Literature Review: Bias in Machine Learning With A Focus On Natural Language Processing And Face Recognition

Priya Jain

Nicolas Lab

Center for Cognitive Science, Rutgers University

Abstract

The application of machine learning is allowing humans to do remarkable things that could not be done before. Predictive decision making, facial recognition in criminal law, self-driving vehicles are all examples of where machine learning can take us over the coming years. A severe threat to the validity of machine learning algorithms and systems is the inherent biases deep-rooted in society. The issue is that these prejudices are captured and represented through machine learning models and then further perpetuates biases and stereotypes in society. Breaking from this cycle is very crucial to having fair and valid models that humans can trust to make decisions. This paper will discuss research on biases in natural language processing and face recognition, social and ethical implications, as well as recent algorithmic debiasing solutions and methodologies.

Ethical Artificial Intelligence and User Trust

Before delving into specifically natural language processing and face recognition, a general discussion of ethical artificial intelligence (AI) is pertinent. There are multiple components that factor into the ethical AI dialogue. One critical cognitive mechanism of human-computer interactions is trust. As defined by Jacovi et al. (2020), Human-AI trust is when humans perceive the AI model as trustworthy to the contract (any explicit agreement on what rules and standards the AI developer must abide by) and accepts vulnerability to the AI model's actions in uncertain situations. Trust contracts between users and developers usually encompass data protection, accountability, transparency, diversity, and/or fairness (Jacovi et al., 2020). All of these categories are paramount to consider when designing and building AI and machine learning systems. Explainable AI has become a principal point of conversation in order to

facilitate trust and transparency between AI users and developers (Bartneck et al., 2020). The idea is that machine learning algorithms should be transparent and explainable so that humans can evaluate it, build trust, and make sure all the processes are moral as opposed to building black box algorithms that may be arriving at wrongful conclusions. In this paper, the focus will be on the fairness and diversity components of ethical AI, particularly in machine learning.

Biases in Word Embeddings

Natural language processing is a subset of machine learning that studies how computers can understand human language. Word embeddings provide an approach to represent unstructured text data with vectors. Each word is mapped to a vector in a neural network. The vectors between words represent a relationship so that word vectors closer together have similar semantic meaning (Bolukbasi et al., 2016). Word embeddings can shed light on racial and gender biases that are implicitly captured in text data. In this context, bias means prejudice in society that is reproduced by machine learning models. For example, some words are gender specific like “sister” which will objectively be closer to “woman” than “man”. However, there are some words that are not gender specific and therefore should be equidistant between “man” and “woman” relaying there is no gender direction. The relationship man : woman :: computer programmer : homemaker implies an implicit gender bias that men and are semantically closer in meaning to computer programmers while women are semantically closer to homemakers (Bolukbasi et al., 2016). This wrongful relationship is produced because data shows there are more male computer programmers than female computer programmers; the model learns this and develops a word embedding that reflects this. Biases like these are so deeply ingrained in human

culture that it cannot be easily separated from the text's objective semantic meaning (Caliskan et al., 2017).

Types of Biases

Before delving into how natural language processing encodes biases and exploring debiasing solutions, it is worthwhile to discuss what bias means in the context of machine learning. In machine learning and artificial intelligence, bias just means prior information. Bias becomes problematic when it is derived from a wrongful precedent causing misrepresentation of information. This can be described as prejudice (Caliskan et al., 2017). There are different kinds of biases and prejudices that present themselves latently in text. One is demographic bias in training data; this occurs when it is assumed that all languages are identical causing lowered performance for other demographics. This assumption is insensitive to certain groups, can lead to demographic exclusion, reduce machine objectivity, and make technology unfriendly for some demographics. One way to algorithmically solve this problem is by downsampling the over represented group in the training data (Hovy & Spuit, 2016). By doing this, the data will be more balanced and equally represent all demographic groups. In fact, a major driver of bias is the quality of training data (Wang & Deng, 2020). In order to prevent under informed training, Caliskan et al. (2017) recommends having diversity among AI developers to create inclusive algorithms. The hypothesis is that if the team of developers are diverse in terms of gender, race, color, etc., then those unique experiences can inform inclusive training, testing, and comprehensive algorithms.

Another source of bias is overgeneralization of false positives, which is a modelling side effect. This is when models produce false positives like receiving an automatic email that

misgenders the recipient or congratulates them on retirement when it's actually their 30th birthday. Although these errors may seem mild, if these errors are overgeneralized in a model, it can create less milder problems like incorrectly identifying someone's religious view or sexual orientation. A way to algorithmically address this is by error weighting and the use of confidence thresholds (Hovy & Spuit, 2016).

Overexposure of topics can also lead to discrimination. Humans follow the availability heuristic, which says information that is easier to recall or more prevalent will be considered more important (Tversky and Kahneman, 1973). This becomes particularly problematic when the heuristic becomes ethnically charged. An example is hearing a lot of news about violence by a certain ethnic group. Just because negative news is amplified more, negative emotions may be associated with this group (Slovic et al., 2007). Overexposure is different from oversampling because oversampling is a data problem, while overexposure is a psychological affair. The solution to overexposure problem is not algorithmic but begs deeper analysis into research design and intrinsic social biases fused with natural language. According to Hovy & Spuit (2016) it is imperative to design research questions in a way that doesn't feed existing biases and doesn't overexpose certain populations.

Power of Word Embeddings and Sentiment Analysis

It is important to recognize that word embeddings not only capture social biases but have the power to perpetuate existing cultural biases even further. Bolukbasi et al.(2016) eloquently presents an example of this by imagining a search query for computer science PHD students at a certain university. Let's say the directory has a 100 student web pages of identical content and relevance except for differing names. Because computer science terms are closely related to

"man" and male names, male student web pages would be ranked higher than a woman's web page that is of equal relevance. This example of direct bias makes it continuously more difficult for female computer scientists to be recognized for their work because their pages are always ranked lower and contributes to the ongoing gender gap in the field of computer science (Bolukbasi et al., 2016). This reaffirms that eliminating bias in word embeddings and NLP can reduce bias in society.

One question that may arise is how does one know that these prejudiced associations between word vectors are not in fact due to chance. Caliskan et al. confirms that associations between word vectors reflect cultural biases (2017). The authors define a score, WEAT, for evaluating wrongful correlations with sentiment for different demographics in text data. This study replicates findings from the Implicit Association Test (IAT) (Greenwald et al., 1998) using the GloVe word embedding (Pennington et al., 2014), a machine learning model trained on a corpus of text from the World Wide Web. The IAT is a psychological test that uncovered how similarly or differently people associate two words depending on the differences in response times. It was found that when words were strongly associated, the response time was faster (Greenwald et al., 1998). Successfully replicating the results of the 1998 study with a word embedding model shows that the associations between word vectors match up to human biases and stereotypes. In the Caliskan et al. study, shorter vector distances would mean semantic nearness just as shorter response times demonstrated semantic similarly in the IAT. This study demonstrated that European American names are associated with "pleasant" terms like "love", "peace", "cheer", "friend", "loyal" and "honest" at a higher rate than African American names are.

Another key result is replicating the results of Bertrand & Mullainathan (2004) Resume Study where they found that among 5000 identical resumes, the candidates with African American names are 50% less likely to receive an interview. In Caliskan et al. (2017) study, using the same names as the previous study, they found that European American names are more likely to be associated with being pleasant than African American names and therefore European American names would be invited to interview (assuming semantic similarity between pleasantness and invitation to interview). This kind of sentiment analysis allows researchers to understand empirical information about the world by classifying language as positive, negative, or neutral. In this case, the results underscore the cultural stereotypes that African Americans are less pleasant than Caucasians. Word embeddings capture these associations that are not easily separable from the semantics of the word.

NLP Challenges for Colloquial Language

Different types of text data pose unique challenges. Language found on Google News or the Web is different from the language on social media like Twitter. Platforms like Twitter include more colloquial forms of language and many times include abusive language (Bartneck et al., 2020). Natural language processing systems may not be robust enough to accurately parse and understand such language generating more negative bias. African American Vernacular English (AAVE) is a good example of language on Twitter that may be inaccurately processed. Jorgensen et al. (2015) finds that part-of-the-speech (POS) tagging does not accurately represent AAVE. To clarify, POS tagging is used to tokenize parts of speech like nouns, verbs, adverbs, and adjectives (Vangara et al. 2020). For example, “brotha” is tagged as an adverb, verb, or a foreign word. Contractions like “finna” (meaning “going to”) and “gimme” (meaning “give me”)

are oftentimes tagged incorrectly as well. This demonstrates that the POS tagging model fails when it comes to social media dialects because they are not explicitly considered which is problematic for creating robust analyses of colloquial natural language (Jorgensen et al., 2015).

Similarly, abusive language detection models are used to regulate hateful, discriminative content. However, many of these NLP models are biased towards identity words due to imbalanced training data (Park & Shin & Fung, 2018). For instance, claims like "You are a good woman" can be considered sexist in models. To avoid these kinds of gender biases, debiasing algorithms must be incorporated.

Debiasing NLP Algorithms

There have been a number of studies that discuss debiasing methodology and frameworks. Bolukbasi et al. (2016) trained the Word2Vec embedding on Google News text. The results suggested gender biases consistent with biases found by the GloVe algorithm as well. The first step in their debiasing algorithm is to identify the gender direction of the subspace that captures bias. From here, two debiasing algorithms are tested. The Neutralize and Equalize option enforces that all neutral words are equidistant from all other words in and out of the subspace by assigning the value of 0 to neutral words. The Equalize function can pose disadvantages by removing certain associations that should not be equalized in some unique cases. For instance, a model may want to assign a higher probability to the phrase "grandfather a regulation" than to "grandmother a regulation", but after equalizing the set this distinction would be removed. The second algorithm that Bolukbasi et al. (2016) explains is called Soften. This function preserves the relationship between words but reduces the projection of gender neutral words on the subspace. Therefore, it is a less extreme version of Neutralize and Equalize.

Another metric for unintended demographic bias, RNSB (Relative Negative Sentiment Bias), can perhaps offer a solution to the somewhat blanket approach above. In this algorithm, a logistic classification model is trained to predict the probability of any word being a negative sentiment word. This can be used to predict unfair sentiment for neutral words to discover if they are entangled with negative sentiment in word embeddings (Sweeney & Najafian, 2019). This approach projects similar results to as the WEAT score shows but can measure discrimination with respect to multiple demographics (Sweeney & Najafian, 2019).

Ethics of Face Recognition

Another application of machine learning is face recognition (FR). Unintended social biases in face recognition systems can lead to discriminative, unethical decision making. The following statistics helps understand current uses of FR technology. 16 states allow the FBI to use face recognition technology to compare the suspected criminal's face to peoples' drivers license photos. Multiple police departments, including Chicago, Dallas, and Los Angeles departments, collect real-time face recognition data from live surveillance cameras. At least 26 states allow law enforcement to run face recognition searches with little regulation (Garvie, 2016). These statistics demonstrate the widespread and growing use of artificial intelligence to inform critical decisions. Because African Americans have disproportionately high arrest rates, face recognition databases of mug shots include a disproportionate number of African Americans. Overrepresentation of certain groups causes imbalance in data and perpetuates problematic biases in algorithms.

Because face recognition is prevalent, it is crucial to discuss its adverse effects in order to make it better for the future. In a paper by Raji et al. (2020), a group of prominent ethical AI

researchers develop a benchmark dataset called CelebSET (Rothe, 2016), a subset of IMDB-Wiki. This dataset is composed of 80 celebrity identities split up into 4 categories: darker male, darker female, lighter male, lighter female. The data set is used to evaluate a variety of recognition tasks on Microsoft, Clarifai, and Amazon's APIs. The 4 tasks are gender, age, name, and smile recognition. They found that all APIs performed the worst on darker complexions and females. The darker female category had the worst accuracy, while the lighter male category had the highest accuracy across all APIs (Raji et al., 2020). This finding is consistent with Klare et al. (2012) finding that facial recognition technology used by US law enforcement is systematically lower for Blacks and females. In addition, Klare et al. (2012) claims the US law enforcement facial recognition accuracy is lower for groups ages 18-30 as well. It is well proven that there is a problem at hand: marginalized groups are at a disadvantage when face recognition is applied to law enforcement, justice systems and automated hiring systems to name a few (Gebru, 2019). In fact, a predictive policing software, Predpol, primarily predicts Black neighborhoods to be crime hotspots. This causes over policing in these disadvantaged neighborhoods and with every crime additional data amplifies existing social inequalities (Kristian & Isaac, 2016).

When it comes to finding a solution to imbalanced populations, trying to purposely increase representation of marginalized groups can lead to tokenism. Population monitoring and tokenism put more visibility and responsibility on the tokenized individuals than there should be, in essence, further aggravating marginalization (Raji et al., 2020). Additionally, while AI needs to be fairer, eliminating categorical information, like race, is not the solution. While, researchers want to reduce systematic biases in technology, the data must still reflect the actual situation in order to reach useful conclusions (Kleinberg et al., 2018). Some companies use AI audits and moral test cases to test for algorithmic bias (Bartneck et al., 2020). In fact, just 6 months after the

CELEBset study (Raji et al., 2020) with Microsoft, Clarifai, and Amazon's APIs was released, which uses algorithmic auditing, Microsoft and IBM built new versions of their API and Google instituted a fairness organization (Gebru, 2019).

Face Recognition In Humans

Perhaps not surprisingly, humans have a built-in cognitive architecture that can recognize and perceive faces efficiently and almost all the time. A complicated process like face recognition is done so effortlessly in our cognitive systems. In fact, seven types of information are derived from faces: pictorial, structural, visually derived, identity specific, name, expression and facial speech coding (Bruce & Young, 1986). In human perception, faces are perceived holistically, as a whole opposed to a collection of separate parts. In artificial intelligence, faces are mapped according to biometrics and recognition is based more on matching features (Schroff et al., 2015). In fact, O'Toole et al. (1990) explains the process of this feature extraction through an unsupervised learning model using backpropagation.

The Fusiform Face area (FFA) is the localized area involved with facial recognition in the human brain. It is less active in cross-race faces than its own race faces which supports the other-race effect theory (Johnson & Fredrickson, 2005). It is worth discussing face processing in humans as face recognition software is fundamentally derived from the human cognitive ability. Biases that exist in human systems can be reproduced in technology as well and so it is necessary to have an understanding of human cognition before designing fair systems (Lake et al., 2016). The other race effect theory says that people recognize faces of their own race more accurately than faces of other races. The contact hypothesis provides an explanation to this theory claiming that the more contact there is with other races, the smaller the other race effect really is. This has

been a controversial perspective as there have been many studies with evidence for and against it (Furl et al., 2002). Nonetheless, one fact that is agreed upon is that racial categorization significantly changes how individual features are represented in memory - the features may be stored in memory as more stereotypic of that race than they really are (Maclin & Malpass, 2003). Discussing facial recognition and perception biases in humans abstracts away from the discussion of AI. However, it is relevant to briefly understand human perception and processes as it can inform and inspire human-centered machine learning algorithms.

The Role of Experience and Sentiment in Face Recognition

The Furl et al. (2002) study considers various face recognition algorithms that computationally represent psychological models like generic contact hypothesis and non-contact hypothesis using the Face Recognition Technology (FERET) program (Phillips et al., 2000). As a commonality among all face recognition models, any face representation is a vector in a multidimensional space. The coordinates represent feature values while the distances between points quantify similarities between two faces. Quantification of similarities and differences between facial features is known as photogrammetry or photo-anthropometry (Mann & Smith, 2017). In this study (Furl et al., 2002), the basis of all algorithms was principal component analysis (PCA) which is a statistical method to learn a set of faces by encoding the face based on a training set (Craw & Cameron, 1992). The face recognition models were tested on 48 Caucasian and Asian faces, some of which were from the training set and others that were novel. The results showed that the generic hypothesis models fared better for minority race faces, the developmental contact model was more accurate for majority race faces, and the non-contact hypothesis models did not show any consistent advantages. The main takeaway presented by

these findings is that although humans are more adept at recognizing faces of their own race, there is no clear evidence that it is fully due to the amount of contact or experience with diverse races (Furl et al., 2002). This is relevant to artificial intelligence because understanding different models of human face processing of own race and other race faces can perhaps help determine what human-like models should be incorporated into algorithms.

While researchers are still trying to understand what degree experience truly plays in accurate face recognition across races, one study links positive emotions with increased accuracy in other-race holistic face recognition. This is a relevant study because other studies in natural language processing have shown the pervasive role of positive and negative sentiment in condoning biases and stereotypes (Caliskan et al., 2017). In this study (Johnson & Fredrickson, 2005), participants viewed brief videos to induce emotions of fear, joy, and neutrality. In the first experiment these emotions were induced prior to the learning stage (face encoding) and in the second experiment the videos were shown prior to the testing stage (face recognition of Black and White people). There were 89 Caucasian participants and their emotions were documented as per their indication. The results showed that the joy state improved recognition of Black faces and therefore reduced the own-race bias. It also confirmed that recognition was better when participants were feeling neutral rather than feeling fear. These results suggest that the effect of positive emotions can boost cross-race face recognition even after the faces have been learned. This could be because positive emotions foster inclusive social categories and mitigate away the salience of categories based on race (Isen et al., 1992). Another explanation is that positive emotions allow for holistic perceptions (Basso et al. 1996). This result is applicable to machine learning because positive and negative correlations can be learned by a machine, which can be comparable to the onset of positive emotions in humans.

Algorithmic Solutions to Mitigate Bias in Face Recognition

Machine learning researchers are continuously improving FR algorithms to be more inclusive and debiased like Microsoft researchers who created a new FR API after an audit that pointed out algorithmic flaws (Gebru, 2019). The fact of the matter is that even with balanced training data, feature separability among minorities is inferior to that of Caucasians (Wang & Deng, 2019). In an effort to product equitable recognition performance, this study implements deep reinforcement learning to set adaptable margins to balance race performance. Mimicking human learning and decision making, deep Q-learning is implemented to train the model to balance distances among races and remove demographic biases (Wang & Deng, 2019). This is achieved through the two key components of Q-learning, the maximization of rewards and state updating, to learn high causal relationships and understand biases present in the learning process. It was found that this method did provide an equitable algorithm and better performance; however, when reducing the quality of images with the application of a gaussian blur, non-white faces were more adversely affects than white faces. This illustrates a key point that even with balanced training data and an equitable algorithm, minorities are disproportionately affected (Wang & Deng, 2019).

A solution to the problem of low-quality pictorial data or ambiguous features could be probabilistic face embeddings (PFEs). This method converts existing face embeddings into PFEs, which use a gaussian distribution to represent face images (Shi & Jain, 2019). This captures the uncertainty and ambiguity of feature values. According to the authors, this approach is more realistic in that it can endure uncontrolled environments. Another face recognition system that is known for its exceptional performance is FaceNet. FaceNet uses a deep convolutional neural network to optimize face embeddings (Schroff et al., 2015). Instead of

using an intermediate bottleneck layer - as do most deep learning methods - it utilizes a unique mining method that achieves a 95.12% accuracy on Youtube Faces database.

Conclusion

Through this investigation of biases in machine learning, it is clear to see the permeating effect that inequitable, poorly trained technologies can have on already standing cultural, racial, and gender inequalities in society. This paper considered the ethical and social implications of NLP and FR which are important to recognize in order to engineer technologies that are morally sound. By using the methods discussed in this paper, bias can be mitigated which can lead to increased trust in machine learning technologies and overall better human-AI interactions.

References

Bartneck, C., Lutge, C., Wagner, A., & Wels, S. (2020). *Introduction to Ethics In Robotics And AI*. S.l.: Springer. doi:10.1007/978-3-030-51110-4.

Basso, M., Schefft, B., Ris, M., & Dember, W. (1996). Mood and global-local visual processing. *Journal of the International Neuropsychological Society, 2*(3), 249-255. doi:10.1017/S1355617700001193.

Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems* (pp. 4349-4357).

Bruce, V., & Young, A. (1986). Understanding face recognition. British journal of psychology, 77(3), 305-327. https://doi.org/10.1111/j.2044-8295.1986.tb02199.x.

Buolamwini, Joy, and Timnit Gebru. "Gender shades: Intersectional accuracy disparities in commercial gender classification." In Conference on fairness, accountability and transparency, pp. 77-91. 2018.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334), 183-186. DOI: 10.1126/science.aal4230.

Clare Garvie. *The perpetual line-up: Unregulated police face recognition in america.* Georgetown Law, Center on Privacy & Technology, 2016.

Craw, I., & Cameron, P. (1992). Face recognition by computer. In *BMVC92* (pp. 498-507). Springer, London.

Fleming, M. K., & Cottrell, G. W. (1990, June). Categorization of faces using unsupervised feature extraction. In *1990 IJCNN International Joint Conference on Neural Networks* (pp. 65-70). IEEE. DOI: 10.1109/IJCNN.1990.137696.

Furl, N., Phillips, P. J., & O'Toole, A. J. (2002). Face recognition algorithms and the other-race effect: computational mechanisms for a developmental contact hypothesis. *Cognitive Science,* 26(6), 797-815. https://doi.org/10.1016/S0364-0213(02)00084-8.

Gebru, T. (2019). Oxford Handbook on AI Ethics Book Chapter on Race and Gender. *arXiv preprint arXiv:1908.06165*.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology, 74*(6), 1464–1480. https://doi.org/10.1037/0022-3514.74.6.1464.

Hovy, D., & Spruit, S. L. (2016, August). The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (Volume 2: Short Papers) (pp. 591-598). **DOI:** 10.18653/v1/P16-2096.

Isen, A. M. (2001). An influence of positive affect on decision making in complex situations: Theoretical issues with practical implications. J*ournal of consumer psychology,* 11(2), 75-85. https://doi.org/10.1207/S15327663JCP1102_01.

Jacovi, A., Marasović, A., Miller, T., & Goldberg, Y. (2020). Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in ai. *arXiv preprint arXiv:2010.07487*.

Johnson, K. *" We all look the same to me": Positive emotions eliminate the own-race bias in face recognition Kareem J. Johnson and Barbara L*. Fredrickson University of Michigan 9/13/2004. Ann Arbor, 1001, 48109-1109. https://doi.org/10.1111/j.1467-9280.2005.01631.x.

Jørgensen, A., Hovy, D., & Søgaard, A. (2015, July). Challenges of studying and processing dialects in social media. In *Proceedings of the workshop on noisy user-generated text* (pp.9-18).

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Ashesh Rambachan. 2018. Algorithmic fairness. Aea papers and proceedings 108: 22–27. DOI: 10.1257/pandp.20181018.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and brain sciences*, *40*. DOI:https://doi.org/10.1017/S0140525X16001837.

Lum, Kristian, and William Isaac. "To predict and serve?." Significance 13, no. 5 (2016): 14-19. https://doi.org/10.1111/j.1740-9713.2016.00960.x.

Mann, M., & Smith, M. (2017). Automated facial recognition technology: Recent developments and approaches to oversight. *UNSWLJ*, 40, 121.

Park, J. H., Shin, J., & Fung, P. (2018). *Reducing gender bias in abusive language detection.* arXiv preprint arXiv:1808.07231.

Phillips,P.J.,Moon,H.,Rizvi, S., &Rauss,P. (2000). The FERET evaluation method for face recognition algorithms. IEEE Transaction Pattern Analysis and Machine Intelligence, 22, 1090–1104. **DOI:** 10.1109/34.879790.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In EMNLP, volume 14, pages 1532–43.

Raji, I. D., Gebru, T., Mitchell, M., Buolamwini, J., Lee, J., & Denton, E. (2020, February). Saving face: Investigating the ethical concerns of facial recognition auditing. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (pp. 145-151). https://doi.org/10.1145/3375627.3375820.

Rasmus Rothe, Radu Timofte, and Luc Van Gool. 2016. Deep expectation of real and apparent age from a single image without facial landmarks. International Journal of Computer Vision (IJCV) (7 2016).

Schroff, F., Kalenichenko, D., & Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 815-823).

Shi, Y., & Jain, A. K. (2019). Probabilistic face embeddings. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 6902-6911).

Slovic, P., Finucane, M. L., Peters, E., & MacGregor, D. G. (2007). The affect heuristic. *European journal of operational research*, *177*(3), 1333-1352. https://doi.org/10.1016/j.ejor.2005.04.006.

Sweeney, C., & Najafian, M. (2019, July). A transparent framework for evaluating unintended demographic bias in word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 1662-1667).

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive psychology*, *5*(2), 207-232. https://doi.org/10.1016/0010-0285(73)90033-9.

Vangara, R. V. B., Vangara, S. P., & Thirupathur, V. K. (2020). *A Survey on Natural Language Processing in context with Machine Learning*.

Wang, M., & Deng, W. (2020). Mitigating Bias in Face Recognition Using Skewness-Aware Reinforcement Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 9322-9331).