The Problem of Identifying Misogynist Language on Twitter (and other online social spaces)

Sarah Hewitt
University Of Southampton
University Road
Southampton, UK
sh9g14@soton.ac.uk

Dr T Tiropanis
University of Southampton
University Road
Southampton, UK
t.tiropanis
@southampton.ac.uk

Dr C Bokhove
University of Southampton
University Road
Southampton, UK
C.Bokhove@soton.ac.uk

CCS Concepts

•Computing methodologies → Machine learning; Cluster analysis;

Keywords

Misogyny, online abuse, Twitter, sentiment analysis, clustering.

1. INTRODUCTION

Misogynist abuse has now become serious enough to attract attention from scholars of Law [7]. Social network platform providers have been forced to address this issue, such that Twitter is now very clear about what constitutes abusive behaviour, and has responded by updating their trust and safety rules [16].

Twitter relies on the community to report abusive and/or threatening tweets using Twitter's online reporting structure, and has a small team of staff to review these manually. The problems for the community in general, and women in particular, are volume and persistence. Lots of abusive tweets can be made by different accounts in seconds, and once the tweets are made, they remain on the web forever through, for example, re-tweets of the original message, or screen grabs.

There is a cost to the platform provider, which could be significant when there is a high volume of complaints. As a public company, Twitter needs to attract advertisers and therefore cannot risk becoming known as a free platform for toxic speech. Given that a high volume of persistent, misogynist tweets can be generated quickly, there are advantages in automated detection. This paper will explore some background research into the area of misogyny online, provide some details of a limited experiment to highlight some of the challenges in carrying out sentiment analysis using tweets from Twitter, and suggest areas for future work which may make it possible to quarantine abusive tweets before they appear online.

WebSci '16 May 22-25, 2016, Hannover, Germany

ACM ISBN 978-1-4503-4208-7/16/05.

DOI: \url{http://dx.doi.org/10.1145/2908131.2908183}

2. BACKGROUND

Twitter is a public, online micro-blogging site where registered users can broadcast and read 140-character status updates, known as tweets. Tweets can be directed at one or more other users by using the '@' prefix. Users can create a network by following other users. Additional information in the form of links to external web sites, short videos and images can be added. The use of the hashtag # in connection with specific text, e.g. #everydaysexism creates a search term. Twitter provides other ways in which the corpus of tweets can be searched, which is explained below. A considerable volume of data generated by the currently active users renders Twitter a rich source of data. As well as the content of tweets, users can add some details about themselves, and provide location data. Twitter also collects other data such as language and time zone.

This data is available to researchers in a number of ways. Access to all public tweets (the firehose) is only available commercially via Twitter's business partners. Twitter does allow free access to the streaming API (application program interface). However, any query is taken as a sample from of all tweets, which can vary. The search API allows users to define search criteria, and will retrieve data from up to 7 days prior the time of the search, although Twitter cannot guarantee that some tweets won't be duplicated and it is more strictly limited than the search API.

Twitter does allow its users to keep their tweets private if they wish, but for many people wishing to build a social network, this option is pointless. Consequently, many women have been the targets of misogynist abuse and threats which may have been amplified by other users joining in either for the 'fun' or as part of a campaign to drive her out, as relevant literature highlights.

In [14] Malpas argued that there is no longer any meaningful boundary between on- and off-line space. This is especially pertinent to online gaming [15] [4, 3, 5]. One of the first papers to study the video game space is [3]. This study looked at the representation of women in video games generally, and concluded that gender roles were often stereotyped.

A recent study [13] using the X-Box Live game Halo 3 found that, as soon as a female gamer's voice was heard, a clear pattern of negative comments ensued and came in for much more abuse than a male one. Phrases including words like 'slut' and 'whore' were used in comments to the female voice.

Table 1: Range of Terms

slag	bitch	whore	cunt
fugly	skank	pig	hysterical
unfuckable	fuckstruggle	rape	hole
slag	bitch	whore	cunt
lesbian	ho	hoe	slut

Table 2: Coded Variables

Disregarded	+ve	-ve	Neutral
21.23%	68.22%	9.34%	1.21%

One of the most recent researchers in the field of misogynist abuse online has been Jane [10, 11, 12]. Her first paper [10] defines e-bile as any communication that relies on technology, and is perceived to be hostile to the recipient. She acknowledges that some e-bile arises as a form of entertainment for some of the perpetrators, or as a game between e-bilers to 'out-bile' each other. However, she points out that, outside of this 'game', the generalised abuse is still extremely problematic. In [11] Jane examines some of the recurring characteristics in the content of the abuse aimed at women. She notes that antagonists are often anonymous, attack women in the public sphere, use sexually explicit language and suggest various sex-acts that could be carried out as a form of correction on the target as a way of punishing them for 'stepping out of line'. Other *ad hominem* remarks made were concerned with appearance. In [12] she argues that *defining* what constitutes misogynist abuse is now moot, and research should move on to find ways of dealing with the problem.

Turning to Twitter, in 2014, Demos [2] published a report in which they examined the volume, degree and type of misogynistic language used on Twitter. Using three key words gathered from the streaming Twitter API from UK accounts, they concluded that the words slut and whore were used in a misogynist way in approximately 18% of tweets. The word rape was used in a threatening manner in 12% of tweets, and in what is classed as a 'metaphoric/casual' way in 29%. Jane would question the labelling of the largest category as 'subversive' [12] because it distracts from the main issue.

More recently, Fox et al [9] analysed the use of a Twitter hashtag, #getbackinthekitchen. The authors investigated whether anonymous participants received more abuse online compared with non-anonymous tweeters, and concluded that this was the case and argued that such attitudes persist even when the individual issuing the abuse engaged in social situations in 'real' life.

3. DATASET

In order to highlight the challenges of identifying misogynist abuse, a small sample of Tweets were gathered using key words drawn from a review of research papers, and are listed in Table 1. From these, the key words 'cunt', 'slut' and 'bitch' were chosen. 'Bitch' was deliberately chosen because represents the most potentially problematic word. It seems to have crept into general use, particularly with young people, as a form of address, although Adams [1] makes the link between this word and misogynist use in popular music culture. The intention was not necessarily to find examples of abusive tweets, but to get a sense of the context in which language, defined as misogynist by the research papers referred to above, is used.

Approximately 5,500 tweets were gathered from Twitter over the course of a week with the search API and utilising NodeXL, a simple extension for Microsoft Excel that stores the data in an easily readable and accessible format. Retweets were treated as having been made by the user themselves (a re-tweet, unless a comment is added to the contrary, is accepted as a signal of agreement with the sentiment of the original tweet). Tweets promoting pornography sites, or containing any other commercial message, wholly or partly in a foreign language, or written in such a way as to be unintelligible were disregarded. The results are shown in Table 2.

Each tweet in this limited study was coded by one researcher using a simple binary model. Only tweets using misogynist language as a form of address or a term of abuse were scored positively. Everything else was either coded negatively, or discarded. Only the content of the tweet was analysed. The tweets were gathered in batches of 200 over the course of a week, at the rate of a few batches during a 24-hour cycle in an effort to try and avoid responses to media events where there may have been a number of retweets around one topic. It is expected that there will be some experimenter bias.

4. DISCUSSION AND PROPOSALS FOR FUTURE WORK

As expected, most of the tweets were coded as positive. However, 'bitch' was used 3,380 times compared with 118 for 'slut' and 228 for 'cunt'. Many of the tweets were repeating the title or lyrics from popular songs, for example the title of Rihanna's single 'Bitch Better Have My Money'.

There are several Natural Language Processing (NLP) techniques that can be used to establish whether the contents of a tweet express a positive or negative sentiment, but, in short, extracting sentiment or opinion from a whole document (consisting of sentences and words) is a complicated process for which there is no complete solution, and a range of methods are best applied. This was discussed at length in [8].

Burnap et al [6] developed an approach to detect racial tension on Twitter which combined conversation analysis methods and text mining to measure tension in a corpus of tweets. The tension detection algorithm classified tweets more accurately than other methods tested. A scalability analysis would be worth conducting to establish if this method could be adapted for detecting misogynist tweets, ranking them from less to more abusive.

However, the problem of defining what constitutes misogynist language remains. With an estimated 5 million tweets per day, Twitter matters. Targeted abuse can destroy lives, as Citron makes clear [7]. Abusive, misogynist tweets can live on the web forever, and work needs to continue to accurately identify the most abusive, and consider quarantining them *before* they appear in the public timeline. One approach would be to use cluster analysis from the field of machine learning. Unlike sentiment analysis, cluster analysis does not necessarily depend on training data or a lexicon. This removes the thorny issue of what some keywords mean when they're used in x context.

5. ACKNOWLEDGEMENTS

This research was funded by the Research Councils UK Digital Economy Programme, Web Science Doctoral Training Centre, University of Southampton. EP/L016117/1.

6. REFERENCES

[1] T. M. Adams. The words have changed but the ideology remains the same: Misogynistic lyrics in rap music. *Journal of Black Studies*, 36(6):938–957, 07 2006.

[2] J. Bartlett, R. Norrie, S. Patel, R. Rumpel, and S. Wibberley. Misogyny on twitter, http://www.demos.co.uk/, 2014, 05.

[3] B. Beasley and T. C. Standley. Shirts vs. skins: Clothing as an indicator of gender role stereotyping in video games. *Mass Communication and Society*, 5(3):279–293, 2002.

[4] A. Braithwaite. 'seriously, get out': Feminists on the forums and the war(craft) on women. *New Media & Society*, 16(5):703–718, 2013.

[5] A. L. Brehm. Navigating the feminine in massively multiplayer online games: gender in world of warcraft. *Frontiers in Psychology*, 2013.

[6] P. Burnap, O. F. Rana, N. Avis, M. Williams, W. Housley, A. Edwards, J. Morgan, and L. Sloan. Detecting tension in online communities with computational twitter analysis. *Technological Forecasting and Social Change*, 95:96–108, 06 2015.

[7] D. K. Citron. Hate crimes in cyberspace. *Harvard University Press*, 09 2014.

[8] R. Feldman. Techniques and applications for sentiment analysis. *Communications of the ACM*, 56(4), 2013.

[9] J. Fox, C. Cruz, and J. Y. Lee. 'perpetuating online sexism offline: Anonymity, interactivity, and the effects of sexist hashtags on social media'. *Computers In Human Behaviour*, 10.(52):436âĂŞ442, 2015.

[10] E. A. Jane. 'your a ugly, whorish, slut'. *Feminist Media Studies*, 14(4):531–546, 2012.

[11] E. A. Jane. 'back to the kitchen, cunt': speaking the unspeakable about online misogyny'. *Continuum*, 28(4):558–570, 2013.

[12] E. A. Jane. Flaming? what flaming? the pitfalls and potentials of researching online hostility. *Ethics and Information Technology*, 2015.

[13] J. H. Kuznekoff and L. M. Rose. Communication in multiplayer gaming: Examining player responses to gender cues. *New Media & Society*, 15(4):541–556, 2012.

[14] J. Malpas. The non-autonomy of the virtual. *Ubiquity*, 2008(May):1–5, 2008.

[15] A. Shaw. What is video game culture? cultural studies and game studies, games and culture, 5(4). 10.:1177/1555412009360414, 2010.

[16] Twitter. Abusive behavior policy, https://support.twitter.com/articles/20169997, 2016-01-28, 2015, 12.