CARMA 2026

8th Int. Conf. on Advanced Research Methods and Analytics
1-3 July 2026, Valencia, Spain
DOI: https://doi.org/10.4995/CARMA2026.2026.21975

Modeling human and LLM-generated counter narratives on the hypersphere: a directional analysis of anti-Roma counter speech

Jai Jobe[1], Alex Cucco[1], Stefania Fensore[1], Annalina Sarra[1], Lara Fontanella[1], Marco Di Marzio[1]

[1]Department of Socio-Economic, Managerial, and Statistical Studies, G. d'Annunzio University Chieti-Pescara, Italy

How to cite: Jobe, J.; Cucco, A.; Fensore, S.; Sarra, A.; Fontanella, L.; Di Marzio, M. (2026). Modeling human and LLM-generated counter narratives on the hypersphere: a directional analysis of anti-Roma counter speech. In: 8th International Conference on Advanced Research Methods and Analytics (CARMA 2026). Valencia, 1-3 July 2026. https://doi.org/10.4995/CARMA2026.2026.21975

Abstract

This study investigates structural differences between human-authored and Large Language Model (LLM)-generated counter narratives addressing anti-Roma hate speech in Italian social media. We analyze a dataset of toxic comments and community-produced responses, comparing them with LLM-generated counter narratives. Texts are embedded in a high-dimensional semantic space, normalized onto the unit hypersphere, and modeled using finite mixtures of von Mises–Fisher distributions. Directional statistical inference reveals significant differences in semantic dispersion: human-generated counter narratives exhibit lower concentration parameters, indicating greater heterogeneity and lexical variability, whereas LLM outputs form more concentrated and homogeneous clusters. These findings suggest that, despite stylistic fluency, LLM-generated responses reproduce less of the intrinsic diversity characterizing community-based discourse. The results support the importance of community-in-the-loop approaches when designing computational tools for socially meaningful counter speech generation.

Keywords: *Hate Speech; Counter narratives; Roma people; Large Language Models; Directional Statistics; von Mises–Fisher Mixture*

Article in Press

1. Introduction

Social networks and online platforms have increasingly become sites of hateful, offensive, and abusive discourse, with hate speech proliferating on a scale that has attracted growing scholarly attention (Tontodimamma et al. 2021 for a comprehensive review) and prompted legislative and policy responses at both national and international levels. The phenomenon has fueled a significant debate between those who resist censorship and legal regulation of speech and those

This work is licensed under a Creative Commons License CC BY-NC-SA 4.0
Editorial Universitat Politècnica de València

who maintain that certain forms of expression exceed acceptable limits (Cohen-Almagor, 2016; Strossen, 2018). A widely supported approach to this challenge is to counter hate speech with more speech (Gagliardone et al., 2015) rather than restricting expression, counter speech opposes hateful content directly and has been shown to outperform other moderation strategies in reducing its harmful impact (Benesch, 2014; Schieb & Preuss, 2016). Grounded in factual argumentation and alternative viewpoints, counter speech responds to and challenges hate speech with the aim of discouraging its recurrence. Although the terms counter speech and counter narrative carry different connotations in social sciences, they have been treated interchangeably in NLP research (Bonaldi et al., 2024). Substantial NLP research has emerged around the automatic production of counter speech, spanning classification, selection, and generation tasks (see Bonaldi et al., 2024, and references therein), with LLMs proving particularly capable in both hate speech recognition and counter speech generation (Hengle et al., 2024; Saha & Srihari, 2023; Wadhwa et al., 2025). Yet a critical open question remains: to what extent do LLM-generated counter narratives reflect the linguistic and rhetorical qualities of human-produced ones? Comparative studies of human and machine-generated text point to consistent disparities in lexical richness, syntactic variation, and argumentative specificity (Muñoz-Ortiz et al., 2024). In the counter speech domain, Mun et al. (2023) demonstrate that LLMs tend to default to generic denunciations of hate speech rather than engaging with its underlying stereotypical implications — precisely the strategies that human evaluators find least convincing. Beyond linguistic quality, Kosonen (2025) argues that counter speech authored by members of targeted communities draws on experiential knowledge that no automated system can reproduce — a perspective that shapes the design of the present study, in which counter narratives addressing anti-Roma hate speech were written by Roma community members. To compare LLM and human written texts, our work adopts a geometric framework, where counter narratives are encoded as unit vectors on a hyperdimensional sphere. Much of the literature on high-dimensional text clustering has relied on discriminative approaches, including methods based on cosine similarity and spectral techniques (Steinbach et al., 2000; Zhao & Karypis, 2004). Among these, spherical k-means (Dhillon & Modha, 2001; Strehl et al., 2000) adapts classical k-means to unit-normalized document vectors and has shown strong empirical performance with the caveat of lacking an explicit probabilistic formulation. Generative models based on mixtures of von Mises–Fisher distributions provide a principled alternative for modeling directional data (Banerjee et al., 2005). This framework extends spherical k-means by allowing component-specific concentration parameters, thereby capturing heterogeneous cluster dispersion. Empirical evidence shows that such methods outperform traditional clustering approaches in several high-dimensional text mining settings (Zhong & Ghosh, 2003), and they have been successfully applied to large-scale text and anomaly detection tasks (Srivastava & Zane-Ulman, 2005). Alternatively, mixtures of Watson distributions have also been proposed (Bijral et al., 2007), although their increased modeling flexibility often comes at the cost of more complex parameter estimation in high dimensions (Banerjee et al., 2005). In

Article in Press

this paper, we make a systematic comparison of human- and LLM-generated counter narratives in Italian, examining lexical and syntactic features following the approach proposed by Banerjee et al., 2005. We identify clusters, investigate distributional dispersion, and conduct directional statistical inference to assess the extent and nature of human-machine divergence.

2. Material and methods

2.1. Data description

This study utilized a dataset of 450 manually selected offensive comments targeting Roma communities in Italian extracted from Facebook (n=15,129 total texts), curated to minimize content redundancy and maximize the applicability to counter narrative task. Three annotators from the Italian Romanès communities were recruited to generate counter rhetorical responses to each comment, with 150 randomly assigned comments per annotator. Subsequently, responses generated by two large language models—*Mixtral-8x7B-Instruct-v0.1* and *GPT-5.2*—were compared. Both models were instructed with a core prompt in Italian which translates in: "You are provided with a sentence from online social media. Please provide a counter rhetorical response to counter hate speech." The *Mixtral* model was deployed locally with individual prompts crafted for each comment, while *GPT-5.2* processed the entire dataset in a single batch. To evaluate counter narrative similarities, we *employed sentence-transformers/stsb-xlm-r-multilingual*, which encoded each counter narrative into a 768-dimensional vector. These vectors were subsequently normalized to the unit norm, representing each counter narrative as a point on a unit hypersphere.

2.2. Methodology

We aim to classify high-dimensional data with directional characteristics by investigating differences between human- and LLM-generated counter narratives through lexical and syntactic features. Each counter narrative is encoded as a unit vector on the hypersphere

$$S^{d-1} = \{ x \in \mathbb{R}^d : \|x\| = 1 \},$$

where $\|.\|$ denotes the L_2 norm. This representation naturally motivates the use of directional probability distributions.

Let $x \in S^{d-1}$ be a d-dimensional unit vector. The von Mises–Fisher (vMF) distribution is defined as

$$f(x \mid \mu, \kappa) = c_d(\kappa) \exp(\kappa \mu^T x), \quad (1)$$

where μ is a d-dimensional vector with $\|\mu\| = 1$ representing the mean direction, $\kappa \geq 0$ is the concentration parameter, and

Article in Press

$$c_d(k) = \frac{k^{d/2-1}}{(2\pi)^{d/2} I_{d/2-1}(k)}$$

is the normalizing constant involving the modified Bessel function of the first kind $I_\nu(\cdot)$ of order ν. The parameter κ controls the dispersion around μ in the sense that for $\kappa = 0$ the distribution reduces to the uniform distribution on S^{d-1}, while larger values of κ imply stronger concentration around μ. Given an i.i.d. random sample $X = \{x_1, \dots, x_n\}$ drawn from (1), to obtain the maximum likelihood estimates for the parameters of the vMF distribution, we consider the log-likelihood function

$$\ell(\mu, \kappa) = n \log c_d(\kappa) + \kappa \mu^T r \text{ , with } r = \sum_{i=1}^{n} x_i.$$

Maximization under the constraint $\|\mu\| = 1$ yields

$$\hat{\mu} = \frac{r}{\|r\|},$$

while the concentration parameter estimate $\hat{\kappa} \geq 0$ satisfies the following equation

$$\frac{I_{d/2}(\hat{\kappa})}{I_{d/2-1}(\hat{\kappa})} = \bar{r}, \text{ with } \bar{r} = \frac{\|r\|}{n}.$$

which has not closed-form solution, therefore numerical approximations are typically employed. To model multiple latent discourse regimes, following Banerjee et al. (2005) we consider a mixture of k vMF distributions:

$$f(x \mid \Theta) = \sum_{h=1}^{k} \alpha_h f_h(x \mid \mu_h, \kappa_h), \quad (1)$$

where $\alpha_h \geq 0, \Sigma_{h=1}^{k} \alpha_h = 1$, and each component $f_h(\cdot)$ is a vMF density with parameters (μ_h, κ_h). The full parameter set is

$$\Theta = \{\alpha_h, \mu_h, \kappa_h\}_{h=1}^{k}.$$

Under this generative model, each observation is generated by first sampling a latent component index h with probability α_h, and then sampling from the corresponding vMF distribution. The parameter set Θ can be estimated by using the Expectation–Maximization (EM) algorithm, as follows.

E-step. Compute posterior probabilities:

$$p(h \mid x_i, \Theta) = \frac{\alpha_h f_h(x_i|\mu_h, \kappa_h)}{\sum_{l=1}^{k} \alpha_l f_l(x_i|\mu_l, \kappa_l)}. \quad (2)$$

M-step. Update the mixture weights and the mean directions:

Article in Press

$$\hat{\alpha}_h = \frac{1}{n}\sum_{i=1}^{n} p(h \mid x_i, \Theta), \tag{3}$$

$$r_h = \sum_{i=1}^{n} x_i p(h \mid x_i, \Theta), \text{ and } \hat{\mu}_h = \frac{r_h}{\|r_h\|}. \tag{4}$$

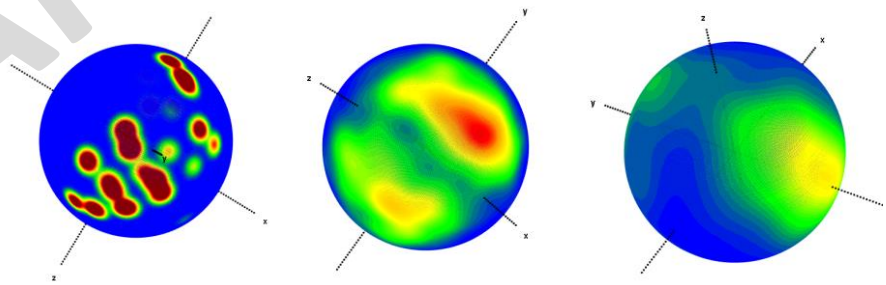
The concentration parameter is approximated by

$$\hat{\kappa}_h = \frac{d\bar{r}_h - \bar{r}_h^3}{1-\bar{r}_h^2}, \text{ with } \bar{r}_h = \frac{\|r_h\|}{n\alpha_h} \tag{5}$$

which provides an accurate and computationally efficient high-dimensional approximation. After a number of iterations alternating between the E-step (3) and the M-step updates (4)–(6), the procedure converges to a local maximum of the incomplete-data log-likelihood and produces the estimated mixture parameters $\hat{\Theta}$, posterior probabilities $p(h \mid x_i, \hat{\Theta})$, and a probabilistic clustering of counter narratives. Differences between human- and LLM-generated counter narratives are then assessed via the component concentration (κ_h), directional separation $(\mu_h^T \mu_h')$ and posterior membership shifts.

3. Results

Directional embeddings were first assessed for spherical uniformity. When pooling all observations, the Rayleigh test (Rayleigh, 1919) rejected the null hypothesis of uniformity on the unit hypersphere $(p = 0.001)$, indicating a statistically significant directional concentration. Figure 1 presents kernel density estimates obtained by projecting the high-dimensional embeddings onto the ordinary unit sphere using the Principal Nested Spheres technique through the fast approach described in Monem et al., 2025. The algorithm applies Principal Component Analysis (PCA) to construct an approximating lower-dimensional hypersphere to initialize the Principal Nested Spheres technique making the approach computationally efficient. We leverage on 20 PCs preserving around 76% of the total variability. The projection provides an exploratory visualization of how the three corpora occupy the semantic space.



Article in Press

Figure 1: From left: Kernel density estimate of ChatGPT, Mixtra, and Human data points projected to the sphere in 3D after using the Principal Nested Spheres.

Density estimates were computed with a von Mises–Fisher kernel using smoothing parameters $\hat{\kappa}$ =240, 65, and 29 for ChatGPT, Mixtra, and human-authored texts, respectively. In directional kernel density estimation, the kernel function has the role of tuning the contribution of the observations in the neighborhood of the estimation point, and $\hat{\kappa}$ regulates the concentration of the kernel: higher values produce sharper, more localized peaks, whereas lower values yield smoother, more diffuse surfaces. The estimates obtained reveal marked differences in semantic concentration across the corpora. Human-authored counter narratives span a broader portion of the sphere, reflecting greater semantic variability and a more heterogeneous set of linguistic strategies. By contrast, LLM-generated texts are more tightly clustered, indicating a pronounced tendency toward semantic centralization (see Figure 1). To further elucidate the latent structure of the semantic space, we fitted a mixture of von Mises--Fisher distributions, according to Equation 2 with $k = 3$, to the pooled dataset in the original 768-dimensional space. Each resulting component can be interpreted as a latent region that organizes the directional distribution underlying counter speech discourse. Table 1 presents the row-normalized composition of each mixture component according to text source (ChatGPT, Mixtra, Human). A pronounced correspondence emerges between mixture components and their predominant text origins. Component 3 is largely dominated by human-authored counter narratives (81.7%), suggesting that this region of the semantic space predominantly reflects community-generated discourse patterns. Component 2 is primarily associated with Mixtra outputs (52.0%), while ChatGPT-generated texts are mostly concentrated in component 1 (51.2%). The marked predominance of human texts within component 3 indicates that this component embodies a semantic regime characterized by directions typical of community-produced counter narratives. In the context of the von Mises--Fisher mixture model, this pattern implies that a substantial fraction of human-authored texts aligns closely with the mean direction of this component, thereby delineating a latent semantic region that is comparatively less occupied by language model outputs.

Table 1. Row-normalized composition of mixture components in terms of text source (ChatGPT, Mixtra, Human) under the EM-estimated von Mises–Fisher mixture model.

Group	ChatGPT	Mixtra	Human
1	0.512	0.326	0.162
2	0.161	0.520	0.319
3	0.051	0.132	0.817

Article in Press

4. Conclusion

This study compared human and LLM-generated counter narratives to anti-Roma hate speech using directional statistics on text embeddings. Human texts showed lower concentration parameters, indicating greater lexical diversity and varied rhetorical strategies, while LLM texts (especially ChatGPT 5.2) formed tighter clusters with more semantic uniformity. Certain semantic regions were predominantly occupied by human discourse, suggesting LLMs only partially capture community-based counter speech patterns. Directional modeling via von Mises–Fisher distributions provided an interpretable framework for analyzing high-dimensional embeddings, with concentration parameters quantifying discourse heterogeneity. The findings indicate that LLM-generated counter speech, while fluent, lacks the structural diversity of community discourse rooted in lived experience and cultural specificity, suggesting automated systems should complement rather than replace community voices through community-in-the-loop frameworks. Finally, we acknowledge that a sensitivity analysis across different embedding models and dimensionalities remains an important open question. Both a systematic sensitivity analysis across embedding models and the standardization of the generation pipeline across models represent our next effort. More broadly, by situating counterspeech within a geometric and probabilistic framework, this study suggests that the divergence between human and machine-generated discourse is not merely stylistic, but structurally embedded in the geometry of semantic space, motivating further investigation into how embedding model choice and generation settings interact with the observed distributional patterns.

References

Banerjee, A., Dhillon, I. S., Ghosh, J., Sra, S., & Ridgeway, G. (2005). Clustering on the unit hypersphere using von Mises-Fisher distributions. *Journal of Machine Learning Research*, *6*(9).

Benesch, S. (2014). Countering dangerous speech: New ideas for genocide prevention (Working paper). United States Holocaust Memorial Museum. Washington, DC. https://www.ushmm.org/m/pdfs/20140212-benesch-countering-dangerous-speech.pdf

Bijral, A. S., Breitenbach, M., & Grudic, G. (2007). Mixture of Watson distributions: A generative model for hyperspherical embeddings. *Artificial Intelligence and Statistics*, 35-42.

Bonaldi, H., Chung, Y.-L., Abercrombie, G., & Guerini, M. (2024). NLP for counterspeech against hate: A survey and how-to guide. *Findings of the Association for Computational Linguistics: NAACL 2024*, 3480-3499.

Cohen-Almagor, R. (2016). Hate and racist speech in the United States: A critique. *Philosophy and Public Issues*, *6*(1), 77-123. https://ssrn.com/abstract=2880753

Dhillon, I. S., & Modha, D. S. (2001). Concept decompositions for large sparse text data using clustering. *Machine Learning*, *42*(1), 143-175.

Article in Press

Gagliardone, I., Gal, D., Alves, T., & Martinez, G. (2015). *Countering Online Hate Speech*. UNESCO Publishing. https://unesdoc.unesco.org/ark:/48223/pf0000233231

Hengle, A., Padhi, A., Singh, S., Bandhakavi, A., Akhtar, M. S., & Chakraborty, T. (2024). Intent-conditioned and non-toxic counterspeech generation using multi-task instruction tuning with RLAIF. In K. Duh, H. Gomez, & S. Bethard (Eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*(pp. 6716-6733). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.naacl-long.374

Kosonen, H. S. (2025). The counterspeech doctrine in communicative capitalism: An intersectional feminist critique. *Feminist Media Studies*.

Monem, M., Dryden, I. L., & George, F. (2025). Principal nested spheres for high-dimensional data. *arXiv preprint arXiv:2511.08398*.

Mun, J., Allaway, E., Yerukola, A., Vianna, L., Leslie, S.-J., & Sap, M. (2023). Beyond denouncing hate: Strategies for countering implied biases and stereotypes in language. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 9759-9777). Association for Computational Linguistics.

Muñoz-Ortiz, A., Gómez-Rodríguez, C., & Vilares, D. (2024). Contrasting linguistic patterns in human and LLM-generated news text. *Artificial Intelligence Review*, *57*, 265.

Rayleigh, L. (1919). Xxxi. On the problem of random vibrations, and of random flights in one, two, or three dimensions. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, *37*(220), 321-347.

Saha, S., & Srihari, R. (2023). Consolidating strategies for countering hate speech using persuasive dialogues. In J. D. Pawar & S. Lalitha Devi (Eds.), *Proceedings of the 20th International Conference on Natural Language Processing (ICON)* (pp. 378-392). NLP Association of India (NLPAI). https://aclanthology.org/2023.icon-1.30/

Schieb, C., & Preuss, M. (2016). Governing hate speech by means of counterspeech on Facebook. *66th ICA Annual Conference*, 1-23.

Srivastava, A., & Zane-Ulman, B. (2005). Discovering hidden anomalies in text reports regarding complex space systems. *IEEE Aerospace Conference*, Big Sky, MT.

Steinbach, M., Karypis, G., & Kumar, V. (2000). A comparison of document clustering techniques.

Strehl, A., Ghosh, J., & Mooney, R. (2000). Impact of similarity measures on web-page clustering. *Workshop on Artificial Intelligence for Web Search (AAAI 2000)*, *58*, 64.

Strossen, N. (2018). *Hate: Why We Should Resist It with Free Speech, Not Censorship: Inalienable Rights*. Oxford University Press.

Tontodimamma, A., Nissi, E., Sarra, A., & Fontanella, L. (2021). Thirty years of research into hate speech: Topics of interest and their evolution. *Scientometrics*, *126*(1), 157-179.

Wadhwa, S., Xu, C., Chen, H., Mahalingam, A., Kar, A., & Chaudhary, D. (2025). Northeastern Uni at multilingual counterspeech generation: Enhancing counter speech generation with LLM alignment through direct preference optimization. In M. Guerini & R. Agerri (Eds.), *Proceedings of the First Workshop on Multilingual Counterspeech Generation*. Association for Computational Linguistics. https://aclanthology.org/2025.mcg-1.3/

Article in Press

Zhao, Y., & Karypis, G. (2004). Empirical and theoretical comparisons of selected criterion functions for document clustering. *Machine Learning*, *55*(3), 311-331.

Zhong, S., & Ghosh, J. (2003). A comparative study of generative models for document clustering. *Proceedings of the Workshop on Clustering High Dimensional Data and Its Applications in SIAM Data Mining Conference*.

Article in Press