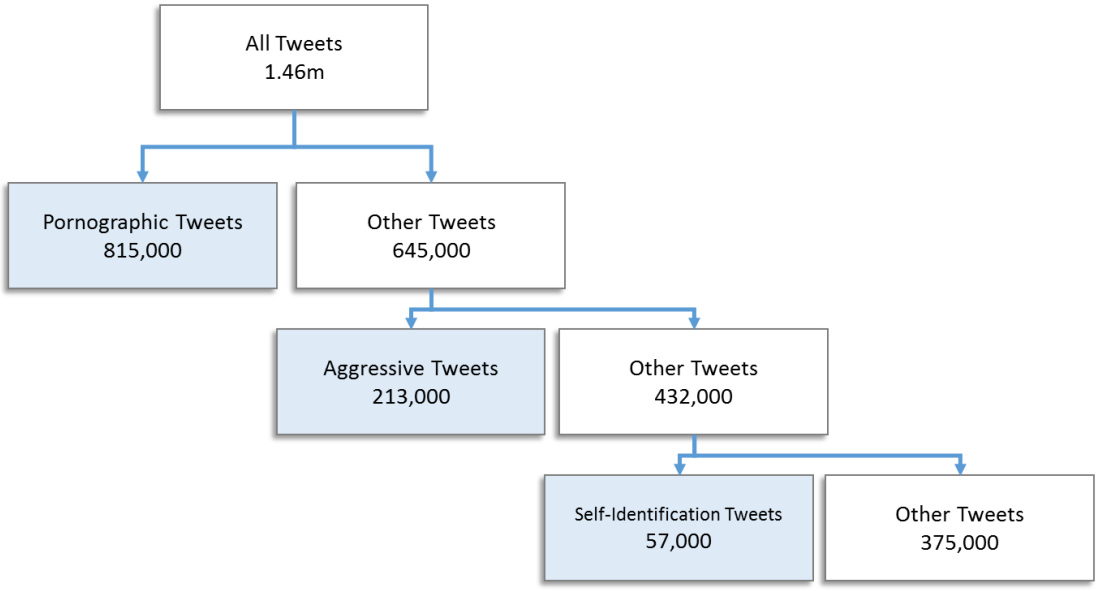
Volumes

We estimated total volumes in each category by applying our completed classifier cascade (see diagram, below) to the overall data set. In total, using the classifiers above, we estimated that pornographic tweets (i.e. advertising explicitly pornographic content, typically linking to pornographic websites) accounted for approximately 56 per cent of the total data set (all tweets containing either of the words). Once these were removed it reduced the dataset from 1.46 million tweets to 645,000.

Of the remaining tweets, we estimated a figure of 213,000 for non-pornographic tweets classified as aggressive. This is 15 per cent of all tweets containing either of the words, and 33 per cent of non-pornographic tweets containing either of the words. Once these were removed it reduced the dataset from 645,000 to 432,000.

Of the remaining tweets (neither pornographic nor aggressive), we estimated that 57,000 tweets could be classified as self-identification: 4 per cent of the total dataset, 9 per cent of the non-pornographic tweets, and 13 per cent of the neither pornographic nor aggressive tweets. The balance of tweets (375,000) fall into none of these categories, are classified as 'other', and are not further analysed in this study.

Image: classifiers in a 'cascade'



When averaged out over the time period, these estimates correspond to over 9,000 aggressively misogynistic tweets sent per day worldwide over the time period, with 80,000 Twitter users receiving those messages (i.e. whether there was a user name included in the body of the tweet). However, we did not try to determine the gender of the individuals who received the messages.

When applying the location algorithm, we estimated that 10,500 aggressive tweets (18 per cent of aggressive tweets), targeted at 6,500 unique users, could be located to tweet senders based in the UK.

We know that not all of the tweets in a class should be in that class (precision < 1) and we know that some tweets were missed from the class (recall < 1). A more sophisticated approach could take into account the imputed precision and recall characteristics of the classifiers (see table above) to adjust the estimates to reflect that. Making these adjustments would change the estimates for each of these categories very slightly: pornographic tweets: 790,000 (54% of total dataset); aggressive tweets: 210,000 (14 per cent of total, 31 per cent of non-pornographic tweets); self-identification tweets: 55,000 tweets (4 per cent of total, 8 per cent of non-pornographic tweets); and other tweets: 380,000 (28 per cent of total, 61 per cent of non-pornographic tweets).

The estimates do rely, however, on the classifiers performing appropriately and we undertook a set of manual analyses to cross check the performance of the 'aggressive' classifier. For the aggressive tweets, a random sample of 150 tweets classified as 'aggressive' were manually reviewed by an analyst. 122 tweets (81 per cent) were also classified manually as aggressive, while 28 tweets (19 per cent) did not appear to be aggressive, in line with the previously estimated overall accuracy of the classifier (see above). A qualitative look at the 28 misclassified tweets suggests the algorithm struggled most on tweets that were extremely offensive descriptions of a user, but may not have been directed at that individual (e.g. *'@xxx if your in maryland look this filthy fucking slut up)* or were highly internally contradictory (e.g. *'@kygirl2675 you're a whore because you took pictures in a bathing suit. nope.'*).

It ought to be noted that even the most apparently aggressive tweet may not actually be aggressive when taken out of context (and vice versa). Judging context for each tweet is not possible when dealing with data of this magnitude. To explore this, the methodology used here could be readily extended by sampling the set of 'aggressive' tweets and exploring the conversational context in which the sample tweets were delivered. Machine learning classifiers as used here will be a useful tool to manage and process the large data sets available, but detailed qualitative research would be necessary to fully understand how these terms are applied.

Gender of posters

We investigated the gender of the posters of the 213,000 'aggressive' tweets, using both an automated analysis on the entire dataset and a manual analysis on a small random sample.

For the automated analysis we used the gender annotator described above and applied it to all 213,000 tweets classed as 'aggressive'. Of users in our dataset who had sent a tweet classed as 'aggressive', 48 per cent were classified using the gender annotator as 'female' and 42 per cent were classified as 'male'. The remaining 10 per cent were classed as 'institution'. Therefore, of those classified as human users, 53 per cent were classified by the gender annotator as female and 47 per cent were classified as male.

For the manual analysis, we took a random sample of 250 users who had contributed to the 'aggressive' dataset, and an analyst estimated gender based on a review of tweets and media associated with each account. The users were split into three categories: male, female and

'institution/unknown'. The latter category was actually a combination of institutions (including bots) and accounts for which gender couldn't be determined.

The 250 accounts were marked up by a human analyst as below.

Gender	Count	%	% Female/Male
Female	118	47.2	55.7
Male	94	37.6	44.3
Institution/unknown	38	15.2	

As a cross-check, we also applied the gender annotator to these 250 accounts. These were marked up by the gender annotator as below.

Gender	Count	%	% Female/Male
Female	110	44.0	49.5
Male	112	44.8	50.5
Institution	28	11.2	

This suggests that the algorithm is unlikely to be over-estimating female participation. For this sample, the algorithm slightly over-estimated the relative proportion of men relative to women, but the difference is small.

Bots

Manual analysis offered the opportunity for a cursory review of automated Twitter accounts or 'bots'. Of the 250 accounts analysed, 19 were found to be bots (7.6 per cent). The algorithm – which is forced to make a choice between either 'male', 'female' or institution had classified eight of these accounts as male, eight as female and three as institution. A refinement of this study would be to attempt to remove bot accounts from the sample before gender classification.

Discussion

Overall, the research suggests that there is a significant volume of tweets where the terms 'slut' and 'whore' are used, and that they are used in a variety of different ways. Interestingly the majority of uses were not found to be aggressive, but rather self-identifying or 'other', such as in discussions of 'slut walks' or 'slut shaming' ('*Y'all preach feminism but at the minute some ones girlfriend does something "problematic" you start slut shamming her #RespectDanielle*').

Data collected over relatively short time periods can be driven by 'surges' or 'spikes' in traffic relating to specific news stories or incidents. We noticed a number of the users most frequently mentioned in association with misogynistic language were celebrities who had recently been the subject of controversy. For example, the US rapper Azealia Banks, who was accused of aiming racist taunts at other celebrities, found herself the subject of tweets insulting her (in defence of Zayn Malik, one of the celebrities she was accused of insulting). The same happened when Beyoncé fans believed they'd identified a woman with whom her husband Jay-Z had had an affair. Both resulted in a significant number of women using the terms slut and whore referring to those individuals.

It is important to note that the categories selected were chosen by an analyst, based on a manual review of the data. This does not mean this was the only (or indeed the best) way to analyse this data set. The very large presence of tweets determined to be 'other' suggests there are many ways it would be possibly to classify the same data. The classifier accuracy scores do not reflect any 'ground truth', but rather how well they performed classifying individual tweets into categories chosen by analysts. Furthermore, without further study of specific context or how terms are being used, it is difficult to make a judgement about how serious these instances are, and how they relate to existing literature on misogyny (online as well as offline). Algorithms like those applied above are very useful in classifying enormous volumes of data into more manageable datasets as determined by an analyst, but they cannot provide deep insight into the very specific and often context dependent uses of terms or phrases. As such, they are best understood as being indicative of broad trends. Further, and more detailed research is necessary, and will form the basis of this ongoing investigation.

Demos, 2016

---

[i] http://www.reclaimtheinternet.com/

[ii] http://www.demos.co.uk/files/MISOGYNY_ON_TWITTER.pdf

[iii] http://www.demos.co.uk/project/vox-digitas/