Detecção de Discurso de Ódio: Homofobia

Andrey O. Souza[1,2], **Eduardo F. Nakamura**[1], **Fabíola G. Nakamura**[1]

[1]Instituto de Computação – Universidade Federal do Amazonas (UFAM)
69080-900 – Manaus – AM – Brasil

[2]Sidia Instituto de Ciência e Tecnologia
69055-035 – Manaus – AM – Brasil

{aos,nakamura,fabiola}@icomp.ufam.edu.br;andrey.souza@sidia.com

Abstract. *Understanding the context and reasons for homophobic actions is primordial to mitigate this type of hatred. This work presents the construction of a dataset taken from Twitter, which contains information on homophobic discourses. The contributions are: (1) the method of building a dataset and anonymous labeling based on their homophobic content; (2) the creation of features of this dataset; (3) the evaluation of machine learning methods for the classification of data regarding homophobic content. Preliminary results show that the Random Forest Classifier model stands out in the identification of homophobic tweets with F1-score of 0.8, recall of 0.9 and precision of 0.7.*

Resumo. *Entender o contexto e os motivos para ações homofóbicas é primordial para mitigar esse tipo de ódio. Este trabalho apresenta a construção de um conjunto de dados retirados do Twitter, os quais contém informações sobre discursos homofóbicos. As contribuições são: (1) o método de construção de um conjunto de dados e a rotulação anônima baseada no teor homofóbico dos mesmos; (2) a criação de características desse conjunto de dados; (3) a avaliação de métodos de aprendizagem de máquina para a classificação dos dados referente ao teor homofóbico. Resultados preliminares mostram que o modelo Random Forest Classifier se destaca na identificação de tuítes homofóbicos com F1-Score de 0.8, revocação de 0.9 e precisão de 0.7.*

1. Introdução

O acesso à internet modificou o mundo e as maneiras como nos relacionamos com ele. Tais relações ocorrem por meio das redes sociais, ferramentas que visam conectar os usuários da internet. Cada rede social tem sua forma de mediar essa conexão, como compartilhamento audiovisual, busca por relacionamentos, expressão de opiniões, contatos profissionais, compra e venda, entre outros. Essas conexões se dão em um período de tempo instantâneo, fato que contribui para o crescimento do número de pessoas que aderem às ferramentas online. No início do ano de 2021, o número de usuários de redes sociais totalizava 4.20 bilhões após um crescimento de 490 milhões de usuários em 12 meses, o que equivalia a mais de 53% da população mundial [Amper 2021].

Dentre as redes sociais ativas, uma das maiores e mais poderosas é o Twitter, o qual é conhecido pelos seus *posts* rápidos contendo mensagens curtas (280 caracteres por mensagem), e em sua maioria, expressando pensamentos e opiniões de seus autores. Em 2021, o Twitter possuía 353 milhões de usuários ativos, dos quais uma parcela de 187

milhões acessava essa rede social diariamente. Dessa parcela, 80% dos usuários estavam localizados nos Estados Unidos da América [Affde 2021].

A capacidade de expressão de pensamentos rápidos do Twitter acaba sendo utilizada como palco para falas ofensivas ou discursos de ódio. A diferenciação dessas duas classes existe em sutis detalhes linguísticos, assim como apontam [Davidson et al. 2017] no seu trabalho. Os termos *fag*, *bitch* e *nigga* são encontrados nas duas classes, enquanto *faggot* e *nigger* são normalmente associados ao discurso de ódio. O Plano de Ação Rabat das Nações Unidas [Council 2013] define diretrizes para distinguir a liberdade de expressão, a linguagem ofensiva e o discurso de ódio, sendo delimitadas entre três tipos de expressão: "expressão que constitua uma infracção penal; expressão não punível criminalmente, mas passível de ação civil ou sanções administrativas; expressão que não dá lugar a sanções penais, civis ou administrativas, mas suscita preocupação em termos de tolerância, civilidade e respeito pelos direitos dos outros.".

A homofobia por sua vez, existe em um subespaço do que engloba o discurso de ódio. Em seu trabalho, [Cazelatto and Cardin 2016] aponta a homofobia como um ato de violência que vai além da agressão física ou psicológica, representando inúmeros comportamentos de repressão, exploração e dominação, que têm como objetivo depreciar pessoas homossexuais a uma condição sub-humana, desprovida de interesses ou direitos. Portanto, podemos afirmar que o discurso de ódio homofóbico se caracteriza como a prática da homofobia através de um veículo linguístico: o ato discursivo.

Como visto anteriormente, o ato discursivo em sua forma escrita vem sendo manifestado fortemente através das redes sociais, o que resulta em um grande número de dados textuais abertos à consulta. Com isso, o crescente interesse em utilizar modelos de classificação textual para detecção de discurso de ódio, especialmente em redes sociais, fez com que muitos trabalhos acerca disso fossem produzidos. Este trabalho tem como objetivo detectar discursos de ódio homofóbicos em *posts* do Twitter, além de disponibilizar um conjunto de dados rotulados e analisar a performance de diferentes abordagens de detecção para esse contexto.

2. Trabalhos Relacionados

A expansão das redes sociais está diretamente ligada ao poder de comunicação, interação e expressão da atual sociedade. [Bermingham and Smeaton 2011] defendem que todas essas interações geram dados com um poderoso poder preditivo, principalmente dados oriundos do Twitter, os quais expressam os sentimentos e as ideias dos usuários no que diz respeito a uma ampla variedade de tópicos.

Baseados nesses dados, vários trabalhos de predição vêm sendo desenvolvidos, porém, [Kwok and Wang 2013] e [Burnap and Williams 2015] apontam que muitas das classificações de tuítes como discurso de ódio são na verdade linguagem ofensiva, o que traz imprecisão a essas classificações. As diferenças linguísticas que distinguem discurso de ódio e linguagem ofensiva dependem muito do contexto [Kwok and Wang 2013].

Atento a isso, [Watanabe et al. 2018] propõem uma abordagem de classificação de tuítes em odiosos, ofensivos ou limpos. A abordagem se baseia em padrões de discurso de ódio, unigramas e características semânticas e sentimentais. O trabalho utilizou o algoritmo *J48graft* na sua conclusão, o qual atingiu 87,4% de precisão para a classificação

binária entre tuítes ofensivos e não ofensivos, ao passo que para a classificação ternária de tuítes entre odiosos, ofensivos e limpos atingiu a precisão de 78,4%. O trabalho não atuou sobre um subgrupo do discurso de ódio, porém, classificou os discursos em 3 classes.

[Davidson et al. 2017] seguiram uma abordagem de categorização baseada em palavras e em características de cada tuíte, como *hashtags*, URL, *retweets* e sentimentos. Os autores utilizaram esses recursos para analisar os desempenhos dos seguintes modelos de aprendizagem: regressão logística, *Naives Bayes*, árvores de decisão, *Random Forest* e SVM linear. O resultado da análise apontou que os modelos de regressão logística e SVM lineares obtiveram melhores desempenhos, sendo a regressão logística o modelo de conclusão escolhido. O modelo resultou em um F1 score de aproximadamente 0,90, contudo, houveram erros de classificação em quase 40% nos casos de discurso de ódio.

Seguindo outra metodologia, [Pereira-Kohatsu et al. 2019], em parceria com o Escritório Nacional Espanhol contra Crimes de Ódio, desenvolveram um modelo de detecção de discurso de ódio chamado *HaterNet*, o qual se baseia em um aprendizado duplo profundo, que combina uma rede neural LSTM (*Long short-term memory*) e MLP (*Multilayer perceptron*) e características de frequência. Para isso, eles compararam 19 estratégias de combinação de características e modelos de classificação, e em todos os casos, o *HaterNet* obteve o melhor desempenho com um F1 score máximo de 0,906 quando utilizado com os mesmos dados do modelo em comparação. O escopo do trabalho também não é definido sob um subgrupo de discurso de ódio, sendo essa uma extensão futura.

Em diferenciação a este trabalho, os trabalhos anteriormente mencionados não direcionam os esforços para uma das subclasses do discurso de ódio, porém, mostram metodologias e abordagens promissoras para a detecção de falas odiosas. A Tabela 1 mostra um comparativo entre os trabalhos relacionados:

Tabela 1. Comparação dos trabalhos relacionados em termos de características, modelos, rótulos e seus respectivos resultados.

Trabalho	Características	Modelos	Rótulos	Resultados
Watanabe, et al. 2018	Bag of Words e Unigramas	J48graft	Clean Offensive Hateful	Recall 0,784 F1 Score 0,784 Precision 0,793
Davison et al. 2017	Frequency representation	Regressão Logística	Clean Offensive Hateful	Recall 0,90 F1 Score 0,90 Precision 0,91
Pereira-Kohatsu, et al. 2019	N-gramas com tamanhos de 1 a 3	LSTM + MLP	Clean Hateful	Recall 0,890 F1 Score 0,906 Precision 0,925

Todos esses trabalhos se baseiam nas expressões dos usuários das redes sociais, expressões essas que tomam forma em sua maioria por meio de palavras. [Miriam J. 2002] explica que todas as palavras são símbolos, exceto algumas imitativas ou onomatopeias, como: *zumzum, ping-pong, tchu etc*. Esses símbolos são classificados em dois grandes grupos: símbolos especiais e símbolos comuns. Símbolos especiais expressam com precisão ideias de um campo específico do conhecimento, além de serem internacionais e livres de tradução. Símbolos comuns são criados para comunicar necessidades corriqueiras do curso da vida, além de serem menos precisos e passíveis à tradução. Portanto, o significado de um símbolo comum é imposto por convenção, logo, ele está

sujeito à ambiguidade, visto que mais de um significado pode ser dado a um mesmo símbolo. Uma palavra se torna ambígua quando é imposto sobre o símbolo intenções diferentes ou eventos recorrentes da história.

Para compreender a construção da palavra "homofobia", é necessário conhecer a origem do termo e da ideia de homofobia. Em 1972, [Weinberg 1975] escreveu no prefácio de seu livro que a doença mental na verdade é a homofobia, e não a homossexualidade, como era vista pela maioria dos psicólogos da época. Porém, ao contrário do que indica a conotação patológica do sufixo "fobia" [Chamberland and Lebreton 2012], a palavra assumiu um significado mais próximo do ódio e aversão a pessoas homossexuais do que de medo a essa classe de pessoas. Tal aversão é baseada na heterossexismo, o que defende a premissa de que a heterossexualidade é superior e normal, e que baseado nessas normas a sociedade deve funcionar. [Chambers 2007] diz que essa normatividade gera práticas regulatórias que produzem e restringem a inteligibilidade de gênero, o que facilita a expressão de alguns gêneros enquanto restringe outros.

Buscando uma definir o contexto onde a homofobia ocorre, [Fraïssé and Barrientos 2016] analisam em seu trabalho a homofobia de uma perspectiva psicossocial, comparando conceitos, formas de medição para possíveis casos de homofobia e como as variáveis dos contextos onde ocorrem homofobia interagem entre si, formando um sistema homofóbico. A heteronormatividade, o sexismo, o preconceito sexual e a dominação masculina são as variáveis que interatuam nesse sistema homofóbico. Com isso, é possível representar de forma mais diversa e abrangente o contexto onde há ou não a homofobia.

3. Material e Métodos

Este trabalho é constituído das seguintes etapas do processo de classificação de texto:

3.1. Coleta de Dados

Os dados coletados para este trabalho foram adquiridos do Twitter via API Rest. As palavras *"fag"*, *"faggot"* e *"bigotry"* foram utilizadas como parâmetros para a API por serem símbolos geralmente utilizados em discursos voltados para pessoas homossexuais. Essa coleta foi executada de Janeiro de 2021 a Abril de 2021, adquirindo um total de 72919 tuítes limitados à língua inglesa. Tal limitação se deu com base na pesquisa de [Statista 2021], a qual aponta o maior uso do Twitter por norte americanos.

3.2. Análise e Filtragem dos Dados

Finalizada a coleta dos dados, iniciou-se o processo de análise e filtragem desses. Dado que o Twitter é uma rede social que permite a publicação de conteúdo adulto, foi necessária uma filtragem de tuítes que continham essa finalidade, pois uma linguagem violenta pode ser usada em âmbitos de conteúdo adulto de forma consensual. A filtragem desses tuítes resultou em um conjunto de dados composto por 62681 tuítes não rotulados.

3.3. Rotulagem

Os tuítes filtrados foram então submetidos a uma rotulagem manual, o que se deu por meio de um site construído para esta finalidade. Usuários pelo mundo podem acessar o site pelo endereço `https://bit.ly/analiseHomofobia` e opinar se o tuíte exibido

possui teor homofóbico. Para auxiliar na decisão, é exibido o conceito de homofobia, um exemplo de um tuíte homofóbico e outro exemplo de um tuíte não homofóbico.

A rotulagem não se resume apenas à classe que foi atribuída a um tuíte, mas também a quem rotula. Conhecer quem rotula possibilita ter um vislumbre das motivações para as escolhas feitas. Embasados nisso, os usuários do site de rotulação podem inserir informações sobre sua sexualidade, identidade de gênero, estado civil e tempo diário gasto utilizando o Twitter, as quais são relevantes para o entendimento do contexto.

Um total de 105 pessoas se interessaram em opinar sobre os tuítes, o que resultou em 1558 tuítes rotulados entre as classes homofóbico, não homofóbico e incerto. A inclusão da classe incerto foi incluída tendo em vista que uma opinião pessoal possui nuances complexas de contexto e linguística, as quais podem causar dúvidas a quem lê. Por fim, a classe final de um tuíte é decidida com base nos votos majoritários atribuídos a ele. Os dados rotulados nessa etapa podem ser acessados pelo link `tinyurl.com/dados-homofobia`.

A Figura 1 demonstra a distribuição dos rotuladores por sexualidade. Nota-se que houve maior interesse por parte de pessoas homossexuais, porém a distribuição entre homossexuais, heterossexuais e bissexuais está próxima. A Figura 2 apresenta a distribuição de frequência de votos de cada classe por sexualidade dos rotuladores. Nota-se nos dados da figura 2 que as proporções de votos para cada classe são aproximadas, mantendo maior número de votos por rotuladores homossexuais em todos os casos.

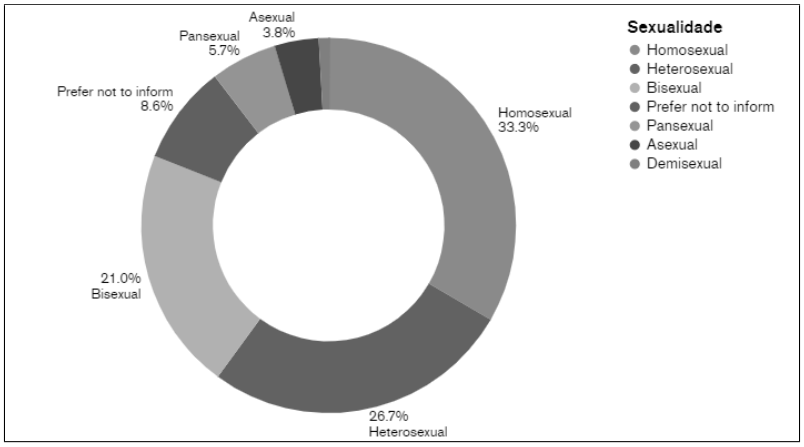

Figura 1. Sexualidade dos rotuladores.

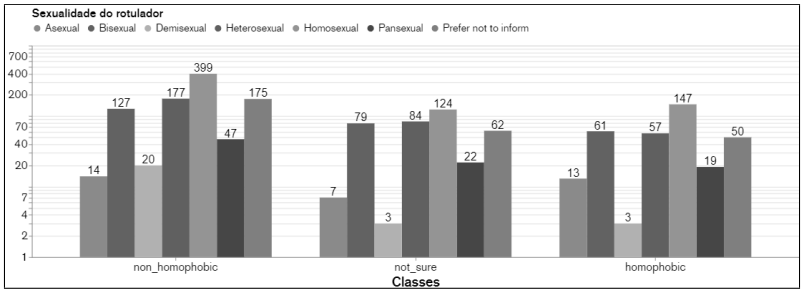

Figura 2. Distribuição de frequência de votos de cada classe por sexualidade dos rotuladores.

3.4. Construção de Características

Junto aos textos dos tuítes foram adicionados ao conjunto de dados a quantidade de caracteres não alfabéticos por tuíte, a quantidade de caracteres totais por tuíte e a análise de sentimentos sobre o texto do tuíte, a qual foi calculada via biblioteca nltk do *python*. Tais informações como o número de retuítes e número de *likes* não foram consideradas por serem dados ausentes em muitos tuítes do conjunto de dados obtidos ao início.

O conteúdo dos tuítes foi filtrado para a retirada das *stop words*, as quais são palavras que aparecem constantemente no texto e que não acrescentam informações relevantes ao contexto. Tais palavras normalmente são conjunções, pronomes, preposições *etc*. Com isso, as palavras mais frequentes nos tuítes se tornam termos com maior relevância para a análise do contexto e do teor do tuíte.

A Figura 3 mostra os termos mais utilizados pelos tuítes rotulados como homofóbicos. Nota-se que a utilização de palavras consideradas pejorativas para pessoas homossexuais como *"faggot"* e *"fag"* são as mais utilizadas, seguidas de *"like"* e *"fuck"* que são palavras utilizadas para proferir comparações e xingamentos. Podemos tomar como exemplo da utilização dessas palavras com um teor homofóbico o seguinte tuíte: *"I HATE faggots cause theyre like a cancer!!"*.

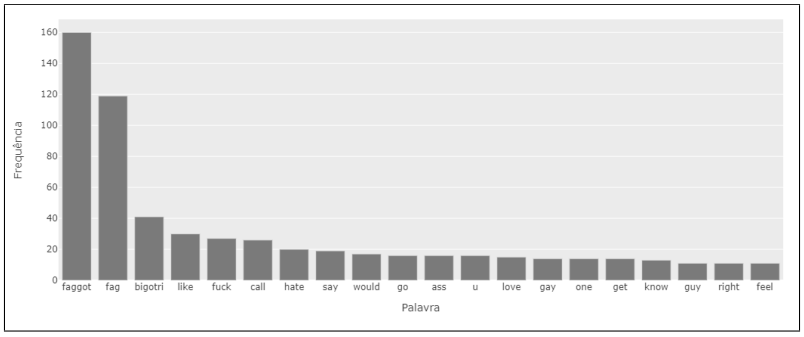

Figura 3. Frequência das 20 palavras mais utilizadas em tuítes homofóbicos.

Além disso, foram retirados dos textos dos tuítes dos conjuntos de dados ngramas de tamanho de 1 a 3 caracteres, além da frequência das palavras de cada tuíte. Para isso foi utilizado o *Term Frequency - Inverse Dense Frequency (TF-IDF)* que consiste em uma técnica utilizada para determinar um peso para a semântica de sentenças, no caso deste trabalho, os tuítes. Essa técnica foi aplicada para cada palavra de um tuíte e para cada *caracter* de cada palavra de um tuíte.

3.5. Treinamento do Modelo de Classificação

Para a etapa de treinamento do modelo de classificação foram utilizados os métodos *Support Vector Machine Linear (SVM Linear), XGBoost (XGB), Logistic Regression (LR) e Random Forest Classifier (RFC)*, todos recebendo as características extraídas na etapa anterior como fonte de dados.

Tendo em vista o pequeno número de tuítes rotulados, foi utilizada a validação cruzada *Leave-One-Out (LOOCV)* na fase de teste e validação, técnica que faz uso de apenas uma observação na fase de validação enquanto o restante dos dados são utilizados para treinar o modelo.

4. Resultados e Discussão

A Tabela 2 mostra que o modelo RFC obteve melhor desempenho.

Tabela 2. Relatório de classificação dos modelos avaliados.

Modelo	Classe	Precisão	Revocação	F1-Score
SVM Linear	Homofóbico	0,47	0,61	0,53
	Não Homofóbico	0,89	0,82	0,86
XGB	Homofóbico	0,58	0,61	0,60
	Não Homofóbico	0,87	0,85	0,86
LR	Homofóbico	0,14	0,81	0,23
	Não Homofóbico	0,99	0,76	0,86
RFC	Homofóbico	0,72	0,93	0,81
	Não Homofóbico	0,98	0,91	0,94

Dos resultados obtidos, observa-se que combinações entre as características poderiam gerar outros resultados, os quais somariam nas conclusões finais. Além disso, outros modelos poderiam ser submetidos a testes para comparações com os modelos atuais.

5. Conclusões e Perspectivas

Neste trabalho foi construído um conjunto de dados rotulados separados entre homofóbicos, não homofóbicos e incerto. Além disso, foram extraídas dos dados características que auxiliam na análise dos votos. Com isso, treinamos modelos de aprendizagem de máquina para identificar quando um tuíte possui teor homofóbico ou não.

Algumas observações feitas neste trabalho mostraram que o interesse em projetos deste cunho é maior na comunidade a que o trabalho se refere, nesse caso, indivíduos homossexuais. Ainda, mostraram que as palavras mais utilizadas de forma odiosa são símbolos que têm como objetivo utilizar a condição de uma pessoa homossexual como um xingamento, motivo de vergonha ou diminuição do valor do alvo.

Alguns aprimoramentos que podem ser executados posteriormente foram identificados, sendo a inclusão de rotuladores com sexualidades mais diversas o principal deles, pois a opinião de cada indivíduo muda de acordo com o seu contexto vivido. Além disso, explorar outras características e combinações entre elas pode mostrar novos resultados possivelmente mais precisos. Levantando resultados melhores e maior diversidade de rotuladores abrimos novas perspectivas sobre um mesmo problema, a homofobia.

Agradecimentos

O trabalho é parcialmente suportado pelo CNPq através do Processo 312175/2021-3. O trabalho também é parcialmente suportado pelo Projeto Samsung-UFAM de Ensino e Pesquisa (SUPER), nos termos do artigo 48 do Decreto nº 6.008/2006 (SUFRAMA), foi parcialmente financiada pela Samsung Eletrônica da Amazônia Ltda., nos termos da Lei Federal nº 8.387/1991, por meio dos convênios 001/2020 e 003/2019, firmados com a Universidade Federal do Amazonas e a FAEPI, Brasil. O trabalho também é decorrente do projeto de Pesquisa e Desenvolvimento (P&D) Sidia, que conta com financiamento da Samsung, usando recursos da Lei de Informática para a Amazônia Ocidental de acordo com o artigo 39º do Decreto nº 10.521/2020.

Referências

Affde (2021). Quantas pessoas usam o twitter em 2021? [novas estatísticas do twitter]. https://www.affde.com/pt/twitter-users.html.

Amper (2021). We are social e hootsuite - digital 2021 [resumo e relatório completo]. https://tinyurl.com/we-are-social-e-hootsuite.

Bermingham, A. and Smeaton, A. (2011). On using Twitter to monitor political sentiment and predict election results. In *Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology (SAAIP 2011)*, pages 2–10, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

Burnap, P. and Williams, M. L. (2015). Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & Internet*, 7(2):223–242.

Cazelatto, C. and Cardin, V. S. G. (2016). O discurso de ódio homofóbico no brasil: um instrumento limitador da sexualidade humana. *Revista Jurídica Cesumar-Mestrado*, 16.

Chamberland and Lebreton (2012). *Réflexions autour de la notion d'homophobie : succès politique, malaises conceptuels et application empirique.* Nouv Questions Feministes.

Chambers, S. A. (2007). An incalculable effect': Subversions of heteronormativity. *Political studies*, 55.

Council, H. R. (2013). Human rights. annual report of the united nations high commissioner for human rights. https://tinyurl.com/rabat-united-nations.

Davidson, T., Warmsley, D., Macy, M. W., and Weber, I. (2017). Automated hate speech detection and the problem of offensive language. *CoRR*, abs/1703.04009.

Fraïssé, C. and Barrientos, J. (2016). The concept of homophobia: A psychosocial perspective. *Sexologies*, 25.

Kwok, I. and Wang, Y. (2013). Locate the hate: Detecting tweets against blacks. In *AAAI*.

Miriam J., M. M. (2002). *The trivium: The liberal arts of logic, grammar, and rhetoric : understanding the nature and function of language.* Paul Dry Books.

Pereira-Kohatsu, J. C., Quijano-Sánchez, L., Liberatore, F., and Camacho-Collados, M. (2019). Detecting and monitoring hate speech in twitter. *Sensors*, 19(21).

Statista (2021). Leading countries based on number of twitter users as of october 2021. https://tinyurl.com/statista-twitter-users.

Watanabe, H., Bouazizi, M., and Ohtsuki, T. (2018). Hate speech on twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection. *IEEE Access*, 6:13825–13835.

Weinberg, G. H. (1975). *Society and the healthy homosexual / [by] George Weinberg*. Gerrards Cross, Smythe.