Dimensions of Abusive Language on Twitter

Isobelle Clarke
Aston University
Birmingham, UK
clarkei@aston.ac.uk

Jack Grieve
University of Birmingham
Birmingham, UK
grievejw@gmail.com

Abstract

In this paper, we use a new categorical form of multidimensional register analysis to identify the main dimensions of functional linguistic variation in a corpus of abusive language, consisting of racist and sexist Tweets. By analysing the use of a wide variety of parts-of-speech and grammatical constructions, as well as various features related to Twitter and computer-mediated communication, we discover three dimensions of linguistic variation in this corpus, which we interpret as being related to the degree of interactive, antagonistic and attitudinal language exhibited by individual Tweets. We then demonstrate that there is a significant functional difference between racist and sexist Tweets, with sexists Tweets tending to be more interactive and attitudinal than racist Tweets.

1 Introduction

With the rise of trolling and other forms of abusive language online, many computational methods for detecting abusive language have been introduced. These classifiers have been trained on a wide range of linguistic features, including specific keywords (Xiang et al., 2012), Bag-of-Words (Warner and Hirschberg, 2012), character *n*-grams (Mehdad and Tetreault, 2016), word *n*-grams (Chen et al., 2012; Yin et al., 2009), part-of-speech *n*-grams (Davidson et al., 2017), and various syntactic features (Burnap and Williams, 2014). A variety of extra-linguistic features have also been considered, including gender (Waseem and Hovy, 2016), location (Waseem and Hovy, 2016), user behaviour and performance (Balci and Salah, 2015; Dadvar et al., 2013), and surrounding posts (Yin et al., 2009). Many of these methods assume that abusive language includes profanity and negative sentiment, but such features are not always present in abusive posts. Including offensive terms in the feature set can even hinder the accuracy of classifiers (Davidson et al., 2017), because profanity can be used for amplification and other non-abusive functions, leading to many false positives (Chen et al., 2012). Trolls have also developed more covert ways of abusing others, such as using creative spelling or avoiding offensive words (Hine et al., 2017). These strategies have been accounted for in part by examining the use of offensive words in context, applying spell-correction algorithms (Chen et al., 2012), consulting WordNet (Chen et al., 2012), and using character *n*-grams to deal with the noisiness of online communication (Mehdad and Tetreault, 2016).

Despite this growing body of research, functional variation in abusive language has yet to be investigated directly. At the most basic level, we do not know what is the general repertoire of styles for abusive language that exists online. One way to understand how the structure of language varies depending on its communicative purpose is multi-dimensional analysis (MDA) (Biber, 1988, 1989). MDA is generally based on the relative frequencies of many lexical and grammatical features measured across a corpus of texts representing a particular variety of language. The most important dimensions of linguistic variation are extracted from this dataset through a factor analysis, and then interpreted functionally based on the linguistic features and the individual texts that are most strongly associated with each dimension. In addition to providing a more complete understanding of the structure of abusive language, incorporating this type of information into abusive language classification systems should lead to more robust and principled methods.

The goal of this study is therefore to use MDA to identify the main dimensions of functional linguistic variation in a corpus of racist and sexist abusive Tweets (Waseem and Hovy, 2016). However, because MDA relies on the multivariate analysis of the relative frequencies of linguistic features, it is not suitable for analysing a corpus of Tweets, which typically include fewer than 30 words, and are therefore too short to allow for the relative frequencies of most features to be measured accurately. Rather than concatenate Tweets to form longer texts (e.g. Passonneau et al., 2014), for example by author, which would obscure text-level patterns, we therefore apply a new form of categorical MDA based on a multiple correspondence analysis of the simple occurrence of a variety of lexical and grammatical forms in individual Tweets to identify common patterns of functional variation in abusive Tweets. Finally, we investigate the degree to which the racist and sexist Tweets in our corpus vary in terms of these dimensions.

2 Method

Our dataset is based on the Twitter corpus used in Waseem and Hovy (2016), which contained 136,052 English Tweets, identified by searching for common racial, religious and sexist slurs and terms, as well as hashtags known to trigger hate speech over a 2 month period. With the help of an outside annotator, they coded 16,914 Tweets as either racist (1,972 Tweets by 9 users), sexist (3,383 Tweets by 613 users) or neither racist nor sexist (11,559). Using 'twitteR' package (Gentry, 2016), we downloaded the Tweets based on the Twitter IDs; however, at the time of download only 2,818 Tweets were still available, presumably because the relevant posts had been deleted. Of these Tweets, 628 had been coded as sexist and 858 as racist. Our analysis focuses on these 1,486 Tweets.

In general, research using MDA has been based on a feature set which has grown over time and which has changed depending on the variety and the language under analysis. There are, however, a core set of features related to basic parts-of-speech and grammatical constructions (Biber, 1988), which we have included in our analysis. These features include tense and aspect markers, place and time adverbials, personal pronouns, questions, nominal forms, passives, subordination, complementation, adjectives and adverbs, modals, specialised verb classes, coordination, negation and other lexical classes, such as amplifiers, downtoners and conjunctions. In addition, as is generally the case in MDA studies (e.g. Grieve et al., 2010), we included additional features to refine our analysis for this particular variety of language, including hashtags, URLs, capitalisation, imperatives, comparatives, and superlatives. We then tagged our corpus for each of the 86 linguistic features. This was achieved by first tagging the Tweets for basic part-of-speech information using the Gimpel et al. (2011) Twitter Tagger. Based on the tagged corpus, we then automatically identified occurrences of our 86 features in the corpus by looking for specific tags, words, and sequences of tags and words, taking into account various exceptional forms found in this corpus.

Rather than measure the relative frequency of these forms across the texts in the corpus, we simply considered whether or not each of these features occurred in each of the texts, retaining the 81 features that occurred in at least 1% of the Tweets in our corpus. We then subjected this 81 feature by 1,486 text binary data matrix to a multiple correspondence analysis (MCA) in R using FactoMineR (Husson et al., 2017). MCA is essentially a dimension reduction method, which aims to represent high dimensional categorical data in low dimensional space, similar to factor analysis as used in traditional MDA for continuous data. MCA is predominantly used to analyse data from questionnaires and surveys (Husson et al., 2010), but it has also been used in linguistics, most notably in lexical semantics (e.g. Tummers et al., 2012; Glynn, 2009, 2014).

The MCA returns a positive or negative coordinate for each linguistic feature on each dimension as well as a value indicating the variables contribution to that dimension (Le Roux and Rouanet, 2010). If the variables' coordinates are of similar value, then this indicates that these variables often co-occur in Tweets. The MCA also assigns a positive or negative coordinate to each Tweet on each dimension, which can then be plotted to visualize the relationship between the Tweets on each dimension. Tweets with similar coordinates on a dimension will share linguistic features. Each dimension was interpreted by considering the functional properties shared by the linguistic features with the strongest contributions. Following Le

Table 1: The positive and negative features strongly contributing to the Dimensions

Dim	Coord	Features
2	+	Question mark (4), Question do (3.9), Accusative case (3.8), absence of Prepositions (3.5), absence of Nouns (3.3), 2nd person pronoun (3.1), absence of Proper nouns (2.9), Emoticons (2.4), absence of Articles (2.4), Nominative case (2.3), Other pronouns (2.2), WH-words (2.1), absence of Attributive adjectives (2.1), Initial DO (2), absence of Be as main verb (1.8), absence of Coordinating conjunctions (1.2), 1st person pronouns (1.2), Subject pronouns (1.1), Initial verbs (.9), WH-clause (.9), Exclamation marks (.8), Quotation marks (.7), absence of Mentioning (.7), Hashtags (.7), Interjections (.6)
	-	Existentials (5.5), Place adverbials (5.4), BE as main verb (3.3), Coordinating conjunctions (2.3), Proper nouns (2.3), absence of Nominative case (2), Articles (1.9), Quantifiers (1.9), Attributive adjectives (1.6), Synthetic negation (1.5), Predicative adjectives (1.2), Contrastive conjunctions (1.2), absence of Other pronouns (1.1), Nominalisations (1.1), Prepositions (1), Numerals (.9), absence of 2nd person pronouns (.9), absence of Accusative case (0.9), Perfect aspect (.7), Determiners (.7), absence of Question marks (.7)
3	+	Question DO (9), Question marks (6.8), 2nd person pronouns (6.8), absence of Subject pronouns (4.4), Initial DO (3.7), Initial verbs (3.2), Determiners (3), Nominalisation (2), Synthetic negation (2), Possessive pronouns (1.9), absence of 1st person pronouns (1.8), Other pronouns (1.7), absence of Nominative case (1.1), absence of Third person pronoun (1), Pro-verb DO (.9), Emoticons (.8), Existentials (.8), BE as main verb (.7)
	-	Subject pronouns (8.7), 1st person pronouns (6.2), Auxiliary BE (3.2), 3rd person pronouns (2.8), Object pronouns (2.5), absence of 2nd person pronouns (1.9), Progressive aspect (1.8), absence of Determiners (1.7), Verbs of perception (1.6), Nominative case (1.3), absence of Mentioning (1.2), absence of Question marks (1.2), absence of Other pronouns (.9), Passives (.8)
4	+	Predicative adjectives (4.5), Existentials (4.4), absence of Prepositions (3.7), absence of Proper nouns (3.5), BE as main verb (3.4), Place adverbials (3), Emoticons (2.5), absence of Nouns (2.3), Synthetic negation (2.3), absence of Capitalisation (2), Subject pronouns (1.9), 1st person pronouns (1.9), absence of Past tense (1.4), Interjections (1.3), absence of Auxiliary BE (1.2), Comparatives (1.1), absence of Articles (1), Requests (.9), absence of URLs (.8), Nominative case (.8)
	-	Auxiliary BE (7.3), Progressive aspect (4.6), Hashtags (3.9), Capitalisations (3.2), By-passives (3.3), URLs (3.1), Proper nouns (2.8), Public verbs (2.1), absence of BE as main verb (1.8), Past tense (1.5), Numerals (1.5), Question DO (1.3), Passives (1), Prepositions (1), Perfect aspect (1), absence of Subject pronouns (1), Articles (0.8), absence of Nominative case (0.7), absence of Predicative adjectives (0.7), Infinitives (0.7)

Roux and Rouanet (2010), we interpreted each dimension by considering all features with a contribution that exceeds 0.62, the average contribution of a feature on a dimension (100/162). In addition, the Tweets with the highest positive and negative coordinates on each dimension were subjected to a micro-analysis to confirm and refine these functional interpretations. Finally, the racist and sexist Tweets were compared on each dimension using Wilcoxon signed-ranked tests to see if there were any functional differences in these two forms of abusive language.

3 Results

We chose to use MCA to extract 4 dimensions based primarily on the functional interpretability of these dimensions. However, because longer Tweets are more likely to contain more features, it is also important to consider whether text length may have confounded our analysis. In standard MDA text length is controlled for by analysing the relative frequencies of features (i.e. by dividing the frequency of a feature in a text by the total number of words in the text), allowing texts of different lengths to be compared. In this case, relative frequencies are not reliable because Tweets are so short, which is why we measured the simple occurrence of forms rather than their relative frequencies and why we used MCA rather than Factor Analysis. To measure the degree to which our analysis was affected by variation in text length, we correlated the dimension coordinates returned by the MCA for each Tweet against Tweet length. Overall, we found that Dimension 1 is strongly positively correlated to Tweet length (r = .72), Dimension 2 is moderately negatively correlated (r = -.33), and Dimensions 3 and 4 are only weakly correlated (r = .02 and r = -.23). The strong correlation between Dimension 1 and Tweet length is reflected by the fact that the positive features that contribute most strongly to this dimension involve the occurrence of a wide range of forms, whereas the negative features that contribute most strongly involve the absence of a wide range of forms. By excluding Dimension 1 from our primary interpretative analysis, because it primarily reflects Tweet length, we were thus able to largely control for text length in our analysis, despite not analysing relative frequencies. The features that contribute the

most to the remaining 3 dimensions, which we interpret below, are presented in Table 1.

3.1 Dimension 2: Interactive

Features with strong contributions and positive coordinates on Dimension 2 have an interactive function. For example, question marks, question DO, WH-words and initial DO are indicative of questions being asked. First and second person pronouns are used to involve the writer and the reader in the discourse. Verb-initial sentences are common in computer-mediated communication when the subject, often the author, is omitted because such information is retrievable from the context (Bieswanger, 2016). Hashtags are used to contribute to and interact with a discussion feed. Quotation marks are used to refer to someone elses speech/words. Interjections are immediate responses to stimuli and emoticons can be used to represent responsive facial expressions.

This interpretation is supported by Examples 1-4, which are Tweets that are strongly associated with positive Dimension 2. All four examples exhibit an interactive style. For example, each Tweet contains at least one second person pronoun. Example 2, 3 and 4 all contain a hashtag and are thus interacting with the feed, whereas Example 1 mentions another user and is therefore interacting directly with another account.

1. *@username Do you think implying someone cant get laid is sexist or abusive?*

2. *#QuestionsForMen Did you know that when you look at a girl - you rape her? http://...*

3. *#QuestionsForMen Did you know that scientists agree that women slut shame to make vaginas more valuable to you? http://t...*

4. *#DontDateSJWs unless you want them to date you, bang you, call you, stalk you THEN cry rape and do performance art. http://t*

Alternatively, features with strong contributions and negative coordinates on Dimension 2 are associated with a more informational style, maximising the amount of information being expressed in 140 characters. For example, existential *there* introduces things or statements. *Be* as main verb and predicative adjectives serve to identify a characteristic, role or attribute of a subject noun phrase. The use of numbers, attributive adjectives, quantifiers, place adverbials, prepositions and proper nouns allow for the expression of detailed descriptions and specific information. Nominalisations are similarly indicative of a high informational load. Contrastive conjunctions emphasise a contrast between two ideas and coordinating conjunctions link two sentences together. Synthetic negation can be used to increase the emphatic force of a statement (Tottie, 1983). This interpretation is also supported by the moderate negative correlation with text length, which reflects the fact that longer Tweets tend to be more informationally dense.

This interpretation is supported by Examples 5-8, which are Tweets that are strongly associated with negative Dimension 2. All four examples exhibit an informational as opposed to an interactive style. For example, each Tweet is made up of 1 or more declarative sentences, headed by the main verb 'to be', which is used to provide identifying information. Synthetic negation can also be seen in Examples 5, 7 and 8, where it is used to present information in an absolutist way.

5. *@username @username @username There is no comparing the vileness of Mohammed to Jesus or Buddha, or Lao Tse. He was simply a criminal*

6. *@username @username Muslims have been raping white girls with Labors approval for 16 years. Any ukip just got there.*

7. *@username @username @username There are no Jews in Saudi or many of the Gulf estates because the Muslims exterminated them.*

8. *@username @username @username There was no golden age. Jews were regularly slaughter by Muslims in pogroms.*

Overall, Dimension 2 is therefore interpreted as representing the degree of interactiveness exhibited by a Tweet. Notably, previous MDA studies (e.g. Biber, 1988, 1989; Grieve et al., 2010) have found a similar primary dimension, which opposes two of the most basic functions of language, namely interacting and informing.

3.2 Dimension 3: Antagonistic

Features with strong contributions and positive coordinates on Dimension 3 have an antagonistic function. For example, several of these fea-

tures are associated with forming questions, including question DO and initial DO, verb initial, and question marks, which can be used to make demands of other users. Second person pronouns are also associated with antagonistic language especially when accompanied by the absence of first and third person pronouns as well as subject pronouns in general, which indicates that these Tweets are targeted at specific users. The co-occurrence of nominalisations with these features is associated with a high degree of specificity, whilst features such as possessive pronouns function to indicate possession, implying that someone is being targeted and challenged on specific information or possessions. A high degree of specificity in questions are common in adversarial discourse, for example in cross-examination questions, which are typically loaded and structured to confuse the witness and discredit their statement (Gibbons, 2008). Furthermore, emoticons and exclamation marks can be associated with more aggressive forms of online communication.

This interpretation is supported by Examples 9-12, which are Tweets that are strongly associated with positive Dimension 3. They all contain questions antagonistically directed to other users. Specifically, in all four cases, something has been noticed by the Tweeter and is now being opposed through questioning.

9. *@username Can you be legally forced into parental obligations? Can your genitals be cut at birth? Does your right to vote have an *?*

10. *If being pro-due process makes you pro-rape, does being anti-death penalty make you pro-murder? http://t...*

11. *#AskAWhiteFeminist Seriously, what rights dont you have, and why can none of you answer that question?*

12. *@username1 Do you approve of your pedophile prophet raping a 9 year old girl, like it says in 7 hadith?*

Alternatively, features with strong contributions and negative coordinates on Dimension 3 are associated with a more conciliatory style. Obviously, abusive language is inherently antagonistic; however, the absence of second person pronouns and the presence of subject and object pronouns, particularly third person pronouns and first person pronouns, indicates that Tweets scoring negatively on this Dimension are not targeting particular individuals. The co-occurrence of the progressive tense indicates that continuing action is being described. Object pronouns suggest that this action is affecting or influencing particular people. The co-occurrence of first person pronouns and verbs of perception suggest that the writer is giving their account of what they are perceiving, rather than opposing people directly.

This interpretation is supported by Examples 13-16, which are Tweets that are strongly associated with negative Dimension 3, all of which reflect an 'us versus them dichotomy', whereby descriptions of the actions of 'them' are either perceived by the person speaking or the actions of 'them' are influencing 'us'. Several of the Tweets are directed to more than one user suggesting that they are part of a conversation between friends/acquaintances. While the people being spoken about may find the messages abusive, they are not targeted to them, hence the language appears to be more collaborative than antagonistic, with people involved in the conversation sharing the same views, even though those views would be considered offensive by others.

13. *@username1 I saw him, but I rarely engage male fems zero point to it. They are just following orders*

14. *@username1 @username2 I actually wish they would just start using egalitarian so we can just let feminist mean the misandrist hypocrites.*

15. *@username1 @username2 Reminds me of Simpsons where grandpa was screaming Death at everything. Now its rape. http://...*

16. *@username1 @username2 @username3 They are breeding us out of existence in the westernised world. Islam will rule the world in time.*

Overall, Dimension 3 is therefore interpreted as representing the degree of antagonism exhibited by a Tweet. Previous definitions of trolling have suggested that such posts tend to be hostile and aggressive (Hardaker, 2010). Moreover, it has been shown that adversarial behaviour, such as anger and accusation are common online, especially when discussing ideological issues because the purpose is to dominate the discourse and such

adversarial behaviour can perform this function (Herring et al., 1995).

3.3 Dimension 4: Attitudinal

Features with strong contributions and positive coordinates on Dimension 4 have an attitudinal function. For example, comparatives are used to describe people or things in relation to others. Predicative adjectives and BE as a main verb function to describe and identify particular attributes or characteristics of the subject. First person pronouns involve the Tweeter in the discourse, marking the post as a personal opinion. The co-occurrence of existential *there* with these features and the absence of nouns suggests that descriptions are being introduced rather than things, or that information is being introduced and then an opinion is given. Synthetic negation indicates that something is being contested.

This interpretation is supported by Examples 16-20, which are Tweets that are strongly associated with positive Dimension 4. All of these examples are expressions of opinions and personal stance through features such as first person pronouns, *be* as a main verb and adjectives.

17. *@username1 @username2 @username3 and we still get payed equally. That stupid myth bothers me to no end because theres really things -*

18. *@username1 No. You have proven your ignorance here to anyone who isnt as dumb as you. Its there for all to see but you dont know it.*

19. *@username1 I have no religion, but I can accommodate Jews, Hindus, Buddhists, Taoist, Atheists. But Islam is too cancerous*

20. *@username1 @username2 Except that there was no such sexual torture and she is a lying bitch*

Alternatively, features with strong contributions and negative coordinates on Dimension 4 do not share this attitudinal function. Instead, several features have a reporting function. For example, public verbs mark indirect or reported speech, the perfect tense is used to report on past events, and the progressive tense refers to continuing action. Passive constructions serve to emphasise the object acted upon, rather than the agent. Agentless passive constructions are common in ideological discourse as they can be used to reduce the agents prominence and therefore the blame or cause, whereas *by*-Passives take the information that would be typically new information and present it at the beginning of the sentence as given information and the agent is then moved to the end of the clause and presented as new information (Fairclough, 1992). URLs function to direct the reader to more information, including a website or an image. Numbers serve to add additional, specific information. The co-occurrence of URLs and numbers with ideological features suggests that they are functioning to support a point or provide proof, either in the form of additional textual or quantitative information. Capitalisation suggests that the writer is either emphasising a point or raising their voice.

This interpretation is supported by Examples 21-24, which are Tweets that are strongly associated with negative Dimension 4.The examples support this interpretation as the speakers point of view is not explicitly marked, but rather the action of others is being reported and supported with numbers and URLs.

21. *@username1 @username2 The Jews are trying to defend themselves against Muslims trying to exterminate them. http://...*

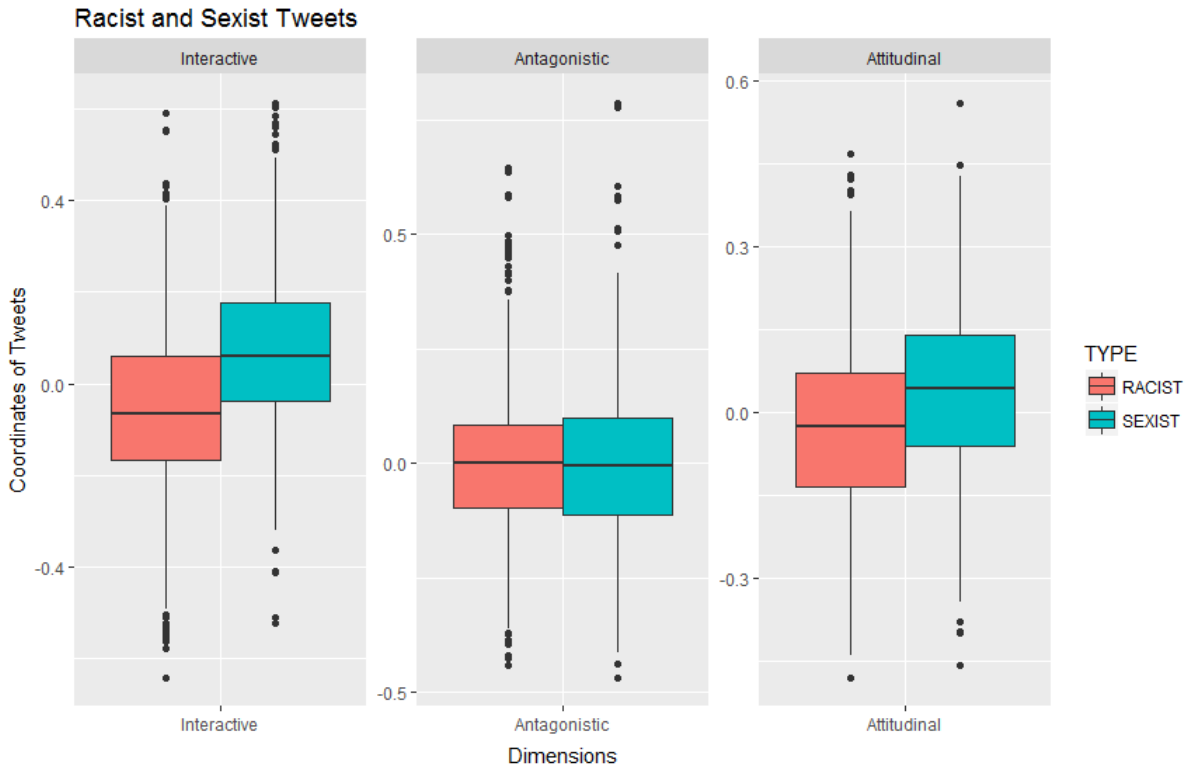
22. *In Islam women must be locked in their houses, and Muslims claim this is treating them well. http://...*

23. *@username1 @username2 The world is doing nothing. Islam is producing the terrorist activities and has been for 1400 years*

24. *Following the example of the pedophile prophet Mohammed in every detail, one ISIS militant is marrying a 7 year old child in Mosul. #Islam*

Overall, Dimension 4 is therefore interpreted as representing the degree of attitudinal judgment exhibited by a Tweet. Abusive language is by nature attitudinal and ideological. However, it has been shown that such beliefs can be realised in various ways, such as through explicit opinions or by telling stories, which present the other in a negative light (e.g. van Dijk, 1993). Thus, the degree of attitudinal judgement reflects the way in which the attitude is encoded.

Figure 1: Boxplots of Racist and Sexist Tweets for Dimension 2, 3, and 4



3.4 Racist versus Sexist Tweets

Following the interpretation of our three primary functional dimensions, we tested the extent to which racist and sexist Tweets differ along each of these dimensions. In particular, we used a Wilcoxon signed-ranked tests to test if there were any functional differences between the coordinates of racist and sexist Tweets on each dimension. In addition, we produced boxplots to help visualise each comparison (see Figure 1). Overall, we found significant differences ($p < .01$) between racist and sexist Tweets on Dimensions 2 and 4, with sexist Tweets tending to be more interactive and attitudinal than racist Tweets. We did not, however, find a significant difference in the degree of Antagonism between racist and sexist Tweets.

To interpret these findings, we considered previous studies on sexist and racist language and strategies. For example, in a study examining sexist strategies on two email lists, Herring et al. (1995) found that one silencing strategy employed in sexist language is to dismiss the points raised by others by referring to their 'triviality'. It is possible to see in Examples 17 and 18 that the significance of a point is being disputed. In Example 17, "stupid myth" not only represents something as nonfactual through the word "myth", but also represents it as trivial and benign through "stupid". In Example 18, the intelligence of the speaker is being called into question, thereby discrediting the original posters statement and presenting it as trivial. Thus, it may be that expressions of attitudinal judgement, specifically by encoding that the previous post is trivial, are serving the over-arching aim to silence the individual. Another silencing strategy employed in sexist discourse is to regain control over the conversation by introducing new topics (Herring et al., 1995). This strategy may provide a reason for why sexist tweets are more interactive as the over-arching aim may be to regain control and therefore they may ask new questions and interact by introducing new topics.

In regards to racist language, van Dijk (1993) describes that racist ideologies have been shown to be reproduced through story-telling and argumentation. Specifically, stories are told by people from majority groups about minority groups in the form of complaints or negative events (van Dijk, 1993). Although stories are often associated with personal expression and opinion, stories are used to inform people, and can take the form of news reports. These stories are functionally less entertaining, but serve more to argue a point or

persuade (van Dijk, 1993). It is possible to draw similarities with what van Dijk (1993) says here with the informational and reporting function of the racist Tweets. The argumentative and persuasive function of racist discourse is apparent in the racist Tweets with the use of URLs and numbers in Examples 21-24, which function here to provide supporting evidence. Furthermore, story-telling involves introducing a complication and providing contextual information, rather than interacting. This can be seen in Examples 5, 7 and 8 through the existential *there*, which functions to introduce new information. Thus, it may be suggested that racist Tweets are less interactive and attitudinal because the aim is to persuade and argue a point by reporting on events which presents minority groups in a negative light. In other words, it presents racist opinions as facts as a way to legitimate racist ideologies.

4 Conclusion

Many classifiers used to detect abusive language are trained on offensive terms. In this study, we aimed to avoid using offensive terms, and instead examined a wide range of functionally-significant grammatical features to identify the main dimensions of functional linguistic variation that occur in racist and sexist Tweets. Although we do not apply our results directly to the task of abusive language detection here, such linguistic co-occurrence patterns could in all likelihood be usefully incorporated into future classification models. Furthermore, the general patterns we have identified in this paper should help to explain why some features work better than others for detecting and distinguishing forms of abusive language online and suggest new directions for feature selection.

In summary, based on the analysis of Waseem and Hovys (2016) data, and using a novel categorical approach to MDA, we have identified 3 dimensions of linguistic variation in racist and sexist Tweets: *interactive*, *antagonistic*, and *attitudinal*. Although there is no absolute distinction between racist and sexist Tweets, by plotting each Tweets dimension coordinates, we have revealed that racist and sexist Tweets do differ functionally in respect to Dimension 2 and Dimension 4, with sexist Tweets tending to be more interactive and attitudinal, perhaps reflecting a somewhat different intent for racist and sexist Tweets. These results suggest that certain features used for classifiers may be biased towards particular types and functions of abusive language. For example, studies selecting the word tri-gram "you are [adjective: offensive word]" (e.g. Chen et al., 2012) are likely to find Tweets that have an interactive and attitudinal function. As a result, other linguistic co-occurrence patterns that represent other functions of abusive language may be missed.

The *antagonistic* and *attitudinal* dimensions are perhaps the most obvious because abusive language is by nature hostile, opinionated and controversial. Nevertheless, we have demonstrated that abusive language can be discussed amongst acquaintances meaning that the function of the interaction changes to be less antagonistic and more collaborative, at least to its immediate audience. Additionally, we have shown that abusive language does not have to be attitudinal as the speakers point of view can be suppressed and the Tweets can function to report on action and provide evidence to such reports.

Without relying on profanity, we have highlighted the value of such research in identifying particular linguistic co-occurrence patterns and functional variation in abusive language. Unfortunately, we have only looked at these particular racist and sexist Tweets and therefore the dimensions could change with more data. However, in the future we aim to gather a larger corpus of different types of abusive language and improve the feature set in order reveal further and more detailed dimensions of linguistic variation of abusive language. Moreover, we aim to validate these dimensions by collecting a corpus of non-abusive language and making comparisons between the two.

References

K. Balci and A. A. Salah. 2015. Automatic analysis and identification of verbal aggression and abusive behaviors for online social games. *Computers in Human Behavior* 53:517–526.

D. Biber. 1988. *Variation across speech and writing*. Cambridge University Press, Cambridge.

D. Biber. 1989. A typology of english texts. *Linguistics* 27:3–43.

M. Bieswanger. 2016. Electronically-mediated englishes: Synchronicity revisited. In L.Squires, editor, *English in Computer-Mediated Communication:*

Variation, Representation, and Change, De Gruyter, Berlin Boston.

P. Burnap and M. L. Williams. 2014. Hate speech, machine classification and statistical modelling of information flows on twitter: Interpretation and communication for policy decision making. In *Proceedings of the Internet, Policy and Politics Conference*. Oxford.

Y. Chen, Y. Zhou, S. Zhu, and H. Xu. 2012. Detecting offensive language in social media to protect adolescent online safety. In *Proceedings of the 2012 ASE/IEEE International Conference on Social Computing and 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust*. IEEE Computer Society, Washington, DC, USA, SOCIALCOM-PASSAT '12, pages 71–80. https://doi.org/10.1109/SocialCom-PASSAT.2012.55.

M. Dadvar, D. Trieschnigg, and F. de Jong. 2013. Expert knowledge for automatic detection of bullies in social networks. In *Proceedings of the 25th Benelux Conference on Artificial Intelligence, BNAIC 2013, Delft, the Netherlands*. TU Delft, pages 57–64.

T. Davidson, D. Warmsley, M. Macy, and I. Weber. 2017. Automated hate speech detection and the problem of offensive language. In *In Proceedings of the 11th International AAAI Conference on Web and Social Media ICWSM17*. World Academy of Science, Engineering and Technology. https://arxiv.org/pdf/1703.04009.pdf.

N. Fairclough. 1992. *Discourse and Social Change*. Blackwell Publishing Ltd., Cambridge.

J. Gentry. 2016. Package twitter pages 1–30. https://cran.r-project.org/web/packages/twitteR/twitteR.pdf.

J. Gibbons. 2008. Questioning in common law criminal courts. In J. Gibbons and M. Teresa Turell, editors, *Dimensions of Forensic Linguistics*, John Benjamins Publishing Company, Amsterdam/Philadelphia, pages 115–130.

K. Gimpel, N. Schneider, B. OConnor, D. Das, D. Mills, J. Eisenstein, M Heilman, D Yogatama, J Flanigan, and N. A Smith. 2011. Part-of-speech tagging for twitter: Annotation, features, and experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Short Papers*. Association for Computational Linguistics, pages 19–24. www.aclweb.org/anthology/P11-2008.

D. Glynn. 2009. Polysemy, syntax, and variation: A usage-based method for cognitive semantics. In V. Evans and S. Pourcel, editors, *New directions in Cognitive Linguistics*, John Benjamins, Amsterdam, pages 77–106.

D. Glynn. 2014. Correspondence analysis: Exploring data and identifying patterns. In D. Glynn and J. A. Robinson, editors, *Corpus Methods for Semantics: Quantitative studies in polysemy and synonymy*, John Benjamins, Amsterdam, pages 443–485.

J. Grieve, D. Biber, E. Friginal, and T Nekrasova. 2010. Variation among blogs: A multi dimensional analysis. In A. Mehler, S. Sharoff, and M. Santini, editors, *Genres on the web: Computational Models and Empirical Studies*, Springer-Verlag, New York, pages 45–71.

C. Hardaker. 2010. Trolling in asynchronous computer-mediated communication: From user discussions to academic definitions. *Journal of Politeness Research* 6:215–242.

S. Herring, D. A. Johnson, and T. DiBenedetto. 1995. This discussion is going too far! male resistance to female participation on the internet. In K. Hall and M. Bucholtz, editors, *Gender Articulated: Language and the Socially Constructed Self*, Routledge, New York and London, chapter 3, pages 67–96.

G. E. Hine, J. Onaolapo, E. De Cristofaro, N. Kourtellis, I. Leontiadis, R. Samaras, G. Stringhini, and J. Blackburn. 2017. Kek, cucks, and god emperor trump: A measurement study of 4chan's politically incorrect forum and its effects on the web. In *Proceedings of the 11th International AAAI Conference on Web and Social Media (ICWSM-17)*. https://arxiv.org/abs/1610.03452.

F. Husson, J. Josse, S. Le, and J. Mazet. 2017. Factominer: Multivariate exploratory data analysis and data mining pages 1–96. https://cran.r-project.org/web/packages/FactoMineR/FactoMineR.pdf.

F. Husson, S. Lł, and J. Pags. 2010. *Exploratory Multivariate Analysis by Example Using R*. Chapman & Hall CRC, London.

B. Le Roux and H. Rouanet. 2010. *Multiple Correspondence Analysis*. SAGE Publications, Inc., California.

Y. Mehdad and J. Tetreault. 2016. Do characters abuse more than words? In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Association for Computational Linguistics, Los Angeles, pages 299–303. http://www.aclweb.org/anthology/W16-3638.

R. J. Passonneau, N. Ide, S. Su, and J. Stuart. 2014. Biber redux: Reconsidering dimensions of variation in american english. In *COLING 2014, 25th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, August 23-29, 2014, Dublin, Ireland*. pages 565–576. http://aclweb.org/anthology/C/C14/C14-1054.pdf.

G. Tottie. 1983. *Much about not and nothing: a study of the variation between analytic and synthetic negation in contemporary American English*. CWK Gleerup, Lund.

J. Tummers, D. Speelman, and D. Geeraerts. 2012. Multiple correspondence analysis as heuristic tool to unveil confounding variables in corpus linguistics. In *Proceedings of the 11th International Conference on Statistical Analysis of Textual Data*. JADT, pages 923–936.

T. A. van Dijk. 1993. Stories and racism. In D. K. Mumby, editor, *Narrative and Social Control: critical perspectives*, Sage, Newbury Park, CA, chapter 5.

W. Warner and J. Hirschberg. 2012. Detecting hate speech on the world wide web. In *Proceedings of the Workshop on Language and Social Media*. The Association for Computational Linguistics, pages 19–26. https://aclweb.org/anthology/W/W12/W12-21.pdf.

Z. Waseem and D. Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of Proceedings of NAACL-HLT 2016*. The Association for Computational Linguistics, pages 88–93. https://aclweb.org/anthology/N/N16/N16-2.pdf.

G. Xiang, B. Fan, L. Wang, J. Hong, and C. Rose. 2012. Detecting offensive tweets via topical feature discovery over a large scale twitter corpus. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*. ACM, New York, NY, USA, CIKM '12, pages 1980–1984. https://doi.org/10.1145/2396761.2398556.

D. Yin, Z. Xue, L. Hong, B. D. Davison, A. Kontostathis, and L. Edwards. 2009. Detection of harassment on web 2.0. In *Proceedings of the Content Analysis in the WEB*. Dublin City University and Association for Computational Linguistics, pages 1–7. http://www.cse.lehigh.edu/ brian/pubs/2009/CAW2/harassment.pdf.