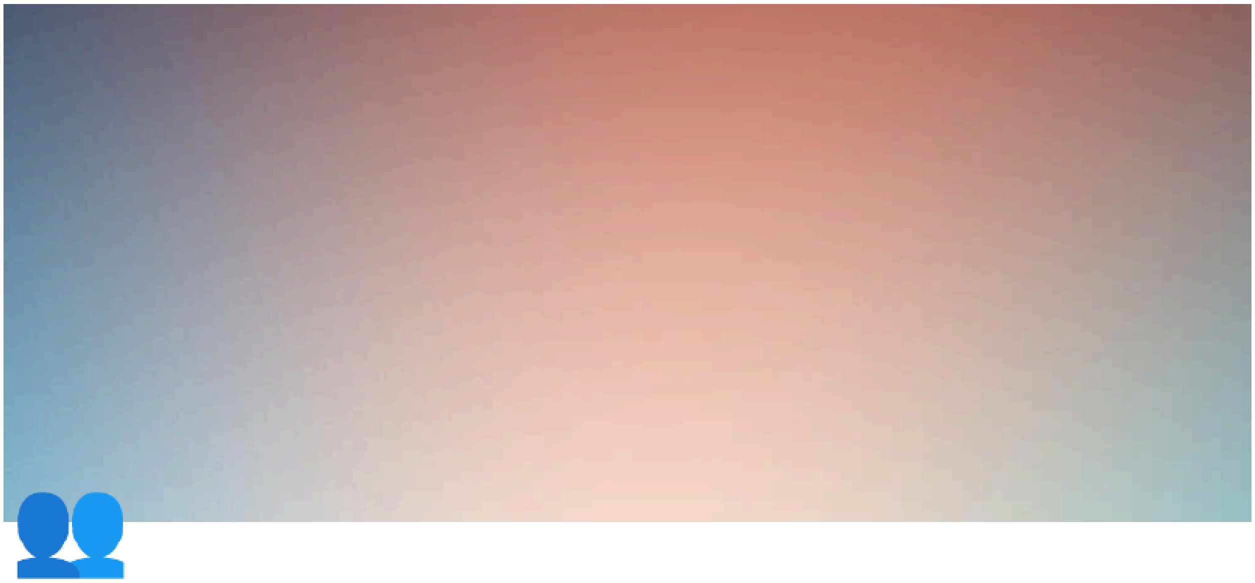


👥

Gender Bias in Word Embeddings: A Deep Dive into NLP and Ethical AI

FEBRUARY 28, 2025

VRISHAB PRSANTH DAVEY
vdave048@uottawa.ca

#UO - 300438343

📌 Abstract

Word embeddings are numeric representations of meaning derived from word co-occurrence statistics in human-produced texts. These embeddings, foundational to many Natural Language Processing (NLP) applications, have been found to encode and perpetuate social biases, including gender bias. This blog explores gender bias in widely used static English word embeddings trained on internet corpora (GloVe 2014, fastText 2017) [58]. While prior research has focused on specific gender associations, this analysis broadens the scope to include word frequency disparities, parts-of-speech tendencies, semantic clustering, and sentiment dimensions. Findings indicate that 77% of the 1,000 most frequent words in these embeddings are more associated with men [58], reinforcing a masculine default in online language. Male-associated

words are predominantly verbs, aligning with perceptions of agency and action, while female-associated words are often adjectives and adverbs, reinforcing descriptive stereotypes. Semantic clustering reveals that men are linked with professions, technology, and power, whereas women are frequently associated with appearance, relationships, and even explicit content. These findings underscore the necessity of mitigating biases in NLP models to create fairer AI applications.

Introduction & Background

All credits for this research go to **Aylin Caliskan et al. (2022)** for their hard work in conducting this study. I have tried my best to reference and summarize the key insights from their paper. This blog aims to provide an accessible understanding of their findings and the implications of how gender bias in word embeddings, foundational to many **Natural Language Processing (NLP) applications**, encode and perpetuate gender biases [58].

The role of Natural Language Processing (NLP) in daily life has grown exponentially, powering applications from machine translation to automated resume screening [7]. A key component of these applications is word embeddings—compressed, numeric representations of word meanings based on large-scale language data. However, research has shown that these embeddings inherit biases present in human-generated text, subtly reinforcing societal stereotypes[9-12].

Among the most pervasive biases is gender bias, which affects a wide range of NLP applications. Prior studies have revealed that word embeddings associate men with career-related terms and women with family-oriented words, perpetuating traditional gender role [12, 17, 26]. This blog takes a comprehensive approach to gender bias in word embeddings by analyzing not just direct word associations but also syntactic structures, semantic clusters, and psychological dimensions such as valence, arousal, and dominance.

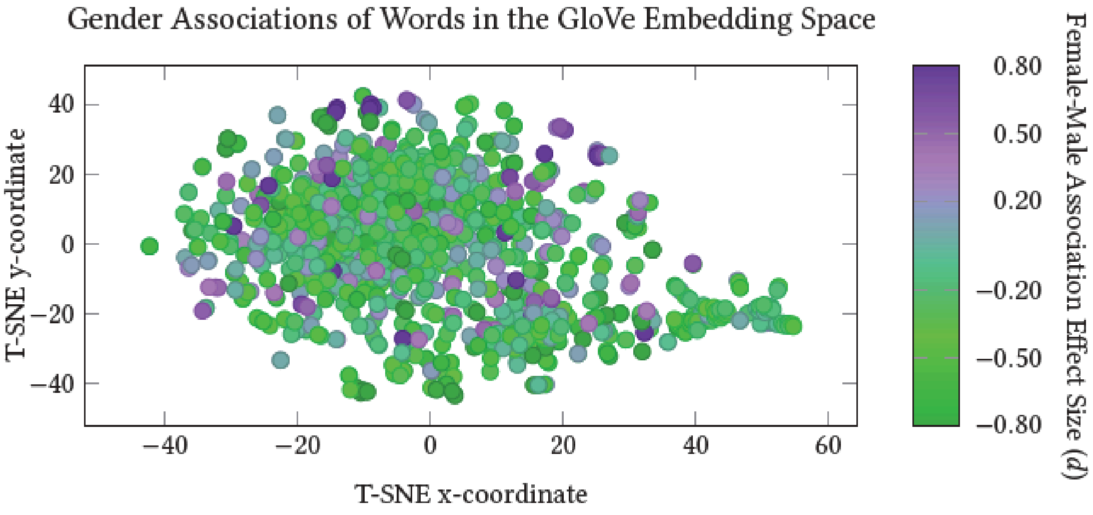


Figure 1: Visualization of Gender Associations in Word Embeddings

As demonstrated in **Figure 1**, the gender associations of the 1,000 most frequently occurring words in GloVe embeddings exhibit a clear imbalance, with a stronger association toward male-related words. This pattern persists even at broader scales, affecting NLP systems at a foundational level. The implications of these biases are far-reaching—from hiring algorithms that favor male candidates to AI-driven content moderation systems that reinforce gendered language disparities.

In this blog, I will dissect these findings, explain their implications, and discuss potential mitigation strategies for reducing gender bias in NLP models.

Understanding Gender Bias in Word Embeddings

Examining past work in the subject helps one to understand the level of gender bias in NLP. Previous research has found bias in many different spheres, from more general language structures to direct word associations. Based on their co-occurrence patterns in text corpora, **word embeddings** including those employed in **GloVe and fastText** offer continuous-valued vector representations of words [58]. These embeddings reflect cultural norms and prejudices present in training data, hence promoting **gender stereotypes in NLP applications** even if they efficiently capture links between words.

The **Word Embedding Association Test** and its variation, **Single- Category WEAT** [58] are among the most often used techniques for spotting bias in embeddings. These assessments assess the relative association of words with various attribute categories (e.g., male vs. female terms), and past studies have shown that such prejudices **fit** real-world occupational gender inequalities. Comparisons with **human psychological judgments** have confirmed that male-associated terms are typically related with **dominance and arousal** while female-associated words tend to score higher on **valence (pleasantness**). This has important ramifications for **AI models that process and interpret human language** since these prejudices affect automated decision-making in content creation, recommendations, and hiring as well as in content development.

Research has also repeatedly shown that **gender stereotypes** endure in word embeddings, so supporting conventional relationships between **men and leadership roles** and **women with domestic and appearance-related terms**. Although several approaches have been suggested to reduce these prejudices, most debiassing strategies fall short in **completely eliminating the underlying gender associations**. Moreover, **word frequency** is quite important for bias transmission; male-associated terms show **more often** in embeddings. This overrepresentation of male-linked words can lead to decreased visibility of underrepresented groups in AI-driven applications, hence adding to **systemic bias in NLP models**. Establishing **fairer language models** that more precisely reflect linguistic variation depends on addressing these discrepancies.

By means of this investigation of related work, it is evident that **gender bias in word embeddings is a well-documented problem with far-reaching consequences**. These prejudices will be further broken out in the next parts, together with possible remedies to lessen them.

Data Collection and Experimental Approach

We need datasets measuring human qualities, gender associations, and emotional aspects of language if we are to examine gender bias in NLP models. Our investigation centers on a few fundamental components. Trained on the **Common Crawl corpus**, more especially **GloVe (trained on 840 billion tokens) and fastText (trained on 600 billion tokens),** we employ **300-dimensional word embeddings**. Standard in artificial intelligence research, these extensively utilized embeddings also reflect prejudices in human language. We use **SC-WEAT**, a technique that measures the relative association of words with male and female identities, to evaluate gender correlations. Whereas a **negative value** denotes masculine prejudice, a **positive effect size** implies a closer link to female-associated phrases.

We investigate **emotional dimensions** by using the **NRC-VAD Lexicon**, a psycholinguistic dataset including human-rated ratings for **valence (positivity), arousal (intensity), and dominance (control)** beyond direct word connections. Previous studies indicate that **female-associated words** show **higher valence** whereas **male-associated words** usually show **higher dominance and arousal scores**. Furthermore, greatly affects how bias spreads in artificial intelligence systems is word frequency. We examine the frequency of gendered terms using third-party estimation tools since embeddings store words based on frequency but lack clear count values. Studies indicate that commonly occurring words effect AI behavior: if male-associated words predominate embeddings, **AI-driven decision-making could be disproportionately influenced by male-centric language**.

We curate a **Big Tech lexicon** including businesses like **Google, Microsoft, Amazon, and Facebook** [58] to investigate gender bias in professional environments. Our results show that **62% of Big Tech-related words** are more firmly connected with men, so supporting male dominance in conversations on technology and invention. The next part will investigate to evaluate gender bias in several linguistic structures.

Experimental Methodology

We conduct multiple analyses to assess gender bias across different linguistic structures:

- **SC-WEAT for gender bias effect sizes** across different frequency ranges (top 100, 1,000, and 10,000 most frequent words).
- **Parts-of-speech analysis**, categorizing gendered words as nouns, adjectives, or verbs.
- **Clustering algorithms**, grouping gender-associated words into conceptual domains (e.g., leadership, appearance, technology).
- **Correlation analysis**, measuring how gender bias aligns with valence, arousal, and dominance scores.

These analyses provide a multidimensional view of gender bias in NLP, helping us identify patterns that affect AI applications. The next section will dive deeper into Understanding how Gender Bias Manifests and discuss their implications for AI fairness.

Understanding How Gender Bias Manifests?

Word frequency distributions are one of the main forms in which gender bias shows in word embeddings. As shown in Figure 2, pretrained GloVe and fastText embeddings show a high male connection for most often occurring terms. This suggests that words connected to men show more often in big size text corpora, so supporting a masculine default in NLP models.

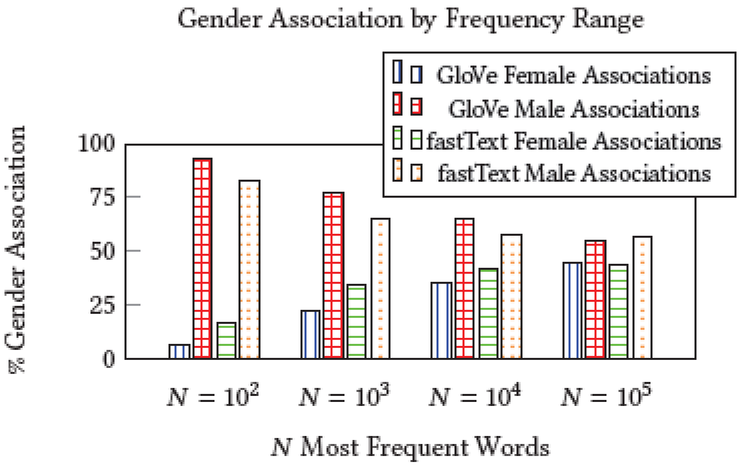
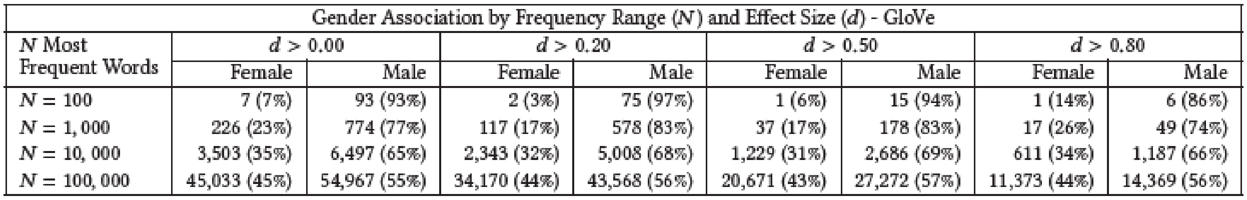


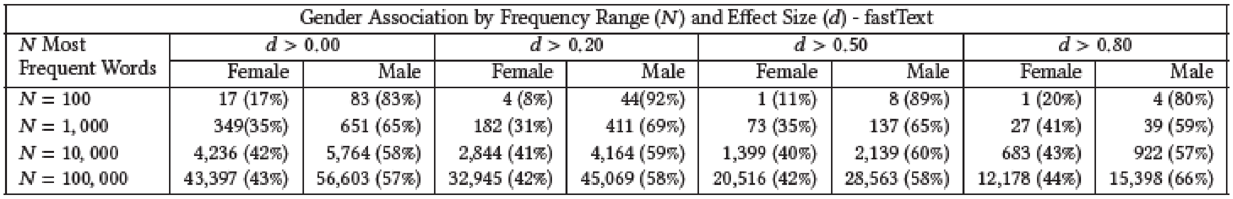
Figure 2: Most Frequent Words in Pretrained GloVe and fastText Word Embeddings [58].

Table 1 and Table 2, measure the distribution of male, and female associated words over several frequency ranges, give a thorough description of these frequency-based gender correlations. Though fastText is somewhat less biassed in terms of frequency percentage, the trends across both GloVe and fastText are constant.



Gender Association by Frequency Range (N) and Effect Size (d) - GloVe								
N Most Frequent Words	$d > 0.00$		$d > 0.20$		$d > 0.50$		$d > 0.80$	
	Female	Male	Female	Male	Female	Male	Female	Male
$N = 100$	7 (7%)	93 (93%)	2 (3%)	75 (97%)	1 (6%)	15 (94%)	1 (14%)	6 (86%)
$N = 1,000$	226 (23%)	774 (77%)	117 (17%)	578 (83%)	37 (17%)	178 (83%)	17 (26%)	49 (74%)
$N = 10,000$	3,503 (35%)	6,497 (65%)	2,343 (32%)	5,008 (68%)	1,229 (31%)	2,686 (69%)	611 (34%)	1,187 (66%)
$N = 100,000$	45,033 (45%)	54,967 (55%)	34,170 (44%)	43,568 (56%)	20,671 (43%)	27,272 (57%)	11,373 (44%)	14,369 (56%)

Table 1: The Most Frequent Words in the GloVe Embedding Vocabulary [58].

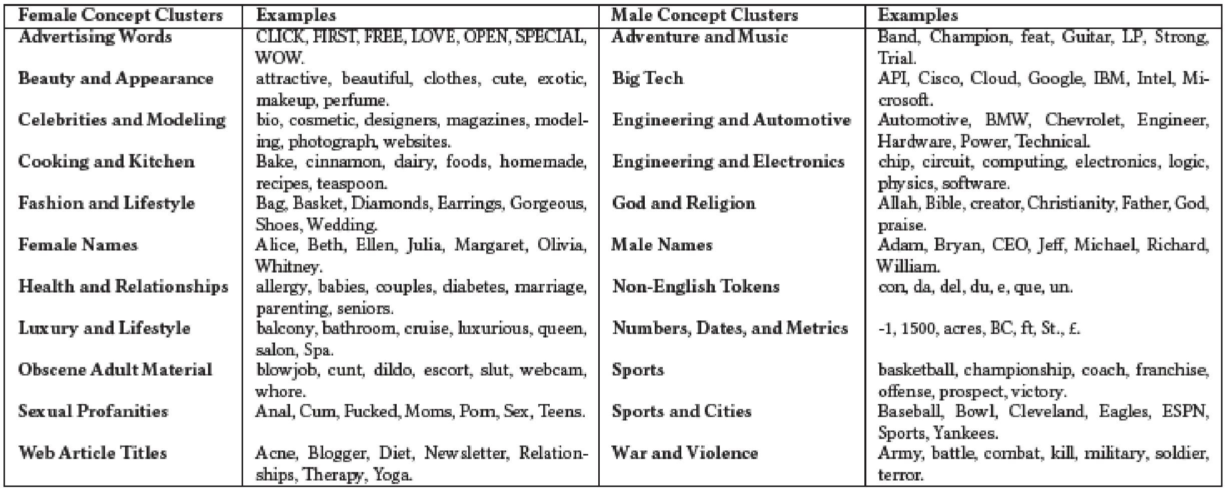


Gender Association by Frequency Range (N) and Effect Size (d) - fastText								
N Most Frequent Words	$d > 0.00$		$d > 0.20$		$d > 0.50$		$d > 0.80$	
	Female	Male	Female	Male	Female	Male	Female	Male
$N = 100$	17 (17%)	83 (83%)	4 (8%)	44(92%)	1 (11%)	8 (89%)	1 (20%)	4 (80%)
$N = 1,000$	349(35%)	651 (65%)	182 (31%)	411 (69%)	73 (35%)	137 (65%)	27 (41%)	39 (59%)
$N = 10,000$	4,236 (42%)	5,764 (58%)	2,844 (41%)	4,164 (59%)	1,399 (40%)	2,139 (60%)	683 (43%)	922 (57%)
$N = 100,000$	43,397 (43%)	56,603 (57%)	32,945 (42%)	45,069 (58%)	20,516 (42%)	28,563 (58%)	12,178 (44%)	15,398 (66%)

Table 2: The Most Frequent Words in the fastText Embedding Vocabulary [58].

Concept Clustering and Gendered Associations

Beyond frequency distributions, gender bias in word embeddings affects conceptual grouping of words. We classed the 1,000 most often occurring male- and female-associated terms using unsupervised clustering methods. Table 3 illustrates the outcomes, which reveal that whilst male-associated words are more likely to be clustered under engineering, leadership, money, and Big Tech-related concepts, female-associated words commonly cluster around themes such as beauty, fashion, relationships, and domestic responsibilities.



Female Concept Clusters	Examples	Male Concept Clusters	Examples
Advertising Words	CLICK, FIRST, FREE, LOVE, OPEN, SPECIAL, WOW.	Adventure and Music	Band, Champion, feat, Guitar, LP, Strong, Trial.
Beauty and Appearance	attractive, beautiful, clothes, cute, exotic, makeup, perfume.	Big Tech	API, Cisco, Cloud, Google, IBM, Intel, Microsoft.
Celebrities and Modeling	bio, cosmetic, designers, magazines, modeling, photograph, websites.	Engineering and Automotive	Automotive, BMW, Chevrolet, Engineer, Hardware, Power, Technical.
Cooking and Kitchen	Bake, cinnamon, dairy, foods, homemade, recipes, teaspoon.	Engineering and Electronics	chip, circuit, computing, electronics, logic, physics, software.
Fashion and Lifestyle	Bag, Basket, Diamonds, Earrings, Gorgeous, Shoes, Wedding.	God and Religion	Allah, Bible, creator, Christianity, Father, God, praise.
Female Names	Alice, Beth, Ellen, Julia, Margaret, Olivia, Whitney.	Male Names	Adam, Bryan, CEO, Jeff, Michael, Richard, William.
Health and Relationships	allergy, babies, couples, diabetes, marriage, parenting, seniors.	Non-English Tokens	con, da, del, du, e, que, un.
Luxury and Lifestyle	balcony, bathroom, cruise, luxurious, queen, salon, Spa.	Numbers, Dates, and Metrics	-1, 1500, acres, BC, ft, St., £.
Obscene Adult Material	blowjob, cunt, dildo, escort, slut, webcam, whore.	Sports	basketball, championship, coach, franchise, offense, prospect, victory.
Sexual Profanities	Anal, Cum, Fucked, Moms, Porn, Sex, Teens.	Sports and Cities	Baseball, Bowl, Cleveland, Eagles, ESPN, Sports, Yankees.
Web Article Titles	Acne, Blogger, Diet, Newsletter, Relationships, Therapy, Yoga.	War and Violence	Army, battle, combat, kill, military, soldier, terror.

Table 3: Concept Clusters for Male- and Female-Associated Words [58].

To further illustrate these findings, Figure **3** provides concrete examples of words within each gender-associated cluster. Notably, two of the most prominent female-associated clusters are related to **sexual profanities and explicit adult content**, reinforcing concerns about how NLP systems may inadvertently perpetuate harmful stereotypes.

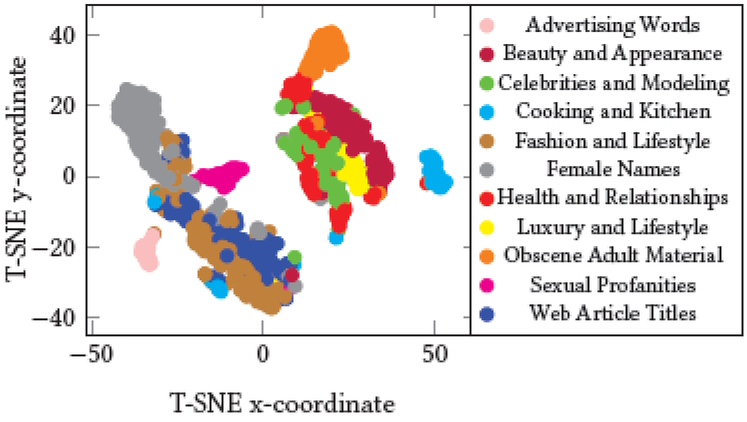


Figure 3: Examples of Concept Clusters for Male and Female Words [58].

Parts-of-Speech and Gender Disparities

The analysis of **parts-of-speech (POS) distributions** within word embeddings reveals notable gender disparities. As depicted in **Figure 4**, words associated with male attributes (d ≥ 0.50) in the GloVe vocabulary tend to cluster into conceptual groups such as **adventure, engineering, science, sports, violence, and war** [58]. These findings suggest that male-associated words are more likely to be linked with active, high-status, and traditionally male-dominated domains.

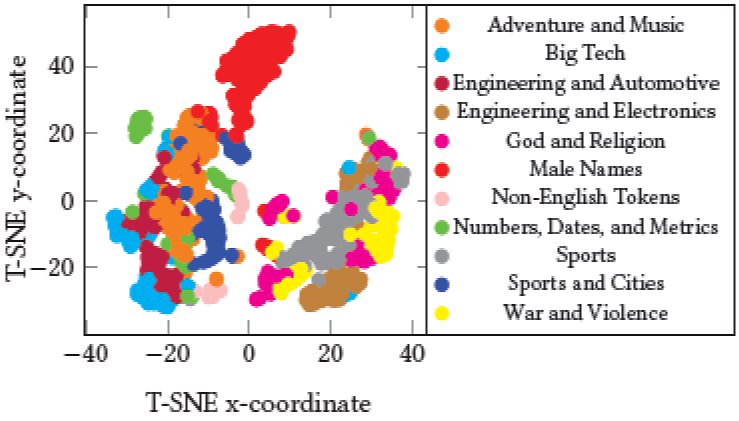
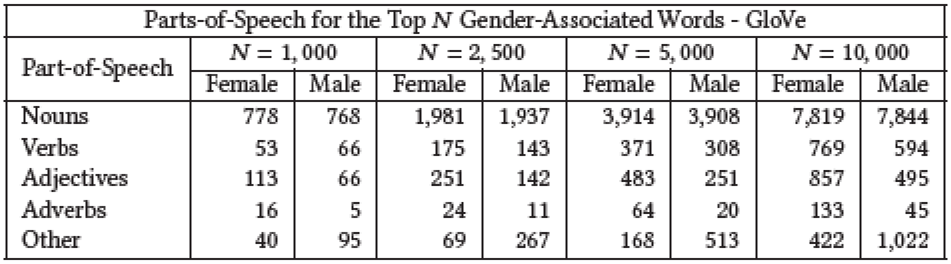


Figure 4: Concept Clusters for Male-Associated Words [58].

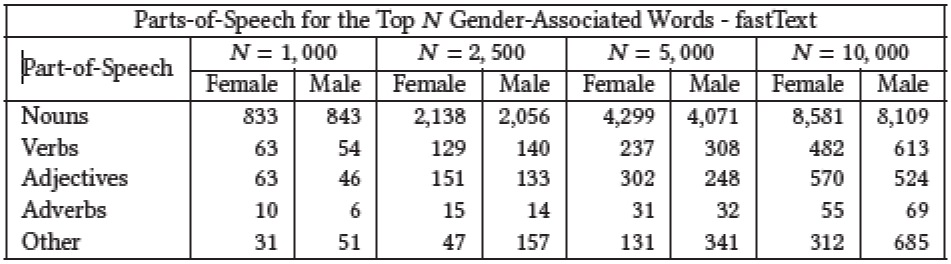
A further breakdown of POS associations in GloVe and fastText embeddings, provided in **Tables 4 and 5**, confirms that **female-associated words are predominantly adjectives and adverbs**, while **male-associated words are more likely to be verbs and nouns**. This pattern reinforces traditional gender roles, where men are depicted as active agents performing actions, while women are often described through their attributes.



Parts-of-Speech for the Top N Gender-Associated Words - GloVe

Part-of-Speech	$N = 1,000$		$N = 2,500$		$N = 5,000$		$N = 10,000$	
	Female	Male	Female	Male	Female	Male	Female	Male
Nouns	778	768	1,981	1,937	3,914	3,908	7,819	7,844
Verbs	53	66	175	143	371	308	769	594
Adjectives	113	66	251	142	483	251	857	495
Adverbs	16	5	24	11	64	20	133	45
Other	40	95	69	267	168	513	422	1,022

Table 4: Gender and Parts-of-Speech Associations in GloVe [58].



Parts-of-Speech for the Top N Gender-Associated Words - fastText

Part-of-Speech	$N = 1,000$		$N = 2,500$		$N = 5,000$		$N = 10,000$	
	Female	Male	Female	Male	Female	Male	Female	Male
Nouns	833	843	2,138	2,056	4,299	4,071	8,581	8,109
Verbs	63	54	129	140	237	308	482	613
Adjectives	63	46	151	133	302	248	570	524
Adverbs	10	6	15	14	31	32	55	69
Other	31	51	47	157	131	341	312	685

Table 5: Gender and Parts-of-Speech Associations in fastText [58].

For instance, in the **GloVe embeddings**, **113 of the 1,000 most frequent female-associated words are adjectives**, compared to **66 male-associated adjectives**[58]. Similarly, the **fastText embeddings** show this same pattern, with **female-associated words being more descriptive than their male counterparts**.

Additionally, **adverbs** follow this trend**—133 of the 10,000 most frequent female-associated words in the GloVe embeddings are adverbs**, compared to just **45 adverbs among male-associated words**[58].

The "Other" Category and Historical Significance

The **"Other" POS category** shows another important difference with regard to pronouns, interjections, and number words. Male-associated words abound in this category, as seen in **Tables 4 and 5** particularly in historical, scientific, and numerical settings.

677 of the 1,252 "Other" male-associated words at N = 10,000 are related in the **GloVe embeddings** to numerical, measurement, and historical issues [58].

Comparatively to merely
482 verbs among female-associated words, **613 of the 10,000 male-associated words** are verbs in the **fastText embeddings**.

This trend points to a **systematic gender bias** whereby female-associated words mostly stress **traits, feelings, and descriptions** while male-associated words fit intellectual, technical, and action-oriented areas.

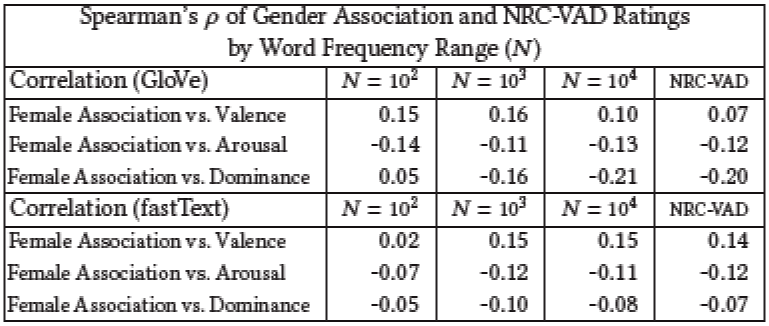
Singular vs. Plural Noun Bias

One last issue of thought is the variation in **singular and plural noun use** in gendered nouns. As noted in the study:

Comparatively to **92 plural male-associated nouns**, 166 of the 1,000 most frequent terms linked with women in GloVe embeddings are plural common nouns [58].
This trend points to a
linguistic tendency whereby women are more commonly described in collective terms while men are referred as individuals, hence strengthening gendered language positioning.

Applications of NLP, like text production, sentiment analysis, and AI-driven content suggestions, depend much on these prejudices. Unconsciously, the over-representation of men in action-oriented roles and women in descriptive, communal, and emotional settings shapes how artificial intelligence models understand and create text. Development of fair and balanced AI-driven communication systems depends on addressing these prejudices.

We investigated the psychological aspects of gender bias in word embeddings by means of correlations with **valence (Pleasantness), arousal (intensity), and dominance (power/control)** employing the **NRC-VAD lexicon**. Whereas male-related terms connect negatively with valence but positively with **dominance and arousal**, female-associated words tend to correlate positively with **valence**, suggesting they are associated with pleasantness.

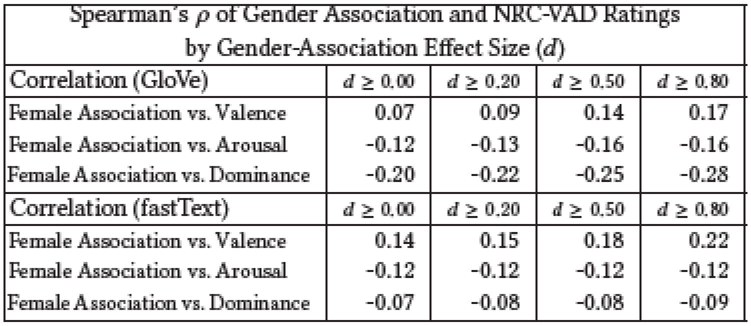


Spearman's ρ of Gender Association and NRC-VAD Ratings by Word Frequency Range (N)

Correlation (GloVe)	$N = 10^2$	$N = 10^3$	$N = 10^4$	NRC-VAD
Female Association vs. Valence	0.15	0.16	0.10	0.07
Female Association vs. Arousal	-0.14	-0.11	-0.13	-0.12
Female Association vs. Dominance	0.05	-0.16	-0.21	-0.20
Correlation (fastText)	$N = 10^2$	$N = 10^3$	$N = 10^4$	NRC-VAD
Female Association vs. Valence	0.02	0.15	0.15	0.14
Female Association vs. Arousal	-0.07	-0.12	-0.11	-0.12
Female Association vs. Dominance	-0.05	-0.10	-0.08	-0.07

Table 6: Female-Associated Words Correlate More Strongly with Valence, While Male-Associated Words Correlate with Arousal and Dominance [58].

This suggests that **words related to men in word embeddings are more likely to evoke feelings of power and intensity**, whereas words related to women tend to be linked with positive emotions but lower dominance. The same pattern persists when broken-down by-**word frequency ranges**, as observed in **Table 7**, which examines correlations by gender-association effect size.



Spearman's ρ of Gender Association and NRC-VAD Ratings by Gender-Association Effect Size (d)

Correlation (GloVe)	$d \geq 0.00$	$d \geq 0.20$	$d \geq 0.50$	$d \geq 0.80$
Female Association vs. Valence	0.07	0.09	0.14	0.17
Female Association vs. Arousal	-0.12	-0.13	-0.16	-0.16
Female Association vs. Dominance	-0.20	-0.22	-0.25	-0.28
Correlation (fastText)	$d \geq 0.00$	$d \geq 0.20$	$d \geq 0.50$	$d \geq 0.80$
Female Association vs. Valence	0.14	0.15	0.18	0.22
Female Association vs. Arousal	-0.12	-0.12	-0.12	-0.12
Female Association vs. Dominance	-0.07	-0.08	-0.08	-0.09

Table 7: Gender Association Effect Size and NRC-VAD Ratings [58].

Gender Bias in Big Tech Terminology

One especially startling example of gender bias is seen in **Big Tech terminology**. Words connected with **technology-related professions, leadership, and innovation** most definitely correlate with male-associated words in both GloVe and fastText embeddings, as seen in **Figure 5**.

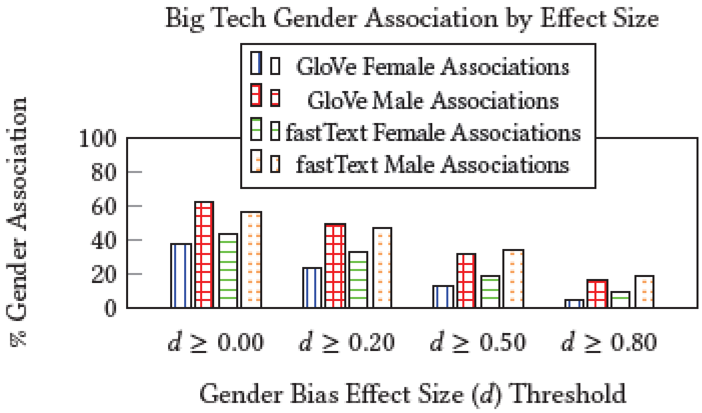


Figure 5: Big Tech Gender Association by Effect Size [58].

This finding is crucial, as **AI-driven hiring platforms, search algorithms, and recommendation systems trained on these embeddings could inadvertently reinforce existing gender disparities in STEM fields**. Addressing such biases is critical for ensuring **fair representation of women in technology-related discussions and AI applications**.

GloVe and fastText static word embeddings, across all levels of analysis, demonstrate a higher association with male-related words than female-related words. Of the **10,000 most frequent words** in the GloVe vocabulary, **1,187** exhibit a large effect size association with men, compared to only **611** with a strong association with women [58]. **fastText embeddings** follow a similar pattern but show a **lower bias**, suggesting that representation of women in corpora has improved over time, particularly from pre-2014 GloVe training data to post-2017 models. However, despite this incremental change, **implicit gender bias remains prevalent** (Figure 6).

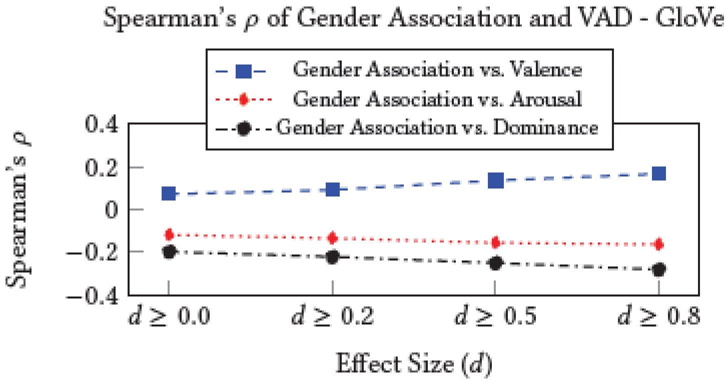


Figure 6: Gender Bias Correlation with Valence, Arousal, and Dominance in GloVe Embeddings [58].

The Role of Bias in AI Applications

The propagation of these **gender representations** in word embeddings significantly impacts downstream AI applications, particularly in **visual-linguistic models** such as OpenAI's **CLIP**. Systems trained on these biased embeddings may associate **non-sexual female-related terms** with explicit content, reinforcing **harmful biases in image classification, content moderation, and hiring algorithms**.

Similarly, an examination of **Big Tech-related words** (Figure 5) reveals that more than **60% of technology-related words in embeddings are associated with men**. This suggests that AI models used in **recruitment, professional networking, and career counseling** may perpetuate existing gender gaps in STEM fields.

Future Directions

Dealing with gender bias in NLP calls for a multimodal strategy combining:

- **Debiasing embeddings** without sacrificing language integrity, hence mitigating bias.
- **Diverse and representative** training data with **gender-balanced word distributions**. Intervention techniques include **ethical artificial intelligence frameworks** and **policy changes** help to stop AI systems from extending gender language prejudices.
- Resolving these problems at both the **algorithmic** and **data** levels will enable researchers and developers to produce **fairer, more inclusive NLP models** reflecting a **balanced representation of gender in AI-driven communications**.

References

[1] Mohamed Abdalla and Moustafa Abdalla. 2021. The Grey Hoodie Project: Big tobacco, big tech, and the threat on academic integrity. In Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. 287–297.

[2] Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual String Embeddings for Sequence Labeling. In COLING 2018, 27th International Conference on Computational Linguistics. 1638–1649.

[3] Mahzarin R Banaji and Anthony G Greenwald. 1995. Implicit gender stereotyping in judgments of fame. Journal of personality and social psychology 68, 2 (1995), 181.

[4] Christine Basta, Marta R Costa-Jussà, and Noe Casas. 2019. Evaluating the underlying gender bias in contextualized word embeddings. arXiv preprint arXiv:1904.08783 (2019).

[5] Emily M Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. Transactions of the Association for Computational Linguistics 6 (2018), 587–604.

[6] Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. 2021. Multimodal datasets: misogyny, pornography, and malignant stereotypes. arXiv preprint

arXiv:2110.01963 (2021).
[7] J Stewart Black and Patrick van Esch. 2020. AI-enabled recruiting: What is it and how should a manager use it? Business Horizons 63, 2 (2020), 215–226.
[8] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics 5 (2017), 135–146.
[9] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. Advances in neural information processing systems 29 (2016), 4349–4357.
[10] Marc-Etienne Brunet, Colleen Alkalay-Houlihan, Ashton Anderson, and Richard Zemel. 2019. Understanding the origins of bias in word embeddings. In International Conference on Machine Learning. PMLR, 803–811.
[11] Aylin Caliskan. 2021. Detecting and mitigating bias in natural language processing. Brookings Institution (2021).
[12] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. Science 356, 6334 (2017), 183–186.
[13] Aylin Caliskan and Molly Lewis. [n. d.]. Social biases in word embeddings and their relation to human cognition. PsyArXiv.
[14] Kaytlin Chaloner and Alfredo Maldonado. 2019. Measuring gender bias in word embeddings across domains and discovering new gender bias word categories. In Proceedings of the First Workshop on Gender Bias in Natural Language Processing. 25–32.
[15] Tessa Charlesworth, Aylin Caliskan, and Mahzarin R. Banaji. 2022. Historical Representations of Social Groups Across 200 Years of Word Embeddings from Google Books. Proceedings of the National Academy of Sciences (2022).
[16] Tessa ES Charlesworth and Mahzarin R Banaji. 2021. Patterns of Implicit and Explicit Stereotypes III: Long-Term Change in Gender Stereotypes. Social Psychological and Personality Science (2021), 1948550620988425.
[17] Tessa ES Charlesworth, Victor Yang, Thomas C Mann, Benedek Kurdi, and Mahzarin R Banaji. 2021. Gender stereotypes in natural language: Word embeddings show robust consistency across child and adult language corpora of more than 65 million words. Psychological Science 32, 2 (2021), 218–240.
[18] Sapna Cheryan and Hazel Rose Markus. 2020. Masculine defaults: Identifying and mitigating hidden cultural biases. Psychological Review 127, 6 (2020), 1022.

[19] Jacob Cohen. 2013. Statistical power analysis for the behavioral sciences. Academic press.

[20] Ronan Collobert, JasonWeston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. Journal of machine learning research 12, ARTICLE (2011), 2493–2537.

[21] Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. 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. 120–128.

[22] Sunipa Dev, Masoud Monajatipoor, Anaelia Ovalle, Arjun Subramonian, Jeff Phillips, and Kai-Wei Chang. 2021. Harms of Gender Exclusivity and Challenges in Non-Binary Representation in Language Technologies. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 1968–1994.

[23] Alice H Eagly and Antonio Mladinic. 1989. Gender stereotypes and attitudes toward women and men. Personality and social psychology bulletin 15, 4 (1989), 543–558.

[24] Alice H Eagly and Antonio Mladinic. 1994. Are people prejudiced against women? Some answers from research on attitudes, gender stereotypes, and judgments of competence. European review of social psychology 5, 1 (1994), 1–35.

[25] Charles Elkan. 2003. Using the triangle inequality to accelerate k-means. In Proceedings of the 20th international conference on Machine Learning (ICML-03). 147–153.

[26] Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. Proceedings of the National Academy of Sciences 115, 16 (2018), E3635–E3644.

[27] Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. arXiv preprint arXiv:1903.03862 (2019).

[28] Anthony G Greenwald, Debbie E McGhee, and Jordan LK Schwartz. 1998. Measuring individual differences in implicit cognition: the implicit association test. Journal of personality and social psychology 74, 6 (1998), 1464.

[29] Wei Guo and Aylin Caliskan. 2021. Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases. In Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. 122–133.

[30] N. Hsu, K. L. Badura, D. A. Newman, and M. E. P. Speach. 2021. Gender,

"masculinity," and "femininity": A meta-analytic review of gender differences in agency and communion. Psychological Bulletin (2021), 987–1011. https:

//doi.org/10.1037/bul0000343

[31] Larry L Jacoby, Colleen Kelley, Judith Brown, and Jennifer Jasechko. 1989. Becoming famous overnight: Limits on the ability to avoid unconscious influences of the past. Journal of personality and social psychology 56, 3 (1989), 326.

[32] Andrew Karpinski and Ross B Steinman. 2006. The single category implicit association test as a measure of implicit social cognition. Journal of personality and social psychology 91, 1 (2006), 16.

[33] Hadas Kotek, Rikker Dockum, Sarah Babinski, and Christopher Geissler. 2021. Gender bias and stereotypes in linguistic example sentences. Language (2021).

[34] Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in Pre-Training Distributed Word Representations. In Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018).

[35] Tomáš Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies. 746–751.

[36] Saif M. Mohammad. 2018. Obtaining Reliable Human Ratings of Valence, Arousal, and Dominance for 20,000 EnglishWords. In Proceedings of The Annual Conference of the Association for Computational Linguistics (ACL). Melbourne, Australia.

[37] Brian A Nosek, Frederick L Smyth, Natarajan Sriram, Nicole M Lindner, Thierry Devos, Alfonso Ayala, Yoav Bar-Anan, Robin Bergh, Huajian Cai, Karen Gonsalkorale, et al. 2009. National differences in gender–science stereotypes predict national sex differences in science and math achievement. Proceedings of the National Academy of Sciences 106, 26 (2009), 10593–10597.

[38] Charles E Osgood. 1964. Semantic differential technique in the comparative study of cultures 1. American Anthropologist 66, 3 (1964), 171–200.

[39] Charles Egerton Osgood, George J Suci, and Percy H Tannenbaum. 1957. The measurement of meaning. Number 47. University of Illinois press.

[40] Amandalynne Paullada, Inioluwa Deborah Raji, Emily M Bender, Emily Denton, and Alex Hanna. 2021. Data and its (dis) contents: A survey of dataset development and use in machine learning research. Patterns 2, 11 (2021), 100336.

[41] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe:

Global Vectors for Word Representation. In Empirical Methods in Natural Language Processing (EMNLP). 1532–1543.
http://www.aclweb.org/anthology/D14-
1162

[42] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. arXiv preprint arXiv:2103.00020 (2021).

[43] Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 2020. Semeval-2020 task 1: Unsupervised lexical semantic change detection. arXiv preprint arXiv:2007.11464 (2020).

[44] Robyn Speer, Joshua Chin, Andrew Lin, Sara Jewett, and Lance Nathan. 2018. LuminosoInsight/wordfreq: v2.2.
https://doi.org/10.5281/zenodo.1443582

[45] Ryan Steed and Aylin Caliskan. 2021. Image representations learned with unsupervised pre-training contain human-like biases. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. 701–713.

[46] Autumn Toney-Wails and Aylin Caliskan. 2021. ValNorm Quantifies Semantics to Reveal Consistent Valence Biases Across Languages and Over Centuries. Empirical Methods in Natural Language Processing (EMNLP) (2021).

[47] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. Journal of machine learning research 9, 11 (2008).

[48] Tianlu Wang, Xi Victoria Lin, Nazneen Fatema Rajani, Bryan McCann, Vicente Ordonez, and Caiming Xiong. 2020. Double-Hard Debias: Tailoring Word Embeddings for Gender Bias Mitigation. In Association for Computational Linguistics (ACL).

[49] Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. 2013. Norms of valence, arousal, and dominance for 13,915 English lemmas. Behavior research methods 45, 4 (2013), 1191–1207.

[50] Laura Wendlandt, Jonathan K. Kummerfeld, and Rada Mihalcea. 2018. Factors Influencing the Surprising Instability of Word Embeddings. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). Association for Computational Linguistics, New Orleans, Louisiana, 2092–2102.

https://doi.org/10.18653/v1/N18-1190

[51] Robert Wolfe, Mahzarin R. Banaji, and Aylin Caliskan. 2022. Evidence for Hypodescent in Visual Semantic AI. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT).
[52] Robert Wolfe, Mahzarin R Banaji, and Aylin Caliskan. 2022. Evidence for Hypodescent in Visual Semantic AI. arXiv preprint arXiv:2205.10764 (2022).
[53] Robert Wolfe and Aylin Caliskan. 2021. Low Frequency Names Exhibit Bias and Overfitting in Contextualizing Language Models. Proceedings of the 2021 conference on empirical methods in natural language processing (EMNLP) (2021).
[54] RobertWolfe and Aylin Caliskan. 2022. Markedness in Visual Semantic AI. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT).
[55] Robert Wolfe and Aylin Caliskan. 2022. VAST: The Valence-Assessing Semantics Test for Contextualizing Language Models. In Proceedings of the 36th AAAI Conference on Artificial Intelligence (AAAI).
[56] Robert B Zajonc. 2001. Mere exposure: A gateway to the subliminal. Current directions in psychological science 10, 6 (2001), 224–228.
[57] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender bias in contextualized word embeddings. arXiv preprint arXiv:1904.03310 (2019).

[58]Aylin Caliskan, Pimparkar Parth Ajay, Tessa Charlesworth, Robert Wolfe, and Mahzarin R. Banaji. 2022. Gender Bias in Word Embeddings: A Comprehensive Analysis of Frequency, Syntax, and Semantics. In Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (AIES '22). Association for Computing Machinery, New York, NY, USA, 156–170. https://doi.org/10.1145/3514094.3534162

Gender Bias in Word Embeddings: A Comprehensive Analysis of...

The statistical regularities in language corpora encode well-known social biases into word embeddings. Here, we focus on gender to provide a comprehensive analysis of group-based biases in...

https://arxiv.org/abs/2206.03390

Appendix

ContentWarning: The authors of the paper have mentioned the following clusters contain the most frequent 1,000 female-associated and male-associated words in the lexicon with effect sizes $d \geq 0.50$. The results might be triggering [58].

A.1 Female Associated Clusters

Advertising Words: ABOUT, BEST, CLICK, CONTACT, FIRST, FREE, HERE,HOT, LOVE,MAC, MORE, NEXT,NOTE,NOW, OPEN, OR, OTHER, PLUS, SAVE, SEE, SPECIAL, STAR, TODAY, TWO, VERY, WITH, WOW

Beauty and Appearance: accessories, adorable, adore, attractive, beads, beautiful, beauty, boutique, bracelet, bridal, bride, butterfly, candles, ceramic, charming, chic, clothes, clothing, coats, cocktail, colorful, colors, colours, coral, costume, costumes, crafts, crystal, crystals, cute, dancers, decor, decorating, decorations, delicate, delightful, designer, designs, doll, dolls, dress, dresses, earrings, elegance, elegant, ensemble, exotic, exquisite, fabric, fabrics, fabulous, fairy, fashion, fashionable, feminine, floral, flower, flowers, footwear, fragrance, gorgeous, gown, hair, handbags, handmade, heels, invitations, jewellery, jewelry, knit, knitting, lace, ladies, lip, lovely, makeup, metallic, necklace, nylon, outfit, outfits, paired, pale, pattern, pearl, pendant, perfume, pillow, pink, platinum, polish, princess, prom, purple, purse, quilt, ribbon, romantic, roses, satin, scent, sewing, sheer, silk, skirt, sleek, stitch, stunning, styling, stylish, sweater, themed, tiny, trendy, vibrant, Vuitton, waist, wardrobe, wear, wedding, weddings, wonderful, yarn

Celebrities and Modeling: AMI, authors, availability, bio, blogger, bloggers, blogs, bookmark, browse, cell, checkout, class, classes, clicks, cluster, collections, contacting, coordinates, cosmetic, coupon, curves, date, designers, dot, engagement, giveaway, goodies, inexpensive, info, invites, layout, libraries, litter, magazine, magazines, markers, matching, measurements, membrane, model, modeling, models, ms, newest, null, on-line, patterns, peek, photograph, photos, photostream, pictures, pumps, registry, reserved, royalty, sample, samples, scans, separated, shipping, shopping, shops, spanish, stamps, stores, strips, supermarket, swap, temp, template, templates, trends, triangle, updated, websites, widget

Cooking and Kitchen: bake, Bake, baked, baking, cake, cakes, chocolate, Chocolate, cinnamon, coconut, Cookies, Cooking, cream, creamy, crust, cups, dairy, delicious, Delicious, dessert, dressing, egg, eggs, foods, Ginger, homemade, honey, Ingredients, lemon, milk, Milk, nutrition, organic, pumpkin, recipe, Recipe, recipes, Recipes, Salad, spice, spicy, sugar, sweet, tea, teaspoon, tsp, vanilla, vegan, vegetables, vegetarian, veggies, yogurt, yummy

Fashion and Lifestyle: Autumn, Bag, Bags, Ballet, Barbie, Basket, Bathroom, Beads, Beautiful, Beauty, Bed, Bedroom, Bee, Bottom, Boutique, Bracelet, Bridal, Bride, Butterfly, Cake, Candle, Candy, Carnival, Carpet, Ceramic, Charm, Cherry, Clearance, Clothes, Colors, Compact, Contemporary, Cookie, Coral, Costume, Cottage, Covers, Crafts, Cream, Crystal, Daisy, Dance, Dancing, Decor, Designer, Designs, Desk, Diamonds, Dining, DIY, Doll, Dolls, Dreams, Dress, Dresses, Earrings, Egg, Emerald, Evening, Fabric, Fairy, Fancy, Fashion, Favorites, Fiber, Floor, Floral, Flower, Flowers, Giveaway, Gorgeous, Hair, Halloween, Heart, Heated, Honey, Inspired, Jeans, Jewelry, Kiss, Kitty, Lace, Ladies, Laundry, Layer, Lovely, Loving, Luxury, Makeup, Mesh, Metallic, Mint, Mirror, Mirrors, Nail, Natural, Necklace, Nylon, Passion, Pattern, Patterns, Pearl, Perfect, Picture, Pillow, Pink, Platinum, Plus, Powder, Pretty, Princess, Printed, Pump, Queen, Rack, Ribbon, Romance, Romantic, Rose, Roses, Ruby, Salon, Satin, Shades, Shape, Shipping, Shoes, Shoulder, Shower, Silk, Simply, Skin, Sleeping, Smile, Soap, Soft, Spice, Split, Style, Sugar, Summer, Sunny, Swan, Sweet, Swim, Tea, Tops, Tote, Trend, Trim, Tropical, Twilight, Unique, Valentine, Vampire, Vanity, Venus, Victorian, Vintage, Wear, Wedding, Weddings, Witch, Womens, Wonderful, Wrap, ~, r

Female Names: Abbey, actress, Actress, Alice, Allison, Amanda, Amber, Amy, Ana, Andrea, Angela, Angie, Ann, Anna, Anne, Annie, Ashley, Barbara, Bella, Belle, Beth, Betty, Beverly, Bonnie, Britney, Brooke, Buffy, Carol, Caroline, Carrie, Catherine, Charlotte, Cheryl, Christina, Christine, Cindy, Claire, Clara, Clare, Courtney, Dana, Dawn, Debbie, Deborah, Denise, Diana, Diane, Donna, Dorothy, Elizabeth, Ellen, Emily, Emma, Erin, Eva, Eve, Gaga, Grace, Heather, Helen, Hilton, Holly, Ivy, Jackie, Jane, Janet, Jen, Jennifer, Jenny, Jessica, Jill, Jo, Joan, Joy, Judy, Julia, Julie, Karen, Kate, Katherine, Kathleen, Kathy, Katie, Katrina, Katy, Kay, Kelly, Kim, Kristen, Lady, Laura, Lauren, Lily, Linda, Lindsay, Lisa, Liz, Louise, Loved, Lucy, Lynn, Madison, Madonna, Mae, Maggie, Mai, Mama, Margaret, Maria, Marie, Marilyn, Marina, Married, Mary, Maya, Megan, Melissa, Mercedes, Met, Michelle, Miss, Molly, Mommy, Monica, Ms, Ms., Nancy, Natalie, Nicole, Nikki, Nina, Olivia, Pam, Patricia, Paula, Penny, Rachel, Rebecca, Rihanna, Rosa, Sally, Samantha, Sandra, Sara, Sarah, Savannah, Sharon, Shirley, singer, Sister, Sisters, Sophie, Spears, Stephanie, Sue, Susan, Tara, Tiffany, Tina, Vanessa, Victoria, Wendy, Whitney, Willow, xx, Yay

Health and Relationships: abortion, acne, addicted, addiction, allergic, allergy, arthritis, aunt, babies, belly, breast, cancer, caring, celebrities, celebrity, chatting, cheating, clinic, complications, counseling, couple, couples, dancer, DD, depressed, depression, diabetes, disabilities, disorder, distress, donor, emotionally, experiencing, girlfriend, grandmother, healthier, her, hers, herself, hips, hormones, inspirational, lady, literacy, lover, loving, marriage, messy, mom, moms, mother, mothers, mum, nurse, nurses, nursing, obsessed, obsession, oral, parent, parenting, passionate, poems, pose, poses, pregnancy, pregnant, protective, relationship, relationships, romance, seniors, sensitive, sexuality, she, She, sister, sisters, skin, stories, stressful, supportive, survivors, syndrome, therapist, therapy, toes, toxic, tumor, vampire, witch, wives, woman, women

Luxury and Lifestyle: accommodations, Apartment, balcony, bath, bathroom, bathrooms, beaches, bedroom, carpet, catering, closet, cottage, cozy, cruise, Enjoy, enjoys, flats, gardening, holidays, intimate, kitchen, laundry, luxurious, luxury, massage, mattress, outdoors, pets, queen, relaxing, rooms, Rooms, salon, sandy, shower, showers, soap, Spa, spa, sunny, swim, swimming, tile, tub, vacations, wellness, yoga

Obscene Adult Material: anal, babe, babes, bikini, bitch, blonde, blowjob, boob, boobs, bra, breasts, brunette, busty, chick, chicks, cum, cunt, dildo, ebony, erotic, escort, facial, flashing, fucked, gal, galleries, gallery, girl, girls, hentai, horny, hot, hottest, juicy, kissing, latex, lesbian, lesbians, lick, licking, lingerie, mature, milf, movies, naked, naughty, nipples, nude, orgasm, panties, penetration, pics, posing, pussy, sexy, shemale, slut, stockings, sucking, teen, teens, tit, tits, webcam, wet, whore, xxx

Sexual Profanities: 00, Amateur, Anal, Asian, Ass, Babe, Blonde, Busty, Cum, Cute, Ebony, Facial, Fucked, Galleries, Girl, Girls, Her, Horny, Hot, Huge, Lesbian, Mature, Mom, Moms, Movies, Naked, Nude, Pics, Pictures, Porn, Pussy, Sex, Sexy, Teen, Teens, Tight, Tits, Wet, Wife, XXX

Web Article Titles: Absolutely, Acne, Across, Addiction, Adelaide, Advertise, Affordable, Alberta, Apply, Aurora, Awareness, Bachelor, Benefit, Biggest, Blogger, Bollywood, Breast, Calendar, Cancel, Cancer, Caribbean, Celebrity, Changing, Choice, Choosing, Classes, Closed, Collections, Compliance, Consumer, Consumers, Contest, Coordinator, Counseling, Created, Cruise, Cure, Czech, Dakota, Dates, Denmark, Designers, Destination, Diabetes, Diet, Disclaimer, Eating, eBook, Editorial, Engagement, ER, Everyday, Exclusive, Explore, Factor, Fiction, Finding, Fitness, Food, Foods, Gallery, Getting, Health, Healthy, Holidays, Inspiration, Languages, Libraries, Lifestyle, Lots, Magazine, Massage, Model, Models, Month, MS, MSN, Multiple, Naturally, Newsletter, non-profit, nonprofit, Novel, Nurse, Nursing, Nutrition, Oral, Parties, Patent, Patient, Platform, Pregnancy, Privacy, Purchase, Readers, Reality, Recently, Reception, Registry, Relationship, Relationships, Reserved, Rica, Runtime, Sample, Scenes, Secret, Secrets, Seller, Shared, Shares, Sharing, Shows, Sierra, Sites, Spotlight, Statement, Student, Target, Teacher, Teachers, Templates, Therapy, Totally, Treat, Trends, Updates, VIP, Virgin, Virtual, Vitamin, Voices, Wellness, Whole, Winners, Women, Write, Yoga

A.2 Male Associated Clusters

Adventure and Music: ", ’, 1972, Against, Answer, Arms, Articles, Back, Band, Bass, Batman, Battle, Bear, Beat, Beer, Blues, Brain, Brother, Brothers, Bull, Camp, Champion, Cold, Comedy, Cool, Count, Crew, Da, Dead, Death, Devil, Die, DJ, Dog, Dragon, Eagle, Empire, End, EP, Essential, Evil, Evolution, Fans, feat, Fight, Fish, Flying, Force, Four, Future, Game, Ghost, Giant, Great, Green, Guitar, Gun, Guys, Hat, Head, Hero, Hood, II, III, Iron, IV, Jazz, Jump, King, Kingdom, Kings, Knight, Late, Leader, Legend, Lincoln, Lion, LP, Major, Man, Mario, Marvel, Master, Max, Military, Motion, Nation, Navy, Numbers, Of, Official, Orchestra, Original, Oxford, Pack, Part, Pass, Points, Prime, Prince, Quote, Rank, Raw, Records, Remix, remix, Reserve, Retrieved, Return, Revolution, Rise, Rock, Rocky, Roll, Rule, Running, Rush, Score, Scottish, Shirt, Shot, Six, Sound, Stand, Strong, Super, Ten, Tiger, Trail, Trial, Ultimate, views, Vol, Volume, Wall, War, Wars, Way, Will, Wolf

Big Tech: .0, 1.1, 2.0, 3.0, Android, Answers, API, App, Applications, Audio, Build, Canon, Cisco, Cloud, Command, Computer, contribs, CPU, demo, developer, Developer, developers, Documents, Error, Firefox, Flash, Forums, Galaxy, Gaming, GMT, Google, GPS, HP, IBM, Install, Intel, Intelligence, Internet, Introduction, iOS, iPhone, Java, JavaScript, Linux, Message, Microsoft, MP3, NET, Nintendo, Notes, OS, PC, Player, plugin, Problem, Programming, PS3, Questions, Re, RE, Remote, replies, RSS, Samsung, Security, SEO, Server, SMS, Software, SQL, Statistics, Test, User, Users, Wii, Windows, Wireless, Xbox, XML, XP, YouTube

Engineering and Automotive: AC, Advance, Audi, Auto, Automotive, Bar, Battery, BMW, Built, Button, Cap, Charger, Chevrolet, Chrome, Circuit, Construction, Contractors, Custom, Dodge, Doors, Driver, Driving, Duty, Economy, Electric, Engine, Engineer, Engineering, Equipment, Extra, Fishing, Fuel, Garage, Gas, Gear, General, GM, Golf, Guard, Hardware, Heating, Heavy, Honda, Industrial, Laser, Logo, Machine, Maintenance, Manual, Manufacturing, Metal, Motor, Nissan, Oil, Pocket, Portable, Power, Premium, Pressure, Printing, Pro, Quick, Racing, RC, Repair, Rod, Signs, Solar, Solid, Speed, Sport, Standard, Steel, System, Tech, Technical, Tool, Tools, Toyota, Trade, Trading, Training, Transfer, Transport, Truck, Universal, Upper, Wood, Yamaha

Engineering and Electronics: assembly, audio, auto, automotive, backup, batteries, blade, brass, build, built, capable, charge, charging, chip, circuit, command, commands, computing, conditioning, console, construction, contractor, contractors, controller, conversion, convert, converted, custom, dealer, dealers, driver, durable, duty, electronics, enabled, engine, engineer, engineering, engineers, engines, enterprise, execution, formation, gate, gear, general, generation, header, install, legacy, lightweight, logic, manual, master, motor, operation, physics, pipe, power, printer, printing, proven, receiver, reference, remote, repair, replace, replacement, restoration, rod, root, scheme, seal, security, setup, software, solution, superior, suspension, tire, transfer, trucks, upgrade

God and Religion: Abraham, according, According, Allah, appointed, authority, bear, bears, believed, Bible, blind, brothers, century, Christ, Christianity, Christians, commentary, composed, creator, evil, evolution, Father, favor, followers, fool, genius, glory, God, god, Gospel, he, He, himself, His, Holy, holy, hundred, Islam, Israel, Jerusalem, Jesus, Jews, king, kingdom, land, Lord, man, mere, Muslims, nations, passage, philosophy, poor, Pope, possession, praise, principle, quote, referred, refers, regard, regarded, respect, reward, Roman, Rome, rule, sacrifice, sheep, sin, sir, Son, sword, temple, theory, tho, thou, Thus, tradition, translation, united, unto, verse, wise, worthy, ye

Male Names: Aaron, actor, Adam, Al, Alan, Albert, Alex, Allen, Andrew, Andy, Anthony, Arthur, Barry, Ben, Bill, Billy, Bishop, Bob, Bobby, Brad, Brandon, Brian, Brown, Bruce, Bryan, Captain, Carl, Carlos, CEO, Chairman, chairman, Charles, Chief, Chris, Christopher, Chuck, Clay, Craig, Dan, Daniel, Danny, Dave, David, Dennis, Dick, Don, Donald, Doug, Duke, Ed, Eddie, Eric, Francis, Frank, Franklin, Fred, Gary, Gates, George, Glenn, Gordon, Governor, Greg, Guy, Harrison, Harry, Henry, Howard, Ian, Jack, Jackson, Jacob, Jake, James, Jason, Jay, Jeff, Jefferson, Jeremy, Jerry, Jim, Jimmy, Joe, Joel, John, Johnny, Johnson, Jon, Jonathan, Joseph, Josh, Jr., Juan, Justin, Keith, Ken, Kevin, Kyle, Larry, Luke, Marc, Mark, Marshall, Martin, Matt, Matthew, Mayor, Michael, Mike, Miles, Morris, Mr, Mr., Murray, Nathan, Neil, Nelson, Nick, Norman, Oliver, Patrick, Paul, Pete, Peter, Phil, Philip, Ralph, Randy, Rich, Richard, Rick, Rob, Robert, Robinson, Roger, Ron, Roy, Russell, Ryan, Sam, Samuel, Scott, Sean, Simon, Sir, Stanley, Stephen, Steve, Steven, Ted, Terry, Thomas, Tim, Tom, Tommy, Tony, Troy, Victor, Vincent, W., Walter, Wayne, William

Non-English Tokens: al, Barcelona, con, da, DE, del, der, des, di, du, e, ed, El, el, et, le, Madrid, o, par, que, se, un, van

Numbers, Dates, and Metrics: -1, 103, 111, 113, 1500, 160, 2.4, 200, 220, 240, 250, 2d, 300, 3000, 320, 360, 400, 450, 500, 51, 600, 700, 73, 77, 900, [, acres, BC, C, c., D, d, ft, ft., G, Given, k, MP, No., O, OF, P, p, p., Per, pp., R, SS, St, U., v, v., W, £

Sports: backs, ball, band, baseball, basketball, bass, bat, beat, beaten, beating, beer, bench, betting, blues, boss, buddy, camp, captain, champion, championship, cheat, coach, coaches, coin, crew, decent, defensive, don, draft, drum, drums, dude, elite, epic, era, fans, fellow, finest, fishing, football, franchise, gambling, game, games, gaming, golf, grand, great, greatest, guard, guitar, guy, guys, heads, hero, hockey, hunting, idiot, injuries, injury, jazz, jersey, jokes, kick, league, legend, legendary, lineup, manager, mark, mate, minor, musicians, offense, offensive, pass, passes, passing, penalty, pit, pitch, player, players, points, pound, premier, prime, pro, prospect, prospects, racing, rally, rank, recruiting, retired, rotation, rush, saves, score, scored, scoring, serving, solid, sport, sports, squad, stadium, starter, stats, suspended, tackle, team, teams, thread, ton, tournament, trade, trading, tribute, tricks, ultimate, versus, veteran, victory, wing, yard, yards, zone

Sports and Cities: 2014, AL, Antonio, Arena, Athletic, Baltimore, Baseball, Basketball, Bay, Bears, Boston, Bowl, Buffalo, Champions, Championship, Chicago, Cincinnati, Cleveland, Columbus, Dallas, Detroit, Diego, Draft, Eagles, England, ESPN, FC, Football, Giants, Highlights, Hockey, Indians, Jersey, Jose, Junior, League, Lions, Liverpool, Louis, Louisville, Manchester, Milwaukee, Minnesota, MLB, MLS, Montreal, NBA, NCAA, NFL, NHL, Nike, Oakland, Orlando, Penn, Philadelphia, Pittsburgh, Players, Premier, Rangers, Saints, San, SEC, Soccer, Sox, Sports, St., Stadium, Tampa, Team, Ticket, Tickets, Tigers, Tournament, United, vs, vs., Yankees

War and Violence: against, Army, army, arrest, arrested, attack, ban, battle, bin, Bush, charges, chief, cited, combat, commit, committed, corruption, crimes, criminal, dead, defeat, defeated, defense, Defense, destruction, enemies, enemy, executed, fight, fighter, fighting, fights, fought, fraud, governor, gun, guns, heroes, illegal, injured, intelligence, Iraq, Iraqi, kill, killed, killing, leader, leaders, leadership, led, march, military, minister, officers, opponents, opposition, personnel, prison, province, racist, regime, revolution, ruled, soldier, soldiers, spokesman, supporters, tactics, terror, terrorist, troops, veterans, violent, war, wars, weapons

A.3 Big TechWords

965 Big Tech Words: 23andMe, 3Com, 3COM, 3Par, 3PAR, 7digital, 9to5Google, 9to5mac, 9to5Mac, AAPL, ABBYY, Accenture, Acer, Acronis, Activision, Acxiom, AdAge, Adaptec, Adidas, AdMob, Admob, Adobe, AdSense, Adsense, AdWords, Adwords, Agilent, Airbnb, Airbus, Airtel, Akamai, Albanesius, Alcatel, Alcatel-Lucent, Alibaba, Alibaba.com, Alienware, AllFacebook, AllThingsD, AltaVista, Altera, Amazon, Amazon.com, AMD, Amdocs, AmEx, AMZN, Anandtech, AnandTech, Andoid, Andreessen, Andriod, Android, ANDROID, android, Android-based, Android-powered, AndroidPIT, anti-competitive, anti-trust, anticompetitive, Antitrust, antitrust, AOL, AOpen, API, APIs, Appcelerator, AppEngine, Apple, APPLE, Apple.com, AppleInsider, AppleTV, Appstore, appstore, AppStore, AppUp, Archos, Ariba, ARM-based, Ask.com, ASRock, AstraZeneca, Asus, ASUS, asus, ASUSTeK, Asustek, ATandT, Atari, Atheros, ATi, ATI, Atlassian, Atmel, Atom-based, Atos, Atrix, AuthenTec, Autodesk, automaker, Automattic, Avanade, Avaya, Avira, Avnet, AWS, Baidu, baidu, Baidu.com, Ballmer, Barclays, Bazaarvoice, BBRY, BenQ, BestBuy, Bestbuy, BetaNews, Betriebssystem, Bezos, BIDU, Bing, Biogen, Bitcoin, Bitdefender, BitTorrent, BlackBerry, Blekko, Blinkx, bloatware, BloggingStocks, Bloomberg, BlueStacks, Boeing, BofA, Box.net, Boxee, Brightcove, Broadcom, Brocade, BSkyB, Bungie, BusinessWeek, Buy.com, BuzzFeed, BYD, Canalys, Canonical, Capgemini, carmaker, Carphone, CCleaner, CentOS, ChannelWeb, Chegg, China, China-based, Chinavasion,

chip-maker, Chipmaker, chipmaker, chipmakers, chipset, chipsets, Chipzilla, Chitika, Chromebook, ChromeBook, Chromebooks, ChromeOS, CinemaNow, CIO.com, Cisco, CISCO, CISPA, Citi, Citibank, Citigroup, Citrix, Cleantech, Clearwire, closed-source, cloud-computing, Cloudera, CNET, CNet, Cnet, cnet, Coca-Cola, Cognizant, Comcast, Compal, companies, company, Compaq, CompTIA, ComputerWorld, Computerworld, Computex, ComScore, Comscore, comScore, Conexant, Cooliris, Corp, Costolo, Coursera, Cr-48, crapware, Cringely, CrunchBase, CSCO, CUDA, Cupertino, Cupertinobased, CyanogenMod, Cyanogenmod, CyberLink, Cyberlink, Cybersecurity, cybersecurity, D-Link, DailyTech, Daimler, Danone, DARPA, Datacenter, Deezer, Dell, DELL, Deloitte, DeNA, Dhingana, DigiTimes, Digitimes, DisplayLink, DisplayPort, DivX, DoCoMo, Docomo, DOCOMO, DoJ, DOJ, DoubleClick, Doubleclick, DreamHost, Dropbox, DropBox, E-Commerce, E-Readers, EBay, eBay, Ebay, Ebuyer, eCommerce, Ecosystem, ecosystem,

Electricpig.co.uk, Electronista, Elop, Eloqua, eMachines, Emachines, eMarketer, EMC, Emulex, Endeca, Engadget, engadget, Epson, Ericsson, Erictric, ESET, Esri, Etisalat, Everex, Evernote, EVGA, eWeek, Experian, ExtremeTech, Exxon, ExxonMobil, Exynos, F-Secure, Facebook, FaceBook, Facebooks, FedEx, Feedly, Firefox, Flextronics, Flipboard, Flipkart, Fortinet, FOSS, Foxconn, foxconn, Foxit, FreeBSD, Freescale, Frito-Lay, FTC, Fudzilla, Fujifilm, Fujitsu, Fusion-io, GadgeTell, Gaikai, Gameloft, GameStop, Gartner, Gawker, GE, Geek.com, GeekWire, GeForce, Geforce, Gemalto, Genentech, Geohot, GetJar, Gigabyte, GIGABYTE, GigaOm, GigaOM, GitHub, Github, Gizmodo, Glassdoor, GlaxoSmithKline, Gmail, GMail, GMAIL, go-tomarket, GoDaddy, Godaddy, GoGrid, GOOG, Google, GOOGLE, Google-owned, Google.com, Googler, Googlers, Googles, GoogleTV, Googleâ, Goolge, GoPro, gOS, GottaBeMobile.com, Gowalla, Gphone, GPU, GPUs, Groupon, GSK, GSMA, GSMArena, H-P, Hackathon,

Hackintosh, hackintosh, Hadoop, Haier, Hanvon, HD-DVD, Heroku,
Hewlett-Packard, Hisense, Hitachi, Honeywell, Hootsuite, Hortonworks,

HotHardware.com, Hotmail, HP, HPQ, HSBC, HTC, htc,
HTML5, Huawei, huawei, HUAWEI, HubSpot, Hulu, Hynix, I.B.M.,
i7500, IaaS, iAd, iAds, IBM, ibm, IBMs, Icahn, iClarified, iCloud,
Ideapad, IDEOS, IDG, IE10, IE8, IE9, iFixit, iMessage, Informatica,
InformationWeek, Infosys, InfoWorld, Inktomi, INTC, Intel,
INTEL, Intel-based, Intels, InterDigital,
internetnews.com, Internet-

News.com, InterVideo, Intuit, Inventec, Iomega, iOS, iPad3, iPhone,
iPhone5, iPhones, IPO, iRobot, iTablet, ITProPortal, iTWire, ITworld.
com, iWatch, iWork, JBoss, JetBlue, Jolicloud, Joyent, JPMorgan,

JR.com, Kaltura, Kaspersky, KDDI, Kinect, Klout, Kobo,
KPMG, LastPass, Lenovo, lenovo, LENOVO, LePhone, Lexmark,
LG, Liliputing, LiMo, Lindows, LinkedIn, Linkedin, Linksys, Linspire,
Linux, Lite-On, Livescribe,
LiveSide.net, Lodsys, Logitech,
LogMeIn, Lucasfilm, Lucent, Lufthansa, Lumia, Lytro, MacDailyNews,
MacMall, MacOS, MacRumors, Magento, MakerBot, Malware,
Malwarebytes, Marketshare, marketshare, Marvell, Mashable,
MasterCard, Mastercard, Mattel, McAfee, McKesson, McKinsey,
McNealy, MediaTek, Mediatek, Medion, Meebo, MeeGo, Meego,
Meizu, Mellanox, Mendeley, Merck, MetroPCS, Microchip, Microelectronics,
Micron, Microsft, Microsoft, MicroSoft, MIcrosoft, microsoft,
MICROSOFT, Microsofts, MicroStrategy, Micrsoft, Mircosoft,
MIT, Mitel, mobile-device, MobileCrunch, MobiTV, mocoNews,
Monoprice, Monsanto, Motherboard, Moto, Motorola, motorola,
Motorolla, Mozilla, Mozy, MSFT, multinationals, Multitouch,
multitouch, MVNO, MySQL, Napster, NASA, Nasdaq, NASDAQ,
Navteq, Neowin,
Neowin.net, Nestle, Nestlé, NetApp, Netbook,
Netezza, Netflix, NetFlix, Netgear, NetGear, NETGEAR, Netscape,
NetSuite, Newegg, NewEgg, newegg,

Newegg.com, NewEgg.com,

news.cnet.com, NewsFactor, Nextag, Nexus, Nike, Nimbuzz, Nintendo, Nokia, nokia, NOKIA, non-Apple, Nortel, Novartis, Novell, NSA, NSDQ, Nuance, NVDA, Nvidia, NVIDIA, NVidia, nVidia, nVIDIA, nvidia, NXP, OCZ, OEM, OEMs, OLED, OLPC, Omniture, OmniVision, Onkyo, OnLive, Onlive, Open-Source, open-source, open-sourced, OpenCL, OpenDNS, OpenFeint, OpenSocial, Open-Solaris, opensource, OpenStack, Optus, Oracle, ORCL, Orkut, OSes, OSX, Otellini,
Outlook.com, Ouya, OUYA, Overstock.com, PaaS, paidContent, PalmOne, Panasonic, PandoDaily, Pantech, Papermaster, patent-infringement, PayPal, Paypal, PCMag,
PCMag.com,
PCWorld, Pegatron, Pepsi, PepsiCo, Pepsico, Pfizer, Phablet, phablet, PhoneArena, Phoronix, Pichai, Pixar, Pixel, Plantronics, Plaxo,

Play.com, Playdom, Pogoplug, Polycom, PopCap, post-PC, Postini, PowerDVD, Powerset, PowerVR, pre-IPO, PS4, Psystar, Publicis, PwC, QCOM, Qihoo, QLogic, QNAP, Quad-Core, Qualcomm, qualcomm, QUALCOMM, Quantcast, Quickoffice, QuickOffice, Quora, Rackspace, RackSpace, Radeon, Rakuten, Ralink, Rambus, Raytheon, Razer, Rdio, ReadWriteWeb, RealNetworks, Realtek, Redbox, Reddit, RedHat, Redhat, Redmond-based, Renesas, Renren, RHEL, RightScale, RIM, RIMM, Roku, Rovio, SaaS, Safaricom, Salesforce, SalesForce, salesforce,
Salesforce.com, SalesForce.com, salesforce.com, Samsung, samsung, SAMSUNG, Samsungs, SanDisk, Sandisk, Sanofi, SAP, Scoble, Scobleizer, SDK, SDKs, Seagate, search-engine, Seesmic, Semiconductor, Set-Top, SGX, Shopify, Silicon, SiliconANGLE, SiliconBeat, Singtel, SingTel, Sinofsky, SkyDrive, Skype, Slashdot, Smartphone, smartphone, smartphones, Smartphones, smartwatch, SoftBank, Softbank, Softpedia, software, SolarCity, SonicWall, Sonos, Sony, sony, SONY, Sophos, SoundCloud, SourceForge, SpaceX, Spansion, Splashtop, Splunk, Spotify, Spreadtrum, Sprint-Nextel, Starbucks, Stardock, startups, Startups, STMicroelectronics, Success-

Factors, SugarCRM, SugarSync, Sumsung, Sunnyvale, SunPower, Supermicro, superphone, SUSE, SuSE, SwiftKey, Swisscom, Swype, Sybase, Symantec, Synaptics, Synnex, Synopsys, T-Mobile, T-mobile, TalkTalk, Taobao, tech, TechCrunch, Techcrunch, techcrunch,

techcrunch.com, Techdirt, TechEye, TechFlash, TechHive, Techmeme, TechNewsWorld, TechnoBuffalo, TechRadar, TechSpot, TechWeb, Tegra, telco, Telco, telcos, Telcos, Telefonica, Telefónica, Telenor, TeliaSonera, Telstra, Tencent, Teradata, Tesco, Tesla, ThinkGeek, Thinkpad, ThinkPad, TIBCO, Tibco, Ticketmaster, TigerDirect, TiVo, Tizen, TMobile, TomTom, Torvalds, Toshiba, toshiba, TouchPad, Touchpad, Toyota, Transmeta, TRENDnet, TSMC, Tudou, Turkcell, Twilio, Twitter, Uber, Ubergizmo, Ubisoft, Ubuntu, Udacity, UEFI, Ultrabook, ultrabook, Ultrabooks, ultrabooks, Unilever, Unisys, uTorrent, UX, V3.co.uk, Valleywag, VatorNews, VentureBeat, VeriFone, Verisign, VeriSign, Verizon, Vertica, Vevo, Viacom, Viadeo, Viber, Vidyo, ViewSonic, Viewsonic, VirnetX, VirtualBox, Visa, Vizio, VIZIO, Vlingo, VMware, VMWare, Vmware, Vodafone, Vodaphone, Volusion, VP8, VPN, vPro, VR-Zone, Vringo, Vuze, Vyatta, VZW, W3C, Wacom, Wal-Mart, Walmart, Walmart.com, Waze, WebEx, Webkit, WebKit, WebM, WebOS, webOS, Webroot, Websense, Weibo, WhatsApp, Whatsapp, WiDi, WikiLeaks, Wikileaks, WildTangent, WiMax, WiMAX, Win7, Win8, WinBeta, Windows, WIndows, Windows7, Windows8, WinRumors, Wintel, Wipro, Wistron, WMPoweruser, Woz, WP7, WP8, WSJ, WWDC, x86, X86, Xbox, XBox, XCode, XDA, Xerox, Xiaomi, Xilinx, Xobni, Xoom, XOOM, Xperia, Yahoo, Yammer, Yandex, Yarow, Yelp, YHOO, Youku, YouTube, Zappos, ZDNet, ZDnet, Zenbook, Zendesk, Zillow, Zimbra, Zoho, Zotac, ZTE, Zuckerberg, Zynga