Title	Detecting weak and strong Islamophobic hate speech on social media
Authors(s)	Vidgen, Bertie; Yasseri, Taha
Publication date	2020
Publication information	Journal of Information Technology and Politics, 17 (1): 66-78
Publisher	Taylor & Francis
Item record/more information	http://hdl.handle.net/10197/12720
Publisher's statement	This is an Accepted Manuscript of an article published by Taylor & Francis in Journal of Information Technology and Politics on 13 December 2020, available online: https://doi.org/10.1080/19331681.2019.1702607
Publisher's version (DOI)	10.1080/19331681.2019.1702607

Detecting weak and strong Islamophobic hate speech on social media

Bertie Vidgen[1]
Oxford Internet Institute, University of Oxford
bertievidgen@gmail.com

Taha Yasseri
Oxford Internet Institute, University of Oxford
Alan Turing Institute, London
taha.yasseri@oii.ox.ac.uk

ABSTRACT

Islamophobic hate speech on social media inflicts considerable harm on both targeted individuals and wider society, and also risks reputational damage for the host platforms. Accordingly, there is a pressing need for robust tools to detect and classify Islamophobic hate speech at scale. Previous research has largely approached the detection of Islamophobic hate speech on social media as a binary task. However, the varied nature of Islamophobia means that this is often inappropriate for both theoretically-informed social science and effectively monitoring social media. Drawing on in-depth conceptual work we build a multi-class classifier which distinguishes between non-Islamophobic, weak Islamophobic and strong Islamophobic content. Accuracy is 77.6% and balanced accuracy is 83%. We apply the classifier to a dataset of 109,488 tweets produced by far right Twitter accounts during 2017. Whilst most tweets are not Islamophobic, weak Islamophobia is considerably more prevalent (36,963 tweets) than strong (14,895 tweets).

Our main input feature is a gloVe word embeddings model trained on a newly collected corpus of 140 million tweets. It outperforms a generic word embeddings model by 5.9 percentage points, demonstrating the importan4ce of context. Unexpectedly, we also find that a one-against-one multi class SVM outperforms a deep learning algorithm.

KEYWORDS

Hate speech, Islamophobia, social media, prejudice, extremism

1 Introduction

In recent times, the prevalence, effects and spread of Islamophobic hate speech on social media has received considerable attention from government (APPG on British Muslims, 2018; HM Government, 2012; Home Affairs Select Committee, 2017), Muslim community groups (Ingham-Barrow, 2018; Runnymede Trust, 2017; Tell Mama, 2018), academics (Allen, 2010; Burnap & Williams, 2016) and the platforms themselves (Facebook, 2018; Twitter, 2018). Islamophobic hate speech inflicts considerable harm on both targeted individuals and wider society, and risks reputational damage for the host platforms.

Islamophobia has been variously described as a form of racism (Meer & Modood, 2009), stereotyping (Moosavi, 2015), prejudice (Imhoff & Recker, 2012), fear (Kunst, Sam, & Ulleberg, 2013), exploitation (Beck, Charania, & Al-issa, 2017), exclusion (Bayrakli & Hafez, 2018) and dominance (Jackson, 2018). It can be understood as what W. B. Gallie terms an 'essentially contested concept' – a concept with numerous definitions and descriptions but little consensus as to what the core features are (Gallie, 1956). In the present work, we use Bleich's widely cited definition of Islamophobia, as this reflects much other work undertaken by leading theorists in the field (Allen, 2010; Awan, 2016; Ekman, 2015). Bleich defines Islamophobia as: 'Indiscriminate negative attitudes or emotions directed at Islam or Muslims.' (Bleich, 2011, p. 1581). This definition can be adapted for social media posts:

> "Any content which is produced or shared which expresses indiscriminate negativity against Islam or Muslims."

Recent interventions in social psychology point to the multifaceted nature of prejudice; from behaviors which are explicit, overt and direct to those which are implicit, covert and indirect (Nadal, Griffin, Hamit, Leon, & Rivera, 2012; Pettigrew & Meertens, 1995). Thus, in recent times much research has focused on 'everyday' prejudicial and hateful actions (Dunn & Hopkins, 2016; Moosavi, 2015) as well as 'micro-aggressions' (Haque, Tubbs, Kahumoku-Fessler, & Brown, 2018; Husain & Howard, 2017). Little research has explicitly explored these distinctions with regard to hate speech. This is surprising given that distinguishing between different types of Islamophobic speech offers considerable empirical and theoretical advantages over using a single category of 'Islamophobia'. It enables researchers to better understand the dynamics of Islamophobia (which may differ across different manifestations) and to investigate radicalization processes, whereby individuals progress from being weak to strongly Islamophobic. It is also important for enabling platforms and governments to better regulate and monitor social media and provide support to victims. Ultimately, it could also lead to better detection of hate speech; Waseem and Hovy note that 'in much hate speech research, diverse types of abuse have been lumped together under a single label, forcing models to account for a large amount of within-class variation.' (Waseem & Hovy, 2016, p. 82).

[1] Primary author, email for further information.

2 Classification task

The classification task addressed is to distinguish between non-Islamophobic, weak Islamophobic and strong Islamophobic social media content. Drawing on Bleich's definition of Islamophobia, as well as work undertaken with victims of Islamophobia by the Runnymede Trust, MIND and Tell Mama (Ingham-Barrow, 2018; Runnymede Trust, 2017; Tell Mama, 2018), we define strong Islamophobic hate speech as:

> Speech which explicitly expresses negativity against all Muslims.

This can vary, from expressing explicitly negative *views*, such as describing Muslims as barbarians to calling for prejudicial *actions*, such as demanding that Muslims are forcibly banned from the UK. We define weak Islamophobic hate speech as:

> Speech which weakly expresses negativity against all Muslims
>
> AND
>
> Speech which explicitly expresses negativity against a specific subset of Muslims

An example of the first type of Islamophobia is making a comment about how Muslims are 'different' or have unusual cultural practices. An example of the second type of weak Islamophobia is sharing a new story about a terrorist attack and explicitly foregrounding the fact that the perpetrator is a Muslim.

Whereas blatant Islamophobia is easy to spot, subtle Islamophobia is often harder to observe and may only partially manifest anti-Muslim negativity.

3 Previous work

Previous work in this area demonstrates the challenges of – but also potential for – creating a classification system which distinguishes between weak and strong Islamophobic hate speech. To the authors' best knowledge, no previous research has focused specifically on this task (Schmidt & Wiegand, 2017). Nonetheless, many prior studies are still relevant to the present discussion, not least because most of them have used data from the same source (Twitter). Most previous research has also focused on binary rather than multi-class classification. Classifier performance in the latter task is often far lower. As Salminen et al. note in a recent paper, 'existing works using multi-label classification for online hate speech are extremely rare, and we could not locate prior work that had achieved good results.' (Salminen et al., 2018, p. 331)

Most existing research into multi-class classification has focused on distinguishing between different *targets* of hate rather than different *strengths* (Burnap & Williams, 2016; Park & Fung, 2017; Saleem, Dillon, Benesch, & Ruths, 2017; Salminen et al., 2018; Silva, Mondal, Correa, Benevenuto, & Weber, 2016). It is difficult to complete both tasks at once; distinguishing between different strengths of hate inevitably involves narrowing the domain to just one target (here, Islamophobia)as 'hateful speech classification systems require target-relevant training' (Saleem et al., 2017, p. 7) Classifying content based on strength rather than target poses additional challenges as there is less variation between classes; weak and strong Islamophobic tweets often use similar keywords, grammatical structures and non-verbal features (such as embedded hyperlinks and emojis).

Burnap and Williams train a classifier to distinguish between different levels of cyberhate (divided into 'moderate' and 'extreme' classes) targeted against Black Minority Ethnic (BME) and religious groups on Twitter (Williams & Burnap, 2016), achieving precision of 0.77.

Malmasi and Zampieri distinguish between 'hate' speech, 'Offensive' speech and 'Ok' speech. They achieve 78% accuracy but on an unevenly weighted training/testing dataset – over half of their corpus is 'OK'. Their model struggles to distinguish between non-OK content; of 2,399 'Hateful' instances in their dataset, 1,050 are categorised correctly, 1,113 are miscategorised as 'Offensive' and 236 as 'OK'. They also do not test their model on unseen data, only reporting the results of cross-validation (Malmasi & Zampieri, 2017).

Jha and Mahmidi distinguish between 'benevolent' and 'hostile' sexism. They use Waseem and Hovy's dataset of 16,000 tweets as well as ~7,000 newly collected ones (Waseem & Hovy, 2016). Using SVM they report an F1 score of 0.80 for Benevolent tweets, 0.48 for Hostile and 0.89 for Others. A their data is highly skewed towards Others rather than Hostile, overall performance is strong.

Kumar et al (Kumar, Ojha, Malmasi, & Zampieri, 2018) distinguish between overtly aggressive, covertly aggressive and non-aggressive tweets using a dataset of 15,000 Facebook posts. In a competition entered by 130 teams (of which 20 completed it and provided the technical details of their model), the highest performing obtained a weighted F-score of 0.64. As the authors note, 'the results […] depict how challenging the task is.' (Kumar et al., 2018, p. 1)

Davidson et al. train a model to distinguish between hate speech and offensive speech, and non-offensive speech in tweets. They report impressive results, with precision of 0.91, recall of 0.90 and an F1 score of 0.90. Their work demonstrates the potential for multiclass classification, makes an important theoretical argument apropos the need to separate different types of content, and introduces the use of 'Ease of Reading' metrics as an input feature. However, as they note, their model performs poorly with hate speech, of which almost 40% is misclassified. The high F1 score is largely due to the fact that their classes are very uneven (76% of the data is in the 'offensive speech' category). They also train and test their classifier on a single dataset, which could risk overfitting.

4 Data

We collect a new dataset of 140 million tweets produced by Twitter followers of mainstream and far right UK political parties. The data is collected over the course of 2017 and the first six months of 2018.

It also includes some tweets from before 2017 which are made available by Twitter's Rest API. The tweets come from:

1. 7,500 users randomly selected from followers of UKIP.
2. 7,500 users randomly selected from followers of the Conservatives.
3. 7,500 users randomly selected from followers of Labour.
4. 7,500 users randomly selected from followers of the Liberal Democrats.
5. All ~15,000 followers of the BNP.
6. All ~32,000 followers of Britain First.
7. Every tweet produced by a set of 45 far right accounts (consisting of every group which appear in Hope Not Hate's 2015 and 2017 reports on the far right, and which have a Twitter account (Hope Not Hate, 2015, 2017)).
8. Every @ mention of the same 45 far right accounts (collected from the Twitter Stream API).

4.1 Data annotation

We create a training dataset of 4,000 tweets by sampling from across all 8 of the sources outlined above. Creating a training dataset with sufficient instances of hateful content is a time-consuming endeavor, not least because in most online contexts the prevalence of hate is relatively low overall (Schmidt & Wiegand, 2017, p. 7).To ameliorate this problem, Waseem and Hovy recommend increasing the prevalence of hate speech by sampling data which contains relevant topics (Waseem & Hovy, 2016). This approach is partially adopted here; we sample 1,000 tweets using keyword searches for 'Muslims' and 'Islam'.

All 4,000 tweets are annotated blind by three annotators who are experts in UK politics and the study of prejudice. The annotators all use the same annotation guidelines. The guidelines are based on the definition of Islamophobia offered above and were iteratively developed through two preliminary studies, each consisting of 200 tweets. Across the 4,000 tweets, inter-rater agreement is high. Percentage agreement is 89.9%, Fleiss' kappa is 0.837 and Krippendorf's alpha is 0.895. We also compute category-wise scores for Fleiss' kappa, which range from 0.737 for Weak Islamophobia to 0.907 for Strong Islamophobia. The consistency of these results show the robustness of the annotation guidelines and how they were implemented.

In cases where annotators disagree, tweets are assigned to classes based on the majority decision. In the final dataset, 3,106 tweets are classed as 'Not Islamophobic', 484 tweets are classed as 'Weak Islamophobic', 410 tweets are classed as 'Strong Islamophobia'. To create an evenly-weighted dataset the number of 'Not Islamophobic' tweets is reduced through random sampling to 447 instances (the difference between the number of tweets in the other two classes). This creates a final dataset of 1,341 tweets.

5 Input features

Feature selection refers to the choice of input variables used to train the classifier. In many cases features are selected using 'brute force' computation via a grid search with little consideration for *why* they have been included. Models in which variables are selected without any theoretical justification may perform well in initial testing but risk overfitting, and as such are unlikely to be generalizable, making them unsuitable for empirical research (Domingos, 2012). Thus, it is crucial that the model is not only accurate but that its choice of inputs can be explained and thus avoids becoming a 'black box' (Biran & McKeown, 2017). Accordingly, in the present work, we only consider features which can be theoretically justified.

First, we create a text only model, using one-hot encodings for each term. Second, we create a model using 50 surface-level and derived non-text features. These include sentiment and polarity (Feuerriegel & Proellochs, 2018), count of swear words (Ipsos MORI, 2016) and parts of speech and named entities (Benoit & Matsuo, 2018). We also derive two new input features, mentions of Muslim names and mentions of Mosques, both taken from relevant Wikipedia pages. Third, we create a combined model that uses both one-hot encodings and all 50 of the non-text features. Fourth, we create a model using pre-trained gloVe word embeddings, trained on two billion tweets (Stanford, 2018). Fifth, we create a gloVe model using newly-trained word embeddings on the corpus of 140 million tweets (Pennington, Socher, & Manning, 2014). Finally, sixth, we create a model which uses the newly-trained word embeddings as well as all 50 of the non-text features.

For testing we implement ten-fold cross-validation on the Naïve Bayes algorithm as previous research indicates that it generally outperforms most other off-the-shelf algorithms for text classification tasks (Kotsiantis, 2007; Wainer, 2016; Wang & Manning, 2012) and it is deterministic, producing the same results each time it is implemented. The results are shown in Table 1.

Input feature model	**Accuracy**
Model 1: Text only (one-hot encoding)	30.07%
Model 2: Non-text features	49.96%
Model 3: Text + non-text features	30.36%
Model 4: Pre-trained word embeddings	63.20%
Model 5: Newly trained word embeddings	69.13%
Model 6: Newly trained word embeddings + all non-text features	65.20%

Table 1: Accuracy of models with different input features

The best performing model is the newly trained word embeddings alone (model 5). Interestingly, this considerably outperform the accuracy of the pre-trained word embeddings model (5.9 percentage points, 69.13% compared with 63.2%). This suggests that the benefits of having tweets which are contextually-specific outweighs the cost of having a smaller dataset. This is in line with previous work, such as Lai et al., who report that 'corpus domain is

more important than corpus size.' (Lai, Liu, He, & Zhao, 2016, p. 8). We optimize the newly trained word embeddings model by including additional non-text features, testing for up to ten additional features through an exhaustive grid search. We find that the count of mentions of Mosques is consistently an important input feature, suggesting that this newly engineered feature could also be used in other studies. The final model (model 7), which maximizes accuracy, contains 6 additional non-text features:

> Word embeddings + count of mentions of Mosques + presence of HTML + presence of RT + part of speech: 'conjunction' + named entity recognition: 'location' + named entity recognition: 'organization'

6 Choice of algorithm

We test the newly trained word embeddings model (model 5 in Table 1) on six different algorithms on, selected based on previous research on classification (Kotsiantis, 2007; Wainer, 2016; Wang & Manning, 2012): Naïve-Bayes, Random Forests (with trees = 10, 100 and 1,000), Logistic Regression, Decision Trees, SVM and Deep Learning. We implement multi-class SVM with a one-against-one strategy (Hsu & Lin, 2002). Through an exhaustive grid search we optimize the hyperparameters of the SVM classifier with a 'radial' kernel. 'C' is 2 and gamma is 0.01. We also optimize the Deep Learning model, testing for the activation function, optimization function, learning rate and number of epochs. The results, including optimized hyperparameters, are shown in Table 2.

Algorithm	Accuracy
Naïve-Bayes	69.13%
Random Forests (trees = 10)	65.40%
Random Forests (trees = 100)	68.72%
Random Forests (trees = 1000)	67.94%
Logistic Regression	69.13%
Decision Trees	61.23%
SVM with kernel = 'radial' + 'C' = 2 + gamma = 0.01	72.17%
Deep Learning with epochs = 100 + activation function = 'relu' + optimization function = rmsprop, learning rate = 0.001	71.14%

Table 2: Results of algorithm testing

All six algorithms perform well, with accuracy ranging from 61.23% to 72.17%. The two highest performing are SVM and Deep Learning (using only a feed forward 'shallow' architecture) – the accuracy of SVM is 72.17%, which outperforms Deep Learning by 1.03 percentage points. Thus, contrary to our initial expectations, we opt to use SVM for the classifier. The performance of SVM and Deep learning algorithms for text classification has long been a point of debate within machine learning (Zaghloul, Lee, & Trimi, 2009). Although Deep Learning has been heralded as the future of machine learning, several recent studies suggest that SVM can outperform it in certain applications (Korba & Arbaoui, 2018; Liu, Choo, Wang, & Huang, 2017). Our result contributes to ongoing discussions in this area.

The SVM hyperparameters are set to maximize generalizability (i.e. low 'C' and gamma values), which make the classifier suitable for empirical applications.

7 Performance

7.1 Cross-validated performance

The classifier consists of model 7 implemented with a tuned SVM. We cross-validate the classifier on the training data set (n = 1,341 tweets) using ten-fold classification. The results are shown in Table 3.

Fold	Accuracy	Balanced accuracy	Precision	Recall	F1 score
1	0.796	0.846	0.795	0.798	0.797
2	0.76	0.808	0.75	0.736	0.743
3	0.736	0.808	0.74	0.75	0.745
4	0.721	0.792	0.714	0.724	0.719
5	0.718	0.774	0.686	0.685	0.686
6	0.746	0.808	0.74	0.742	0.741
7	0.702	0.785	0.699	0.721	0.71
8	0.79	0.845	0.793	0.793	0.793
9	0.756	0.809	0.736	0.736	0.736
10	0.735	0.798	0.733	0.729	0.731
Mean	0.746	0.807	0.739	0.741	0.740

Table 3: Performance of classifier over ten folds

For the accuracy, recall and precision scores (and, as such, F1 scores) we use the macro-aggregation strategy described by Sokolova and Lapalme, in which values are calculated for each class and then the per-class agreement is averaged, with each class treated equally (Sokolova & Lapalme, 2009). The classifier performs similarly for recall and precision (0.741 and 0.739 respectively), and as such has a comparable F1 score (0.74). This is encouraging as it means that the classifier does well at balancing the need to identify relevant instances with minimizing misclassifications, and as such can be applied to real world 'wild' data. We also test for balanced accuracy. This is a relatively new metric put forward by Velez et al. which combines specificity and sensitivity (Velez et al., 2007). They argue that it helps to overcome imbalanced classes and is well-suited to smaller datasets where even small differences in class size can have considerable impact. We report high balanced accuracy (0.807), which provides further evidence that the classifier does well at balancing identifying relevant instances with minimizing misclassifications.

7.2 Performance on unseen data

To check the classifier's performance in 'the wild' we apply it to an unseen dataset of 109,488 tweets produced by 45 far right Twitter accounts during 2017. 100 tweets are randomly sampled from tweets assigned to each of the three classes (None, Weak Islamophobia and Strong Islamophobia) to create a new combined dataset of 300 tweets. This is annotated blind by the three annotators who annotated the original training dataset, using the same annotation guidelines. As before, we take the majority decision to decide the annotation (in 95% of cases all three annotators are in perfect agreement). The results of this testing, as well as how it compares with the previous 10-fold testing, are shown in Table 4. Interestingly, the classifier performs better across all metrics on the unseen data, with accuracy of 77.3%. The uplift in performance, and consistency of the results, indicates the robustness of our approach and its generalizability, which is most likely due to our selection of theoretically-informed input features. Importantly, these results suggest that the classifier is suitable for implementation in empirical research as performance is well above the 70% minimum precision recommended by van Rijsbergen (van Rijsbergen, 1979).

	Accuracy	Balanced accuracy	Precision	Recall	F1 score
Results on unseen data	0.773	0.83	0.778	0.773	0.776
Difference with ten-fold testing	0.027	0.023	0.039	0.032	0.036

Table 4: Performance of classifier on unseen data

The classifier performs well at distinguishing None Islamophobic from Strong Islamophobic. However, it struggles with distinguishing Weak from both Strong and None. For instance, out of 100 tweets which are labelled as Strong Islamophobic, 23 are actually Weak. Similarly, out of 100 predicted Weak Islamophobic tweets, 22 are actually None. This is shown in Figure 1, a contingency table of the classifier's performance on unseen data.

Qualitative investigation of the 300 tweet dataset shows that, in many cases, the None Islamophobic tweets express hatred and prejudice against other groups, such as immigrants. Some also discuss Muslims and Islamic practices but without expressing any negativity. Distinguishing between instances such as these is a challenge as they often have similar input features.

		Predicted Islamophobia			
		None	Weak	Strong	
Actual	None	91	22	4	117
	Weak	8	68	23	99
	Strong	1	10	73	84
		100	100	100	**300**

Figure 1: Contingency table for performance on unseen data

8 Application to far right tweets

To demonstrate the utility of distinguishing between different classes of Islamophobic hate, we show the results of applying the classifier to the 109,488 tweets produced by 45 far right accounts. This is in Figure 2. Noticeably, whilst most tweets are not Islamophobic (57,630 tweets), weak Islamophobia is considerably more prevalent (36,963 tweets) than strong Islamophobia (14,895 tweets). In future empirical research, the classifier could be used to better understand the dynamics of these respective types of Islamophobic hate speech, such as how they fluctuate over time.

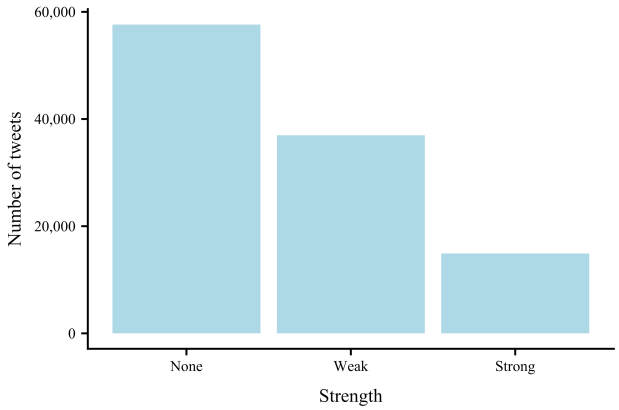

Figure 2: Prevalence of tweets for different strengths of Islamophobia

9 Conclusion

The multi-class Islamophobic hate speech classifier developed in the present work marks an important step forward in developing quantitative methods to provide detailed insight into online Islamophobia. The findings are also relevant for classifying and studying other forms of hate, such as misogyny, racism and anti-Semitism. Whilst more work needs to be undertaken, particularly in making nuanced distinctions between different strengths of hate, the results reported here are promising (particularly, accuracy of 77.3% and balanced accuracy of 83%). In our future work we plan on improving the classifier's performance by increasing the size of the training dataset and engineering additional input features.

REFERENCES

Allen, C. (2010). *Islamophobia*. Surrey: Ashgate.

APPG on British Muslims. (2018). *The many faces of Islamophobia, submission by Tahir Abbas*.

Awan, I. (2016). Islamophobia on social media: A qualitative analysis of the facebook's walls of hate. *International Journal of Cyber Criminology*, *10*(1), 1–20. https://doi.org/10.5281/zenodo.58517

Bayrakli, E., & Hafez, F. (2018). *European Islamophobia Report 2017*. Istanbul: SETA.

Beck, E., Charania, M., & Al-issa, F. A. (2017). Undoing Islamophobia : Awareness of Orientalism in Social Work Undoing Islamophobia : Awareness of Orientalism in Social Work. *Journal of Progressive Human Services*, *28*(2), 58–72. https://doi.org/10.1080/10428232.2017.1310542

Benoit, K., & Matsuo, A. (2018). *R Package: 'spacyr.'* London.

Biran, O., & McKeown, K. (2017). Human-centric justification of machine learning predictions. *IJCAI International Joint Conference on Artificial Intelligence*, 1461–1467. https://doi.org/10.24963/ijcai.2017/202

Bleich, E. (2011). What is islamophobia and how much is there? theorizing and measuring an emerging comparative concept. *American Behavioral Scientist*, *55*(12), 1581–1600. https://doi.org/10.1177/0002764211409387

Burnap, P., & Williams, M. L. (2016). Us and them: identifying cyber hate on Twitter across multiple protected characteristics. *EPJ Data Science*, *5*(11), 2–15. https://doi.org/10.1140/epjds/s13688-016-0072-6

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, *55*(10), 78. https://doi.org/10.1145/2347736.2347755

Dunn, K., & Hopkins, P. (2016). The Geographies of Everyday Muslim Life in the West The Geographies of Everyday Muslim Life in the West. *Australian Geographer*, *47*(3), 255–260. https://doi.org/10.1080/00049182.2016.1191138

Ekman, M. (2015). Online Islamophobia and the politics of fear : manufacturing the green scare Online Islamophobia and the politics of fear : manufacturing the green scare. *Ethnic and Racial Studies*, *38*(11), 1986–2002. https://doi.org/10.1080/01419870.2015.1021264

Facebook. (2018). *Facebook: Community Standards*. *Facebook*. San Francisco. Retrieved from https://secondlife.com/corporate/cs.php

Feuerriegel, S., & Proellochs, N. (2018). *R Package: 'SentimentAnalysis.'* London.

Gallie, W. B. (1956). Essentially Contested Concepts. *Proceedings of the Aristotelian Society*, *56*, 167–198.

Haque, A., Tubbs, C. Y., Kahumoku-Fessler, E. P., & Brown, M. D. (2018). Microaggressions and Islamophobia: Experiences of Muslims Across the United States and Clinical Implications. *Journal of Marital and Family Therapy*, *44*(2).

HM Government. (2012). *Deputy Prime Minister extends funding to tackle hate crime against Muslims*. London.

Home Affairs Select Committee. (2017). *Home Affairs Committee Hate crime: abuse, hate and extremism online*. London.

Hope Not Hate. (2015). *Hope Not Hate: State of Hate 2015*. London: Hope Not Hate.

Hope Not Hate. (2017). *Hope Not Hate: State of Hate 2017*. London: Hope Not Hate.

Hsu, C. W., & Lin, C. J. (2002). A comparison of methods for multiclass support vector machines. *IEEE Transactions on Neural Networks*, *13*(2), 415–425. https://doi.org/10.1109/72.991427

Husain, A., & Howard, S. (2017). Religious Microaggressions : A Case Study of Muslim Americans Religious Microaggressions : A Case Study of Muslim. *Journal of Ethnic & Cultural Diversity in Social Work*, *26*(1–2), 139–152. https://doi.org/10.1080/15313204.2016.1269710

Imhoff, R., & Recker, J. (2012). Differentiating Islamophobia: Introducing a New Scale to Measure Islamoprejudice and Secular Islam Critique, *33*(6). https://doi.org/10.1111/j.1467-9221.2012.00911.x

Ingham-Barrow, I. (2018). *More than words: approaching a definition of Islamophobia*. (I. Ingham-Barrow, Ed.). London: MEND.

Ipsos MORI. (2016). *Attitudes to potentially offensive language on TV and radio*. London. Retrieved from http://www.ipsos-mori.com/terms.

Jackson, L. R. (2018). *Islamophobia in Britain: the making of a Muslim enemy*. London: Palgrave Macmillan UK.

Korba, K. A., & Arbaoui, F. (2018). SVM Multi-Classification of Induction Machine's bearings defects using Vibratory Analysis based on Empirical Mode Decomposition. *International Journal of Applied Engineering Research*, *13*(9), 6579–6586.

Kotsiantis, S. B. (2007). Supervised Machine Learning: A Review of Classification Techniques. *Informatica*, *31*, 249–268. https://doi.org/10.1115/1.1559160

Kumar, R., Ojha, A. K., Malmasi, S., & Zampieri, M. (2018). Benchmarking Aggression Identification in Social Media. *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying*, 1–11. Retrieved from http://aclweb.org/anthology/W18-4401

Kunst, J. R., Sam, D. L., & Ulleberg, P. (2013). International Journal of Intercultural Relations Perceived islamophobia : Scale development and validation. *International Journal of Intercultural Relations*, *37*(2), 225–237. https://doi.org/10.1016/j.ijintrel.2012.11.001

Lai, S., Liu, K., He, S., & Zhao, J. (2016). How to generate a good word embedding. *IEEE Intelligent Systems*, *31*(6), 5–14. https://doi.org/10.1109/MIS.2016.45

Liu, P., Choo, K. R., Wang, L., & Huang, F. (2017). SVM or deep learning ? A comparative study on remote sensing image classification. *Soft Computing*, *21*(23), 7053–7065. https://doi.org/10.1007/s00500-016-2247-2

Malmasi, S., & Zampieri, M. (2017). Detecting Hate Speech in Social Media. *ArXiv Preprint:1712.06427v2*. https://doi.org/10.26615/978-954-452-049-6_062

Meer, N., & Modood, T. (2009). Patterns of Prejudice Refutations of racism in the ' Muslim question .' *Patterns of Prejudice*, *43*(3), 335–354. https://doi.org/10.1080/00313220903109250

Moosavi, L. (2015). The Racialization of Muslim Converts in Britain and Their Experiences of Islamophobia. *Critical Sociology*, *41*(1), 41–56. https://doi.org/10.1177/0896920513504601

Nadal, K. L., Griffin, K. E., Hamit, S., Leon, J., & Rivera, D. P. (2012). Subtle and Overt Forms of Islamophobia : Microaggressions toward Muslim Americans, *VI*(2), 15–37.

Park, J. H., & Fung, P. (2017). One-step and Two-step Classification for Abusive Language Detection on Twitter. https://doi.org/10.18653/v1/W17-3006

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. https://doi.org/10.3115/v1/D14-1162

Pettigrew, T. F., & Meertens, R. W. (1995). Subtle and blatant prejudice in Western Europe. *European Journal of Social Psychology*, *25*(1), 57–75.

Runnymede Trust. (2017). *Islamophobia: still a challenge for us all*. London.

Saleem, H. M., Dillon, K. P., Benesch, S., & Ruths, D. (2017). A Web of Hate: Tackling Hateful Speech in Online Social Spaces. Retrieved from http://arxiv.org/abs/1709.10159

Salminen, J., Almerekhi, H., Milenkovi, M., Jung, S., An, J., Kwak, H., & Jansen, B. J. (2018). Anatomy of Online Hate: Developing a Taxonomy and Machine Learning Models for Identifying and Classifying Hate in Online News Media. *12th International AAAI Conference on Web and Social Media (ICWSM)*, (Icwsm), 330–339.

Schmidt, A., & Wiegand, M. (2017). A Survey on Hate Speech Detection using Natural Language Processing. *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, (2012), 1–10. https://doi.org/10.18653/v1/W17-1101

Silva, L., Mondal, M., Correa, D., Benevenuto, F., & Weber, I. (2016). Analyzing the Targets of Hate in Online Social Media, (June). Retrieved from http://arxiv.org/abs/1603.07709

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing and Management*, *45*(4), 427–437. https://doi.org/10.1016/j.ipm.2009.03.002

Stanford. (2018). Stanford GloVe. Retrieved October 12, 2018, from https://nlp.stanford.edu/projects/glove/

Tell Mama. (2018). *Beyond the incident: outcomes for victims of anti-Muslim prejudice*. London. https://doi.org/10.1053/j.jvca.2010.06.032

Twitter. (2018). Twitter: Docs. Retrieved October 17, 2018, from https://developer.twitter.com/en/docs.html

van Rijsbergen, C. J. (1979). *Information Retrieval*. London: Butterworths.

Velez, D. R., White, B. C., Motsinger, A. A., Bush, W. S., Ritchie, M. D., Williams, S. M., & Ã, J. H. M. (2007). A Balanced Accuracy Function for Epistasis Modeling in Imbalanced Datasets using Multifactor Dimensionality Reduction, *315*, 306–315. https://doi.org/10.1002/gepi

Wainer, J. (2016). Comparison of 14 different families of classification algorithms on 115 binary datasets, (2014). Retrieved from http://arxiv.org/abs/1606.00930

Wang, S., & Manning, C. (2012). Baselines and Bigrams: Simple, Good Sentiment and Topic Classification. *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, (July), 90–94.

Waseem, Z., & Hovy, D. (2016). Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. *Proceedings of the NAACL Student Research Workshop*, 88–93. https://doi.org/10.18653/v1/N16-2013

Williams, M. L., & Burnap, P. (2016). Cyberhate on Social Media in the aftermath of Woolwich: A Case Study in Computational Criminology and Big Data. *British Journal of Criminology*, *56*(2), 211–238. https://doi.org/10.1093/bjc/azv059

Zaghloul, W., Lee, S. M., & Trimi, S. (2009). Text classification: neural networks vs support vector machines. *Industrial Management & Data Systems*, *109*(5), 708–717.