Masculine Defaults via Gendered Discourse in Podcasts and Large Language Models

Maria Teleki, Xiangjue Dong, Haoran Liu, James Caverlee
Texas A&M University
{mariateleki, xj.dong, liuhr99, caverlee }@tamu.edu

Abstract

We define *masculine discourse words* as discourse terms that are both socially normative and statistically associated with male speakers. We propose a twofold framework for (i) the large-scale discovery and analysis of gendered discourse words in spoken content via our *Gendered Discourse Correlation Framework*; and (ii) the measurement of the gender bias associated with these words in LLMs via our *Discourse Word-Embedding Association Test*. We focus our study on podcasts, a popular and growing form of social media, analyzing 15,117 podcast episodes. We analyze correlations between *gender* and *discourse words* – discovered via LDA and BERTopic. We then find that gendered discourse-based masculine defaults exist in the domains of business, technology/politics, and video games, indicating that these gendered discourse words are socially influential. Next, we study the representation of these words from a state-of-the-art LLM embedding model from OpenAI, and find that the masculine discourse words have a more stable and robust representation than the feminine discourse words, which may result in better system performance on downstream tasks for men. Hence, men are rewarded for their discourse patterns with better system performance – and this embedding disparity constitutes a representational harm and a masculine default.

Masculine defaults are a type of gender bias "in which characteristics and behaviors associated with the male gender role are valued, rewarded, or regarded as standard, normal, neutral, or necessary aspects of a given cultural context" (Cheryan and Markus, 2020), and hence result in the *other-ing* of women (Beauvoir, 1949).

There is a research gap in identifying and analyzing masculine defaults that arise through *gender differences*[1] *in discourse*. Specifically, we focus on patterns of discourse in spoken communication, including fillers (e.g., *uh*, *um*), discourse markers (e.g., *well*, *you know*, *I mean*), false starts (e.g., *It was, anyways, I went to Target yesterday*) and more (Merriam-Webster, 2024; Shriberg, 1994).

Such discourse words are non-content related words that serve important social purposes with respect to gender, such as to *"hold the floor"* in conversation (Shriberg, 1994, 1996). Previous work notes gender differences in how men and women use specific types of *discourse words* – for example, men use more filled pauses and repeats (Shriberg, 1996; Bortfeld et al., 2001) than women. However, these studies lack an automated method for large-scale discourse word discovery and gender analysis, primarily relying on the Switchboard corpus (Mitchell et al., 1999) – a corpus which is not representative of the range of natural speech patterns, as the phone calls were recorded in the manufactured, awkward situation of randomly-pairing two callers and assigning them a topic to discuss.

Hence, we propose in this paper a twofold framework for (i) the large-scale discovery and analysis of gendered discourse words in spoken content via our **Gendered Discourse Correlation Framework (GDCF, shown in Figure 1)**; and (ii) the measurement of the gender bias associated with these gendered discourse words in LLMs via our **Discourse Word-Embedding Association Test (D-WEAT, shown in Figure 2)**.

Concretely, we focus our study on podcasts, a popular and growing form of social media (Clifton

[1]We consider the binary definitions of sex (female/male) and gender (women/men, feminine/masculine) in our work due to (i) continuity with previous work in the gender debiasing task in the NLP community (Caliskan et al., 2017; Bolukbasi et al., 2016), and (ii) modeling constraints – i.e., *inaSpeechSegmenter* (Doukhan et al., 2018) for gender approximation via audio signal. This definition, however, is not representative of the sex and gender spectrums – and transgender, intersex, intersectional identities, and other identities are also not represented in this binary definition (Ghai et al., 2021; Ovalle et al., 2023; Seaborn et al., 2023). This is an important direction for future work.

Proceedings of the Third Workshop on Social Influence in Conversations (SICon 2025), pages 93–99

et al., 2020; The Pew Research Center, 2023). We analyze 15,117 podcast episodes from the Spotify Podcast Dataset (Clifton et al., 2020), to discover the *rewards* associated with *masculine discourse words* in terms of (i) correlated domains with substantial economic rewards, and (ii) more stable LLM representations. The presence of rewards for these *masculine discourse words* means that they indeed constitute *masculine defaults* (Cheryan and Markus, 2020).

Research Question 0: How are women and men's discourse different? We first introduce our *Gendered Discourse Correlation Framework (GDCF)* as shown in Figure 1, a framework for discovering gendered discourse words, with features which are centered around spoken content – specifically, an audio-based GENDER SEGMENTER (Doukhan et al., 2018), a TOPIC MODELER via LDA (Blei et al., 2003) and BERTopic (Grootendorst, 2022), and a specialized CONVERSATIONAL PARSER (Jamshid Lou and Johnson, 2020). We analyze correlations between *gender* and *discourse words* to automatically form gendered discourse word lists, as shown in Tables 1 and 2. Additionally, GDCF is a flexible framework which can be extended to other forms of audio speech data – such as short videos that are prevalent on TikTok, Instagram, and YouTube, long videos on YouTube, streamers on Twitch, and more.

Research Question 1: Are discourse-based masculine defaults present in domain-specific contexts? We then study the prevalence of these gendered discourse words in domain-specific contexts, as shown in Table 3. We find that masculine discourse words are positively correlated with the business domain, the technology/politics domain, and the video games domain. Participation in these domains grants economic *rewards* (Cheryan and Markus, 2020), hence there are indeed discourse-based masculine defaults present.

Research Question 2: Are discourse-based masculine defaults present in LLM embeddings? Finally, we study the representation of these gendered discourse words as shown in Figure 2, using a state-of-the-art LLM embeddings model from OpenAI, `text-embedding-3-large`. We find that the masculine discourse words have a more stable and robust representation than the feminine discourse words, as shown in Figures 3 and 4, resulting in better system performance on downstream tasks for men. Hence, men are *rewarded* (Cheryan and Markus, 2020) for their discourse patterns with better system performance by one of the state-of-the-art language models – and therefore this difference in the embedding representations for women and men constitutes a masculine default (Cheryan and Markus, 2020) and a *representational harm* (Blodgett et al., 2020).

We consider a few key types of implications:

(1) Theoretical Implications: First, the use of gendered discourse words can be considered a type of *gender performativity* (Butler, 1988, 2009; West and Zimmerman, 1987; Unger, 1979; Muehlenhard and Peterson, 2011), wherein the discourse words are part of a *gender schema* (Bem, 1984; West and Zimmerman, 1987). Hence, we identify specific words which are part of the current *hegemonic masculine* strategy (Connell, 1995, 1987) – and in the domain of technology, discourse words which are part of the *technomasculine* strategy (Cooper, 2000; Lockhart, 2015; Bulut, 2020). We contribute GDCF (Figure 1) for the discovery and analysis of gendered discourse words. Second, we contribute D-WEAT as an intrinsic metric which can be used to debias LLMs, broadening the debiasing task in natural language processing.

(2) Policy Implications: Policymakers – in government or platforms such as Spotify – could implement measures by which to mitigate bias in LLMs with respect to gender. Specifically, policymakers could regulate the use of D-WEAT to impose an unbiased representation of discourse words with respect to gender. Broadly, D-WEAT can join *a set of debiasing methods, tools, and datasets* (Bolukbasi et al., 2016; Caliskan et al., 2017; May et al., 2019; Nangia et al., 2020; Nadeem et al., 2020; Guo et al., 2022; He et al., 2022; Cheng et al., 2023; Dong et al., 2023) which can be employed to regulate bias in LLMs.

(3) Ethical Implications: A potential ethical concern is that tools used to remove bias can also be used to exacerbate bias. GDCF and D-WEAT could potentially be used to discover discourse words in audio-text corpora, and then *increase* the gender bias of the LLM embeddings. This abuse of the framework would be a *representational harm* (Blodgett et al., 2020). However, a more important point is that it is hard to undo bias issues without knowing how that bias manifests.

References

Simone de Beauvoir. 1949. *The Second Sex*.

S L Bem. 1984. Androgyny and gender schema theory: a conceptual and empirical integration. *Nebraska Symposium on Motivation. Nebraska Symposium on Motivation*, 32:179–226.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in nlp. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, page 5454–5476.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *arXiv*.

Heather Bortfeld, Silvia D Leon, Jonathan E Bloom, and 1 others. 2001. Disfluency rates in conversation: Effects of age, relationship, topic, role, and gender. *Language and Speech*, 44(2):123–147.

Ergin Bulut. 2020. *A Precarious Game*. Cornell University Press, Ithaca, NY.

Judith Butler. 1988. Performative acts and gender constitution an essay in phenomenology and feminist theory. *Theatre Journal*, 40(4):519.

Judith Butler. 2009. Performativity, precarity and sexual politics. *AIBR. Revista de Antropología Iberoamericana*, 4(3).

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Myra Cheng, Esin Durmus, and Dan Jurafsky. 2023. Marked personas: Using natural language prompts to measure stereotypes in language models. *arXiv*.

Sapna Cheryan and Hazel Rose Markus. 2020. Masculine defaults: Identifying and mitigating hidden cultural biases. *Psychological Review*, 127(6):1022–1052.

Ann Clifton, Sravana Reddy, Yongze Yu, and 1 others. 2020. 100,000 podcasts: A spoken english document corpus. In *COLING*, pages 5903–5917.

R.W. Connell. 1987. *Gender and power: society, the person, and sexual politics*. Stanford University Press.

R.W. Connell. 1995. *Masculinities*. Allen Unwin.

Marianne Cooper. 2000. Being the "go-to guy": Fatherhood, masculinity, and the organization of work in silicon valley. *Qualitative Sociology*, 23(4):379–405.

Xiangjue Dong, Ziwei Zhu, Zhuoer Wang, Maria Teleki, and James Caverlee. 2023. Co^{2}PT: Mitigating bias in pre-trained language models through counterfactual contrastive prompt tuning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5859–5871. Association for Computational Linguistics.

David Doukhan, Jean Carrive, Félicien Vallet, and 1 others. 2018. An open-source speaker gender detection framework for monitoring gender equality. In *ICASSP*. IEEE.

Bhavya Ghai, Md Naimul Hoque, and Klaus Mueller. 2021. Wordbias: An interactive visual tool for discovering intersectional biases encoded in word embeddings. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–7.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Yue Guo, Yi Yang, and Ahmed Abbasi. 2022. Autodebias: Debiasing masked language models with automated biased prompts. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, page 1012–1023.

Jacqueline He, Mengzhou Xia, Christiane Fellbaum, and Danqi Chen. 2022. Mabel: Attenuating gender bias using textual entailment data. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, page 9681–9702.

Paria Jamshid Lou and Mark Johnson. 2020. Improving disfluency detection by self-training a self-attentive model. In *ACL*.

Eleanor Amaranth Lockhart. 2015. *Nerd/Geek masculinity: Technocracy, Rationality, and gender in nerd culture's countermasculine hegemony*. Ph.D. thesis.

Chandler May, Alex Wang, Shikha Bordia, Samuel R Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. *arXiv*.

Merriam-Webster. 2024. Discourse.

Marcus Mitchell, Beatrice Santorini, M Marcinkiewicz, and 1 others. 1999. Treebank-3 ldc99t42 web download. *Linguistic Data Consortium*, 3:2.

Charlene L. Muehlenhard and Zoe D. Peterson. 2011. Distinguishing between sex and gender: History, current conceptualizations, and implications. *Sex Roles*, 64(11–12):791–803.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2020. Stereoset: Measuring stereotypical bias in pretrained language models. *arXiv*.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R Bowman. 2020. Crows-pairs: A challenge dataset for measuring social biases in masked language models. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, page 1953–1967.

Anaelia Ovalle, Palash Goyal, Jwala Dhamala, Zachary Jaggers, Kai-Wei Chang, Aram Galstyan, Richard Zemel, and Rahul Gupta. 2023. "i'm fully who i am": Towards centering transgender and non-binary voices to measure biases in open language generation. *2023 ACM Conference on Fairness, Accountability, and Transparency*, page 1246–1266.

Katie Seaborn, Shruti Chandra, and Thibault Fabre. 2023. Transcending the "male code": Implicit masculine biases in nlp contexts. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, page 1–19.

Elizabeth Shriberg. 1996. Disfluencies in switchboard. In *International Conference on Spoken Language Processing*.

Elizabeth Ellen Shriberg. 1994. *Preliminaries to a Theory of Speech Disfluencies*. Ph.D. thesis.

Christopher St Aubin The Pew Research Center. 2023. Audio and Podcasting Fact Sheet.

Rhoda K. Unger. 1979. Toward a redefinition of sex and gender. *American Psychologist*, 34(11):1085–1094.

Candace West and Don Zimmerman. 1987. Doing gender. *Gender and Society*, 1:125–151.

A Appendix

We provide supplementary figures and tables here.

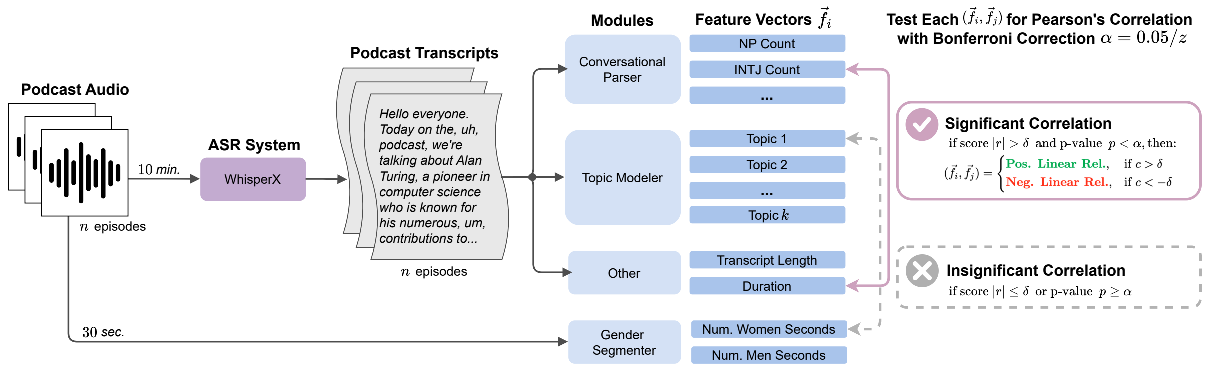

Figure 1: GDCF (Gendered Discourse Correlation Framework) Diagram: Testing for correlations with an example of a significant correlation and an insignificant correlation – all $(\vec{f}_i, \vec{f}_j)$ pairs are labeled *significant* or *insignificant*. $|\vec{f}_i| = 15,117$ podcast episodes. $z = \binom{124}{2} = 7,626$ correlation tests for the 124 total feature vectors.

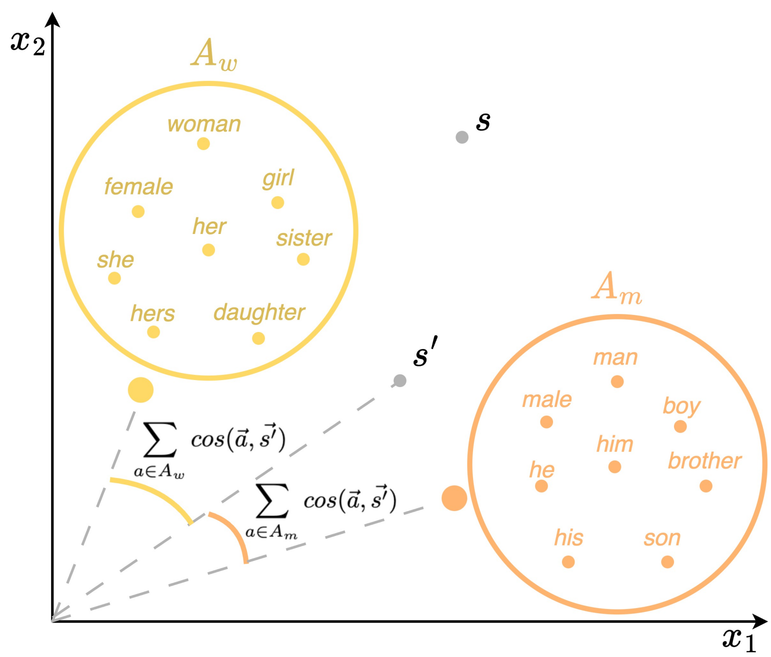

Figure 2: D-WEAT: Plot of the segment vectors $\vec{s}$ and $\vec{s'}$, and the word vectors, $\vec{w} \in A_w$, and $\vec{w} \in A_m$, projected into a two-dimensional space for illustrative purposes. The cosine similarity for s' and A_w, and s' and A_m is depicted; the cosine similarity for s and A_w, and s and A_m is calculated in the same way.

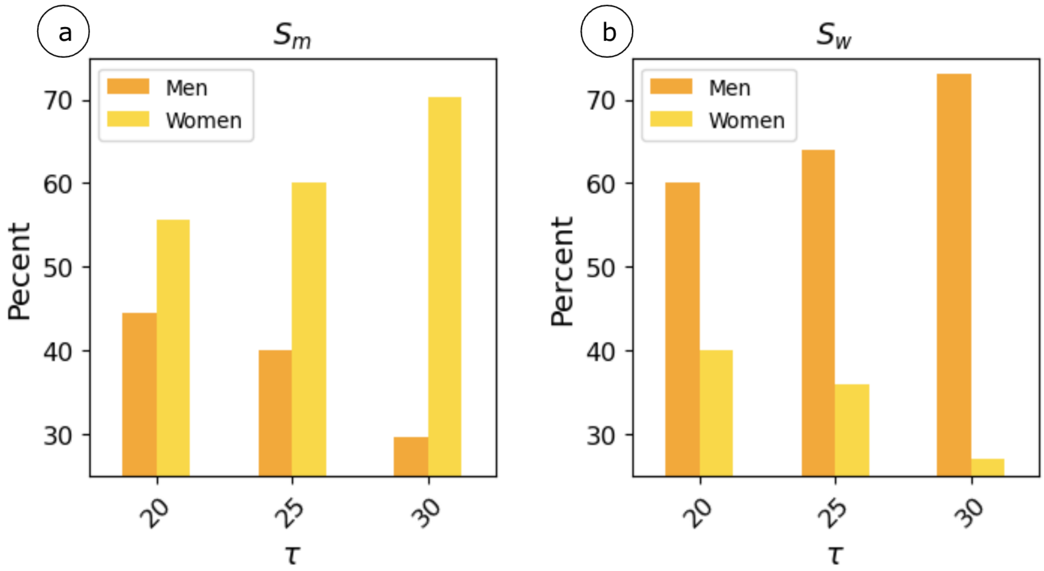

Figure 3: ⓐ Impact of τ on the average percentage of S_m segments which move closer to the *women* concept (A_w) versus the *men* (A_m) concept. ⓑ Impact of τ on the average percentage of S_w segments which move closer to the *women* concept (A_w) versus the *men* (A_m) concept.

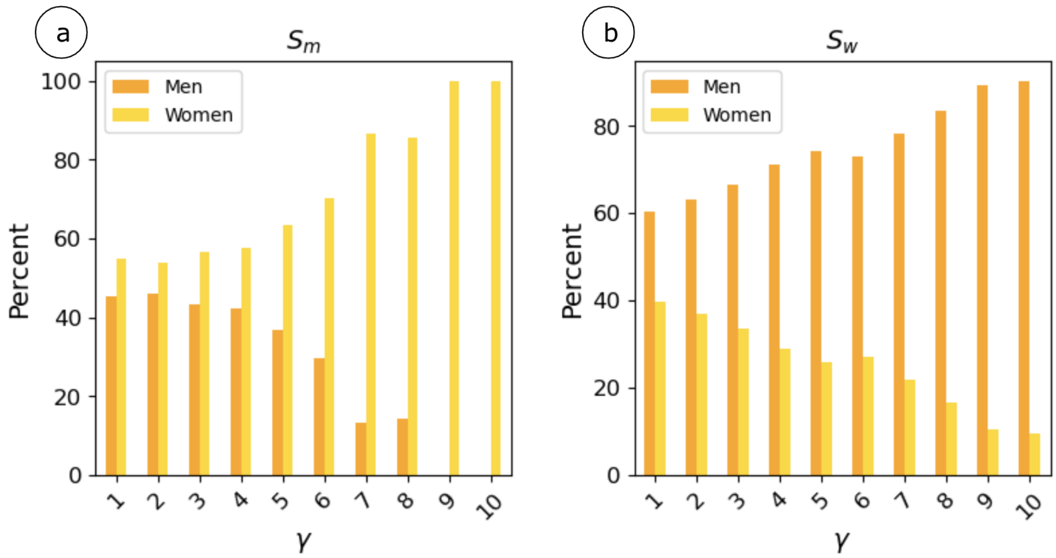

Figure 4: ⓐ Impact of γ on the average percentage of S_m segments which move closer to the women concept (A_w) versus the men (A_m) concept. ⓑ Impact of γ on the average percentage of S_w segments which move closer to the women concept (A_w) versus the men (A_m) concept.

Table 1: **LDA with Non-Contextual Embeddings (Bag-Of-Words):** The complete set of significant correlations between gender features and topic features – *both content topics and discourse topics*. Based on r, the Topic N Gender forms the **gendered discourse word lists** via Topics 54 and 60 (the masculine word lists) and Topic 62 (the feminine word list).

Topic N	Gender	r	Topic N Word List	Topic N Categories	Topic N Gender
Topic 3	Women	0.15	women, woman, men, baby, pregnant, girls, men, doctor, health, birth	Content - Pregnancy	Women
	Men	-0.14			
Topic 10	Women	0.10	energy, body, feel, mind, space, yoga, love, beautiful, feeling, meditation	Content - Yoga	Women
	Men	-0.12			
Topic 49	Women	-0.21	game, know, think, team, going, mean, play, year, one, good	Content - Sports	Men
	Men	0.17			
Topic 71	Women	0.14	christmas, sex, girl, hair, love, get, date, girls, let, wear	Content - Dating	Women
	Men	-0.14			
Topic 54	Women	–	get, like, know, right, people, going, podcast, make, want, one	Discourse	Men
	Men	0.12			
Topic 60	Women	-0.27	going, know, think, get, got, one, really, good, well, yeah	Discourse	Men
	Men	0.20			
Topic 62	Women	0.33	like, know, really, going, people, want, think, get, things, life	Discourse	Women
	Men	-0.28			

Table 2: **BERTopic with Contextual Embeddings (BERT, ChatGPT, Llama):** The complete set of significant correlations between gender features and topic features for *discourse topics only* (content topics are omitted).

Topic N	Gender	r	Topic N Word List	Topic N Categories	Topic N Gender
Topic 0	Women	-0.08	like, yeah, know, oh, right, podcast, got, going, think, really	Discourse	Men
	Men	0.10			
Topic 2	Women	0.08	life, know, things, really, people, feel, like, want, love, going	Discourse	Women
	Men	-0.08			
Topic 5	Women	0.08	like, know, think, yeah, episode, really, going, anchor, kind, right	Discourse	Women
	Men	–			

Table 3: LDA with Non-Contextual Embeddings (Bag-Of-Words): Significant correlations between content topic features and **gendered discourse word lists** (discourse topic features 54, 60, 62, see Table 1) for content topic features which *do not* have direct, significant correlations with gender features, but may broadly be more used by one gender.

Topic N	Topic M	r	Topic N Word List	Topic N Categories	Topic M Word List	Topic M Categories
Topic 11	Topic 54	0.11	data, new, technology, public, bill, theory, science, system, security, article	Content - Technology/ Political	get, like, know, right, people, going, podcast, make, want, one	Discourse (Men)
	Topic 62	-0.20			like, know, really, going, people, want, think, get, things, life	Discourse (Women)
Topic 12	Topic 54	0.24	business, money, company, market, buy, right, million, companies, pay, sell	Content - Business	get, like, know, right, people, going, podcast, make, want, one	Discourse (Men)
Topic 79	Topic 60	0.18	game, games, play, playing, like, played, nintendo, video, fun, switch	Content - Video Games	going, know, think, get, got, one, really, good, well, yeah	Discourse (Men)
	Topic 62	-0.13			like, know, really, going, people, want, think, get, things, life	Discourse (Women)