Türkçe Kelime Temsillerinde Cinsiyetçi Ön Yargının İncelenmesi
Investigation of Gender Bias in Turkish Word Embeddings

Nurullah SEVİM
Elektrik Elektronik Mühendisliği Bölümü
İhsan Doğramacı Bilkent Üniversitesi
Ankara, Türkiye
nurullah.sevim@ug.bilkent.edu.tr

Aykut KOÇ
Elektrik Elektronik Mühendisliği Bölümü
İhsan Doğramacı Bilkent Üniversitesi
Ulusal Manyetik Rezonans Araştırma Merkezi
Ankara, Türkiye
aykut.koc@bilkent.edu.tr

2021 29th Signal Processing and Communications Applications Conference (SIU) | 978-1-6654-3649-6/21/$31.00 ©2021 IEEE | DOI: 10.1109/SIU53274.2021.9477774

***Özetçe* —Doğal Dil İşleme uygulamalarında cinsiyetçi ön yargının incelenmesi, olası bir cinsiyetçi yaklaşımın olumsuz sonuçlarından dolayı son zamanlarda önem kazanmıştır. Özellikle İngilizce kelime temsillerinde bu tür ön yargılar çeşitli bağlamlarda incelenerek birçok araştırma yapılmıştır. Bu çalışmada Türkçe kelime temsillerinin cinsiyetçi ön yargılar açısından durumu incelenmiştir ve Türkçe dil yapısı İngilizce dil yapısı ile cinsiyetçi ön yargılar kapsamında karşılaştırılmıştır. Kelime temsillerinde yapılan cinsiyetçi ön yargıların ölçümü sonucunda Türkçe'nin İngilizce'ye kıyasla dil yapısında cinsiyetçi ön yargıyı daha az barındırdığı sonucuna varılmıştır.**

Anahtar Kelimeler—Doğal Dil İşleme, Cinsiyetçi ön yargı, Kelime temsilleri.

***Abstract*—Investigating gender bias in Natural Language Processing has recently gained importance due to the negative consequences of a possible sexist approach. Especially by examining such biases in English word embeddings in various contexts, many studies have been conducted on these issues. In this study, the status of Turkish word embeddings in terms of gender bias was examined and the Turkish language structure was compared with English within the scope of gender biases. As a result of the measurement of gender bias in word embeddings, it was concluded that Turkish contains less gender bias in language structure compared to English.**

Keywords—Natural Language Processing, gender bias, word embeddings.

I. GİRİŞ

Doğal Dil İşleme; makinelere okuma, yazma ve insan dillerini anlamlandırma yetisini kazandırmaya çalışan, yapay zeka çalışmalarının bir alt alanıdır. Son yıllarda önemli bir ölçüde popülarite kazanan Doğal Dil İşleme'nin birçok uygulama alanı vardır. Bunlardan bazıları makine çevirisi, anlamsal analiz ve konuşma tanımadır. Kelimelerin vektör temsilleri olarak tanımlanabilecek kelime temsilleri (word embeddings), bu tür uygulamalarda Doğal Dil İşleme'nin en önemli araçlarından birisidir [1]. Özellikle *Word2Vec* ve *GloVe* modellerinin tanıtılması, kelime temsillerinin Doğal Dil İşleme uygulamalarında kullanımını bir hayli yaygınlaştırmıştır [2], [3]. Son zamanlarda kelime temsilleri üzerinde yapılan yenilik ve geliştirmeler, kelime temsillerinin çeşitli Doğal Dil İşleme görevlerinde alınan sonuçları iyileştirmek amacıyla daha da yoğun kullanılmasına sebep olmuştur. Bu çalışmaların sonucunda, klasik kelime temsillerinin yanı sıra kelimelerin cümle içindeki bağlamlarını da absorbe edebilen ve birçok dili temsil edebilen dönüştürücü modeller literatüre kazandırılmıştır. [4]–[7].

Kelime temsilleri ve dönüştürücü modelleri halihazırda dilin yapay zeka tarafından kullanılmasına dair birçok problemi çözmekte kullanılmaktadır. Çözülen her problemde yapay zeka algoritmaları gerçek bir insanın düşünme tarzına büyük bir hızla yaklaşmaktadır. Bununla birlikte, yapay zeka algoritmalarının gerçeğe yaklaştıkça etik olma dereceleri de aynı hızla sorgulanmaktadır. Yapay zeka modellerinin özellikle hukuk alanında kullanılmaya başlanmasının hedeflenmesi göz önüne alındığında, oluşturulan algoritmaların tarafsız ve ön yargısız olma zorunluluğu ortaya çıkmaktadır. Öte yandan sadece hukuk alanında değil, dil modellerinin kullanıldığı bütün uygulamalarda herhangi bir şekilde yanlılık ortaya çıkması kabul edilemez bir durumdur. Örneğin özgeçmiş sınıflandırması gibi çalışmalarda, bir kimsenin bir işe uygunluğu kişinin yaşı, cinsiyeti, ülkesi, ırkı gibi etkenlerden tamamen bağımsız olmalıdır. Bu nedenlerle, kelime temsillerinde ya da daha büyük dil modellerinde istemsiz bulunabilecek olan ön yargıların tespit edilmesi büyük önem arz edilmektedir.

Literatürde İngilizce kelime temsillerinde bulunan ön yargıların derecelerini ölçmek için çeşitli çalışmalar yapılmıştır. Fakat benzer bir çalışmanın Türkçe kelime temsillerinde bulunmaması, Türkçe'nin cinsiyetçi ön yargılar açısından nasıl bir durumda olduğunu araştırılmasının ana motivasyon kaynağı olmuştur. Bu çalışma, Türkçe kelime temsillerindeki cinsiyetçi ön yargıların varlığına dair meslek grupları üzerinden bir araştırma içermektedir. Bu araştırmada, ön eğitimi yapılmış bir Türkçe kelime temsili modeli [1] kullanılarak, kelime temsillerinin ne derecede ön yargı içerdikleri ölçülmüştür. Elde edilen sonuçlar, Bölükbaşı ve arkadaşları [8] tarafından İngilizce'deki cinsiyetçi ön yargıyla ilgili yapılan çalışmanın sonuçlarıyla karşılaştırılmıştır ve sonuçlar arasındaki farkların

978-1-6654-3649-6/21/$31.00 ©2021 IEEE

[1] https://github.com/akoksal/Turkish-Word2Vec

olası sebepleri irdelenmiştir. Bununla birlikte, yapılan analizler iki dilin yapısı bakımından detaylı bir şekilde tartışılmıştır.

Bildirinin organizasyonu şu şekilde olacaktır. Literatürdeki benzer çalışmalar Bölüm II'de incelenecektir. Çalışmada kullanılan modeller Bölüm III'te anlatılacaktır. Yapılan deneyler ve elde edilen sonuçlar Bölüm IV'te verilecektir. Deneylerin analizi Bölüm V'te yapılacaktır. VI. Bölüm'de ise çalışmanın tartışılması yapılarak sonuca bağlanacaktır.

II. İLGİLİ ÇALIŞMALAR

Kelime temsillerinin Doğal Dil İşleme'ye yaptığı katkının yanı sıra, aynı zamanda cinsiyetçi ve ırkçı yaklaşımlar gibi insansı ön yargılar taşıyabildiği literatürde gösterilmiştir [8], [9]. İngilizce için yapılan bu araştırmalar, kelime temsillerinin eğitildikleri metinlerde bulunan cinsiyetçi ve ırkçı yaklaşımları taklit ettikleri ve bazen bu yaklaşımları daha da arttırdığı gözlemlenmştir. Örneğin, Bölükbaşı ve arkadaşlarının [8] yaptığı araştırmaların sonucunda, kelime temsillerinin cinsiyet bakımından nötr olan `programmer` (programcı) kelimesini erkeklere özgü bir meslek olarak değerlendirirken `homemaker` (evi çekip çeviren kişi) kelimesini de kadınlara özgü olarak değerlendirmektedir. Bu ön yargılar sadece mesleklerde değil, cinsiyet açısından nötr kalması gereken diğer olgularda da baş gösterebilmektedir. Örneğin `volleyball` (voleybol) kelimesini ve sporunu daha çok kadınlarla ilişkilendirirken, `football` (futbol) sporunu daha ziyade erkeklerle ilişkilendirmektedir. İdeal bir durumda bu kavramların cinsiyetsel bir ayrımı olmamalıdır ve her iki cinsiyetle de kelime temsilleri tarafından eşit bir şekilde ilişkilendirilmelidir.

Kelime temsillerinde bulunabilecek potansiyel ön yargılar sadece cinsiyet alakalı ön yargılarla sınırlı kalamayabilmektedir. Irksal, milliyetsel veya dinsel gibi birçok çeşitli kavramda da ön yargılar bulunabilmektedir. Örneğin kelime temsilleri, `housekeeper` (kahya) mesleğini de daha çok İspanyolların yaptığı bir meslek olarak algılayabilmektedir.

Çalışkan ve arkadaşları [9] tarafından yapılan başka bir çalışmada, kelime temsillerinin çiçeklere ve böceklere karşı olan etik olarak önemsiz ön yargıların yanı sıra, maalesef, ırklara ve cinsiyetlere karşı da ön yargılar içerdiği belirtilmiştir. Kelime temsillerindeki cinsiyetlere ve ırklara karşı olan bu tür kabul edilemez yaklaşımların varlığı, bu önemli Doğal Dil İşleme araçlarının çeşitli günlük uygulamalardaki kullanımlarına yönelik endişeler doğurmaktadır. Yapılan bir araştırmada, arama motorlarında belirli meslekler araştırılınca, bu arama motorlarının bazı sosyal gruplardaki insanların üst sıralarda gösterilmesine öncelik verdiği ve bu durumun meslek gruplarındaki sosyal eşitsizliği arttırdığı gösterilmiştir [10].

Kelime temsillerinde bulunan ön yargıların kökenine dair yapılan araştırmalar, bu ön yargıların nedenine dair çeşitli ipuçları vermektedir. İlk akla gelen sebeplerden biri; kelime temsilleri oluşturulurken eğitilen modellerin, insanlar tarafından üretilen metinler aracılığıyla eğitilmesi olarak gösterilebilir. İnsanların bu metinlere aktardığı ön yargılar doğal olarak bu metinler tarafından eğitilen modellerde de kendini gösterebilmektedir [11]. Ne var ki bu durum modellerin eğitildiği bütün metinlerin ön yargılı olduğunu göstermemektedir. Yapılan başka bir çalışmada, kelime temsillerinde bulunan ön yargıların, bu kelime temsillerinin eğitilmesi sürecinde kullanılan hangi metinlerden kaynaklandığı bilgisine ulaşılmaya çalışılmıştır [12].

III. METODOLOJİ

Bu çalışmada, kelimelerin cinsiyetlere göre ne derece ön yargı içerdiğini; kelimelere karşılık gelen vektörlerin, tanımı yapılan bir cinsiyet vektörü üzerindeki iz düşümünün büyüklüğü ölçülerek belirlenmiştir.

Cinsiyet vektörü, biri eril biri dişil anlam taşıyan iki zıt anlamlı kelimenin farkı alınarak oluşturulur. Burda amaç kelime temsillerinin bulunduğu vektör uzayında olduğu varsayılan cinsiyet alt uzayını belirlemektir. Bu alt uzay, kendi ekseninde seçilen eril-dişil doğrultusundaki herhangi bir vektörle temsil edilebilir.

Türkçe kelime temsillerinin bulunduğu vektör uzayında ise, cinsiyet alt uzayı `adam` ve `kadın` vektörlerinin farkı alınarak elde edilen doğrultu kullanılarak temsil edilmiştir. Cinsiyet vektörünün yönü `adam` vektöründen `kadın` vektörüne doğru tanımlanmıştır. Cinsiyet vektörünün tanımlanmasının ardından, meslek gruplarına dair kelimeler içeren bir kelime listesindeki kelimelerin cinsiyet vektörü üzerindeki iz düşümünün hesaplanmasıyla, bu kelimelerin kullanılan kelime temsili modelinde hangi cinsiyetle ilişkilendirildiği belirlenmeye çalışılmıştır. İz-düşüm büyüklüğü, hedef kelimeye karşılık gelen vektör ile cinsiyet vektörünün iç çarpımı alınarak bulunmuştur.

$$\mathbf{c} \cdot \mathbf{h} = |\mathbf{c}||\mathbf{h}|cos(\theta), \tag{1}$$

Denklem 1'de, $\mathbf{c}$ cinsiyet vektörünü, $\mathbf{h}$ hedef meslek kelimesine karşılık gelen vektörü ve θ iki vektör arasındaki anlamsal uzaydaki açıyı temsil etmektedir. Kelime temsilleri kullanımdan önce normalize edildiği için, hesaplamalarda birim vektörler kullanılmaktadır ve kelime vektörlerinin normları 1'e eşit olmaktadır. Bu sebeple yukarıdaki Denklem 1, Denklem 2'deki hale gelmektedir ve kelime vektörlerinin iç çarpımları aralarındaki açının cosinüs değerine eşit olmaktadır.

$$\mathbf{c} \cdot \mathbf{h} = cos(\theta). \tag{2}$$

Kelimelerin iz düşüm büyüklüğü ve işareti, kelimelerin hangi cinsiyetle ilişkili olduğunu belirlemek için kullanılabilmektedir. Cinsiyet vekötürünün tanımlandığı yöne (`adam-kadın` ya da `kadın-adam`) bağlı olarak iz düşüm sonuçları yorumlanmaktadır. Tanımladağımız cinsiyet vektöründen elde edilen iz düşüm ölçüsü -1'e yakın olduğu takdirde kelimenin eril bir anlam taşıdığı, 1'e yakın olduğun durumda ise dişil bir anlam taşıdığı sonucuna varılmaktadır. İdeal durumda cinsiyetten bağımsız olması gereken kelime vektörlerinin cinsiyet vektörüne olan iz düşümü 0 ya da 0'a çok yakın değerler almalıdır.

IV. DENEYLER VE SONUÇLARI

Deneylerin hazırlık aşamasında gerekli olan, meslek gruplarına dair kelimeleri içeren kelime listesi, Bölükbaşı ve arkadaşları [8] tarafından yapılan çalışmada kullanılan kelime listesinin tarafımızca Türkçe'ye çevirilmesiyle elde edilmiştir. Bu listede bulunan bazı kelimelerin Türkçe karşılığı, kullanılan Türkçe kelime temsillerinin içeriğinde yer almadığı için, bu

kelimeler listenin dışına çıkarılmıştır. Deneylerde kullanılan kelime temsilleri ise, önceden de belirtildiği gibi, önceden eğitilmiş bir Türkçe kelime temsili modeli kullanılarak uygulanmıştır.

TABLO I: MESLEKLER VE İZ DÜŞÜM DEĞERLERİ (TÜRKÇE)

Kelime	İz Düşümü	Kelime	İz Düşümü
anlatıcı	-0.230	radyolog	0.135
patron	-0.217	delege	0.115
davulcu	-0.214	öğrenci	0.114
çavuş	-0.190	çevreci	0.113
kaptan	-0.187	balerin	0.110
yönetmen	-0.182	vatandaş	0.103
badigard	-0.181	misyoner	0.101
haydut	-0.178	aktris	0.089
dedektif	-0.176	eğitmen	0.081
gangster	-0.170	eğitimci	0.078

Mesleki kelime listesindeki her bir kelimenin cinsiyet vektörü üzerindeki iz düşümü hesaplanmış ve iz düşümlerinden yola çıkılarak kelimenin hangi cinsiyetle ilişkilendirildiği bulunmaya çalışılmıştır. Türkçe kelime temsilleri kullanılarak yapılan bu çalışmanın sonuçları Tablo I'de gösterilmiştir. Tabloda iz düşüm büyüklüğü en büyük 10 kelime verilmiştir [2]. Dikkat edilmelidir ki, negatif işaretli iz düşümler kelimenin eril anlam taşıdığını pozitif işaretli iz düşümler ise kelimenin dişil anlam taşıdığını göstermektedir.

Türkçe için elde edilen bu sonuçlar, aynı kelime grubunun İngilizce karşılıkları kullanılarak aynı metodoloji sonrasında İngilizce kelime temsilleri ile alınan sonuçlarıyla karşılaştırılmıştır. İngilizce kelime temsilleriyle alınan sonuçlar Tablo II'de verilmiştir [8]. Tabloda sadece iz düşümleri en büyük olan kelimeler gösterilmektedir.

TABLO II: MESLEKLER VE İZ DÜŞÜM DEĞERLERİ (İNGİLİZCE)

Kelime	İz Düşümü	Kelime	İz Düşümü
maestro	-0.238	businesswoman	0.360
statesman	-0.217	actress	0.352
skipper	-0.208	housewife	0.340
protege	-0.203	homemaker	0.304
businessman	-0.202	registered_nurse	0.304
sportsman	-0.195	nurse	0.281
philosopher	-0.188	waitress	0.275
marksman	-0.181	receptionist	0.273
captain	-0.173	librarian	0.266
architect	-0.168	socialite	0.257

V. DENEYLERİN ANALİZİ

Tablo I incelendiği zaman; tabloda dişil önyargıların gösterildiği bölümde (sağ), `aktris` ve `balerin` dışında üst sıralarda yer alan kelimelerin dilimizde özellikle kadınlarla ilişkilendirilmediği görülebilir. Bu düşünceyi, Türkçe kelime temsilleri kullanılarak elde edilen iz düşümlerinin büyüklüklerinin de dikkat çekecek ölçüde yüksek olmaması desteklemektedir. Fakat, Tablo I'in eril önyargılar içeren kelimelerin sıralandığı bölümünde (sol), `çavuş` ve `badigard` gibi dilimizde ve toplumumuzda erkeklerle sıkı sıkıya ilişkisi olan kelimeler bulunmaktadır. Aynı zamanda, bu kelimelerin izdüşüm büyüklüklerinin dişil kısımdaki izdüşüm büyüklüklerinden bir hayli fazla olması da dikkat çekmektedir. Bu durum, Türkçe kelime temsillerinin yansıttığı dişil cinsiyetçi önyargıların endişe verici ölçüde olmadığını belirtmekle beraber toplumun eril kısmının belli bir kalıpta algılanıyor olabileceğini önermektedir.

Diğer taraftan, Tablo II'de verilen İngilizce kelimelerin cinsiyet vektörü üzerindeki iz düşümleri incelendiğinde, hem iz düşüm büyüklüklerinin Tablo I'de gösterilenlerden daha büyük olması hem de kelime temsillerinin eril algıladığı meslekler ve dişil algıladığı mesleklerin toplumdaki prestijleri arasındaki farkın fazla olması, İngilizce'nin Türkçe'ye göre cinsiyetçi ön yargıları daha fazla barındırdığını gösterebilir. Örneğin, İngilizce kelime temsillerinin dişil olarak gördüğü mesleklerin en üstlerinde `homemaker` (ev hanımı), `nurse` (hemşire), `receptionist` (resepsiyonist) gibi kelimeler yer alırken; eril olarak gördüğü mesleklerin başında `statesman` (devlet adamı), `philosopher` (filozof), `captain` (kaptan, yüzbaşı) gibi kelimeler gelmektedir. Bu durum İngilizce'nin meslekler açısından cinsiyetçi ön yargılar içerdiğine dair kanıt olabilmektedir.

Farklı dillerde yapılan aynı deneyin sonuçları arasındaki bu farklılığın sebebi dillerin yapısal farklılığından kaynaklanma ihtimali göz ardı edilmemelidir. İngilizce dilinde cinsiyet anlamı taşıyan ve sıkça kullanılan `he`, `she`, `her`, `his` gibi birçok kelime varken, Türkçe'de bu tür kelimelerin genelde cinsiyet bakımından nötr `o` ve `onun` gibi karşılıkları vardır.

İngilizce'deki bu tür kelimelerin metinlerde kullanılması, aynı bağlamda geçen başka cinsiyet bakımından nötr olan kelimelerin belirli cinsiyetlerle ilişkilendirilmesine yol açmaktadır. Kelime temsillerinin eğitiminde, ele alınan kelime metin içinde yakın geçtiği kelimelerle eğitildiği için cinsiyet belirten kelimeler nötr kelimeleri etkilemektedir. Öte yandan, Türkçe'de bu kelimelerin nötr karşılıklarının olması, aynı bağlamdaki diğer kelimelerin de nötr kalmasına ve dolayısıyla herhangi bir cinsiyetçi ön yargının metinler aracılığıyla taşınmasına engel olmaktadır. Yine bazı mesleki kelimelerin kendi içinde cinsiyet belirten yapılar içermesi (Ör. `marksman` ya da `bussinesswoman`) cinsiyet vektörü üzerindeki iz düşüm büyüklüğünü artırmaktadır.

VI. SONUÇ

Bu çalışmada bilgimiz dahilindeki literatürde bulunan hiçbir çalışmanın konusu olmayan Türkçe dilinde cinsiyetçi ön yargıların bulunma durumunu incelenmiştir. Bu amaçla Doğal Dil İşleme'de önemli ve popüler bir araç olan kelime temsilleri kullanılmıştır. Kelime temsillerinin ön yargı bulundurma dereceleri, kelime vektörlerinin tanımlanmış cinsiyet vektörü üzerinde bulunan iz düşümü ölçülerek hesaplanmıştır. Türkçe için elde edilen sonuçları, İngilizce için yapılmış benzer bir çalışmanın sonuçları ile kıyaslanmıştır. İki dil arasındaki sonuç farklılıkları dillerin yapıları bağlamında tartışılmıştır. Çalışmanın sonucunda Türkçe'nin İngilizce'ye göre daha az cinsiyetçi ön yargı barındırıyor olduğu görülmüştür. Bu durumun sebeblerinden biri olarak ise İngilizce'de yaygın olarak kullanılan ve cinsiyet iması barındıran birçok kelimenin Türkçe karşılıklarının herhangi bir cinsiyet anlamı barındırmaması olduğu

[2]Bütün kelimelerin olduğu orijinal İngilizce liste şu internet sitesinden bulunabilir: https://github.com/tolga-b/debiaswe/blob/master/data/professions.json

tartışılmıştır. İlerleyen zamanlarda, hem farklı kelime temsilleri kullanılarak hem de farklı ön yargı çeşitleriyle yapılacak deneylerle çalışmanın kapsamının artırılması hedeflenmektedir.

BİLGİLENDİRME

Bu çalışma Türkiye Bilimsel ve Teknolojik Araştırma Kurumu (TÜBİTAK) (1001-120E346) fonuyla desteklenmiştir.

KAYNAKLAR

[1] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, (Red Hook, NY, USA), p. 3111–3119, Curran Associates Inc., 2013.

[2] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings* (Y. Bengio and Y. LeCun, eds.), 2013.

[3] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, 2014.

[4] M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, (New Orleans, Louisiana), pp. 2227–2237, Association for Computational Linguistics, June 2018.

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, (Minneapolis, Minnesota), pp. 4171–4186, Association for Computational Linguistics, June 2019.

[6] D. Tang, F. Wei, N. Yang, M. Zhou, T. Liu, and B. Qin, "Learning sentiment-specific word embedding for Twitter sentiment classification," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, (Baltimore, Maryland), pp. 1555–1565, Association for Computational Linguistics, June 2014.

[7] S. Lai, L. Xu, K. Liu, and J. Zhao, "Recurrent convolutional neural networks for text classification," in *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, p. 2267–2273, AAAI Press, 2015.

[8] T. Bolukbasi, K.-W. Chang, J. Zou, V. Saligrama, and A. Kalai, "Man is to computer programmer as woman is to homemaker? debiasing word embeddings," in *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, (Red Hook, NY, USA), p. 4356–4364, Curran Associates Inc., 2016.

[9] A. Caliskan, J. J. Bryson, and A. Narayanan, "Semantics derived automatically from language corpora contain human-like biases," *Science*, vol. 356, no. 6334, pp. 183–186, 2017.

[10] M. De-Arteaga, A. Romanov, H. Wallach, J. Chayes, C. Borgs, A. Chouldechova, S. Geyik, K. Kenthapadi, and A. T. Kalai, "Bias in bios: A case study of semantic representation bias in a high-stakes setting," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, (New York, NY, USA), p. 120–128, Association for Computing Machinery, 2019.

[11] J. Zou and L. Schiebinger, "Ai can be sexist and racist — it's time to make it fair," *Nature*, vol. 559, pp. 324–326, 2018.

[12] M.-E. Brunet, C. Alkalay-Houlihan, A. Anderson, and R. Zemel, "Understanding the origins of bias in word embeddings," in *Proceedings of the 36th International Conference on Machine Learning* (K. Chaudhuri and R. Salakhutdinov, eds.), vol. 97 of *Proceedings of Machine Learning Research*, pp. 803–811, PMLR, 09–15 Jun 2019.