

Cross-Ideological Hallmarks of Extremist Rhetoric in Social Media

Hasan Rehman Omarzai and Björn Gambäck

Department of Computer Science
Norwegian University of Science and Technology
NO-7491 Trondheim, Norway
hassano9@live.nl, gamback@ntnu.no

Abstract

The paper investigates to what extent linguistic features, primarily emotional and affective signals, serve as a common denominator in the online texts of various extremist groups. Examining social media data from multiple sources, this research focuses on ideologies such as far-right, jihadist, and various manosphere subgroups. Key results demonstrate distinct emotional profiles across ideologies, with “disgust” emerging as a dominant emotion in many extremist discourses, while “joy” is most prominent in the analyzed religious extremist content. Statistically significant dissimilarities in emotional expression patterns were confirmed, allowing for differentiation between ideological groups. Classifiers, particularly the DeBERTa model, demonstrated a strong capability to reliably distinguish between extremist groups based solely on linguistic characteristics.

1 Introduction

The social media space has evolved into a powerful vector for radicalization. A critical aspect of this radicalization pathway is the interplay between and process of misunderstanding, fear, and hatred, which can culminate in violence (Beck and Deffenbacher, 1999; Abuelaish and Arya, 2017). Understanding the emotional and affective dimensions of this process, particularly how language is weaponized to intensify fear, anger, and other emotions, provides the central motivation of this work.

Research indicates that individuals with differing ideological views exhibit variations in personality traits (Sibley et al., 2012), social motivations (Jost et al., 2003), cognitive functioning (Onraet et al., 2015), moral beliefs (Graham et al., 2009), and values (Caprara et al., 2017). These factors can influence an individual toward aggressive behavior, suggesting that different ideological frameworks

may vary in their likelihood to endorse or employ violence. The emotional components of extremist discourse, how fear, anger, pride, or other affective states are invoked and manipulated, play a crucial role in the radicalization process. In particular, an examination of these emotional patterns across a range of extremist groups will consider how this knowledge might offer new perspectives on the persuasive strategies and recruitment practices of such groups through linguistic analysis.

Understanding the online communication of extremist groups is a crucial concern for both theoretical and practical purposes. Upon identifying the linguistic features that characterize the rhetoric of each group, including, in particular, emotional language and effects that emerge in their content, we can better understand how these groups frame their content to convince, radicalize, and mobilize their audiences. To achieve this overarching goal, the following specific research questions need answers:

1. What are the dominant emotional and affective cues present in the extremist and radical discourse of these groups?
2. To what extent do these discourses exhibit similarities and dissimilarities in the way emotions are expressed and evoked?
3. Can state-of-the-art classification models reliably differentiate between extremist groups’ discourse based solely on linguistic features?

The rest of the paper is organized as follows: Section 2 examines previous work in extremism detection, while Section 3 describes the data collections supporting the research. Sections 4 and 5 then respectively details system architecture designed to answer the research questions and the experiments carried out. Section 6 interprets the results and discusses theoretical and practical implications, before Section 7 summarizes the key findings and suggests directions for future work. A final section at the end of the paper elaborates on possible limitations of the study.

2 Related Work

This section aims to build a comprehensive understanding of the current state-of-the-art concerning emotional and affective patterns in extremist discourse, the methods used for their comparison, the computational techniques employed for analysis, and the nature of the datasets utilized. The successes and limitations of existing lexicon-based, ML, and hybrid methods, particularly the advancements in transformer models noted by [Rajendran et al. \(2022\)](#) and [Jamil et al. \(2022\)](#), provide a benchmark and highlight the potential for further improvement in differentiating extremist discourse. Their findings are based on linguistic features, which this work also aims to explore.

Many studies have found that anger, fear, and moral outrage are primary affective markers of extremist language across different ideologies. [Buckingham and Alali \(2020\)](#) undertook a corpus-driven analysis of ISIS propaganda and far-right manifestos, observing that both employed affect-laden language, primarily negative affects such as anger and fear, but in different contexts. This contextual difference is significant; e.g., [Buckingham and Alali's](#) work suggests that while the core negative emotions are shared, the framing, religious justification for ISIS vs. cultural preservation for the far-right, dictates how these emotions are manifested and leveraged. Similarly, [Kaati et al. \(2016\)](#) conducted a LIWC ([Pennebaker et al., 2001](#)) analysis of lone offender manifestos and identified unique emotionality profiles with anger and hostility as key features, highlighting that even within such manifestos the specific blend and intensity of these core emotions can vary, potentially reflecting individual psychological states or particular grievances. [Törnberg and Törnberg \(2025\)](#) delivered important findings about digital media's emotional mobilization of far-right movements. Their study demonstrates how online media creates opportunities for generating and escalating emotional intensity, even in settings where people remain anonymous. They argue that emotional processes are crucial to the formation of far-right online groups and their ability to mobilize, with specific emotional profiles, such as moral indignation, a victim mentality, and group pride, serving as key drivers of radicalization.

Studies have also examined how extremist discourse employs moral foundations and values to evoke emotional responses. [del Real and Araque \(2023\)](#) applied moral values and emotions in their

radical language detection system, which revealed that including moral foundations theory and emotion lexicons improved detection accuracy. They demonstrate that extremist groups often present issues in moral terms to foster greater emotional investment and defend radical stances. [Baele et al. \(2023\)](#) introduced an integrated model for the language of violent extremists, demonstrating how semantic patterns in ISIS and National Socialist Movement (NSM) propaganda lead to social identity polarization and increased perceived threats. According to their study, the language patterns between these groups show comparable features through identity-polarizing language along with hostile rhetoric and violent justifications.

Transformers and deep learning models demonstrate successful results when used to identify emotional content in extremist discourse, e.g., [Rajendran et al. \(2022\)](#) used Bi-LSTM ([Hochreiter and Schmidhuber, 1997](#)) and BERT ([Devlin et al., 2019](#)), RoBERTa ([Liu et al., 2019](#)), and DistilBERT ([Sanh et al., 2019](#)) models to detect extremism in the U.S. Capitol riot with transformer models producing superior results.

Despite significant progress made, several methodological issues still hinder the analysis of extremist discourse. Most current approaches lack a theoretical basis for emotion and radicalization, except [del Real and Araque \(2023\)](#), who applied Moral Foundations Theory. Since computational findings lack a clear link to psychological theories of radicalization, the results have limited interpretability, underscoring the importance of developing hybrid methods that integrate computational linguistics with psychological studies of radicalization.

Current studies also lack adequate longitudinal analysis of extremist discourse because scientists rarely study the developmental trajectories of emotional patterns over time or when faced with external events. Our ability to understand radicalization as a process faces limitations because current models fail to monitor emotional changes within extremist groups during extended periods, indicating the need for dynamic models. Furthermore, studies of emotional contagion and amplification lack sufficient depth because there have been few systematic analyses of how extremist groups transmit and strengthen emotions through their online networks. Current studies also lack a sufficient analysis of the emotional aspects used in counter-extremism messaging and their effectiveness in countering extremist appeals.

3 Data Collections

This research heavily relies on two extensive datasets curated and maintained by the Cambridge Cybercrime Centre (CCC) and its partner laboratories: **CrimeBB** is a collection of data scraped from various online cybercrime forums. It contains approximately 48 million posts across 32 different forums (Pastrana et al., 2018). **ExtremeBB** is a parallel collection focusing specifically on extremist forums. These collections offer a rich, longitudinal view into the evolution of online extremist narratives and community dynamics.

In addition, data from public repositories provided curated datasets cleaned by the wider research community. Public datasets offer valuable snapshots of specific phenomena, and add comparability with prior studies. This dual-channel approach was deliberately chosen to balance the depth and longitudinal nature of large-scale institutional datasets, crucial for understanding nuanced, evolving discourse, with the accessibility and comparability offered by public repositories, facilitating benchmarking against existing research.

A key component is the publicly available **Twitter/X** dataset “How ISIS Uses Twitter” (Fifth Tribe, 2020). This collection, spanning the years 2014–2016, contains 17,000 tweets and retweets. It is particularly valuable as it is a highly cited collection of jihadist propaganda posts. Each record in this dataset provides the tweet text, an anonymized user identifier, and various engagement metrics.

Telegram has emerged as a significant fallback sanctuary for extremist groups following deplatforming actions on mainstream social media (Rogers, 2020). CrimeBB’s data collected from Telegram preserves rich contextual information, including channel descriptions, subscriber counts, post timestamps, references to attachments, and links indicating message replies. Six channels are used: Proud Boys, Patriot Front, PragerU, Benny Johnson and Ben Shapiro gang (far-right) and Kiwi Farms (alt-right).

Discord, initially a platform for gaming communities, has also become a haven for various extremist factions (Oakford et al., 2023), particularly those associated with the manosphere and white supremacist ideologies. CrimeBB retains significant server-level context, allowing for in-depth studies of sub-community dynamics, the impact of moderation, and the specific linguistic and thematic characteristics of these extremist groups.

Online forums and bulletin boards also serve as significant hubs for extremist ideologies and communities. ExtremeBB data from these sources provide another layer of understanding regarding the breadth and nature of online radicalization: Data spanning over two decades from established **far-right forums**, such as Stormfront, White Nations, and the Vanguard News Network, form a big part of the data used for this research by enabling the study of how emotional expressions, ideological positions and rhetoric within these communities shift and evolve. Data from **manosphere communities**, including forums associated with **Incels**, **pickup artists (PUA)**, and **Men Going Their Own Way (MGTOW)** facilitate cross-ideological comparisons. Manosphere communities were selected due to their increasing prominence and documented links to misogynistic and sometimes extremist ideologies (Ging, 2019; Van Eerten et al., 2018).

Ideology labels are assigned at source level: far-right comprises Stormfront, White Nations, Vanguard News Network, the five telegram channels and four small Discord servers; alt-right Kiwi Farms, Lolcow and Kiwi Farms Chat; incel Incels.is, Incels.net, Looksmax, Lookism and one Discord server; MGTOW and PUA two forums each; Jihadist the Twitter corpus; religious extremist the Traditional Catholicism Discord server; and accelerationist the Nazi Weeb Gamers Discord server. The last three are thus single source labels (see Table 1 for details).

The raw data acquired from the various sources arrived in diverse formats, so was first ingested into a PostgreSQL schema (PostgreSQL Global Development Group, 2025). For intermediate storage and efficient data interchange with pandas DataFrames (Reback et al., 2020) during the analysis phase, Apache Feather (Apache, 2025) was selected for its high read/write performance with tabular data. Each record in the final Feather files includes essential metadata (source, platform, creator ID, timestamp, etc.) and both raw and cleaned text, alongside the engineered features.

The depth of the associated metadata adds significant value to the collected data, e.g., timestamps are essential for examining temporal trends and changes in language or topics over time. We utilize author IDs and reply links to conduct network analysis, influence mapping, and identify community structure, role labels, and server/channel information, which enables us to understand the structure and dynamics behind online extremism groups.

The online extremist datasets this research relies on are inevitably incomplete and biased, as platform demographics, self-selection, moderation, temporal gaps, and license constraints distort the data collected, while private or deleted content remains invisible. To reduce these distortions, results are triangulated across multiple platforms, ideologies, and periods. Still, each corpus carries its caveats, the “How ISIS Uses Twitter” dataset reflects tactics from 2014 to 2016 and may miss current behaviors. To mitigate this, this work incorporates more recent data from Telegram and Discord (up to 2024) and contemporary forum data from CrimeBB and ExtremeBB. By analyzing these diverse and temporally varied datasets, the aim is to capture a broader and more current spectrum of extremist communication patterns, allowing for a more nuanced understanding than a single, older dataset could provide.

The study operates within a stringent ethical and legal framework that balances the utility of research with the need for privacy, safety, and regulatory compliance. Personally identifiable details are anonymized or paraphrased in line with the GDPR and Cambridge Cybercrime Centre (CCC) requirements. Extremist excerpts are contextualized and sanitized to minimize harm to researchers and readers. While analytical methods and aggregated results may be published, raw data (especially CCC-provided corpora) remains inaccessible to the public and even peer reviewers without formal CCC approval and a non-disclosure agreement. To maintain transparency, citations, and links are provided for all openly licensed datasets, including the Kaggle “How ISIS Uses Twitter.”

4 Architecture

All raw messages are consolidated from four sources: ExtremeBB forum dumps, Telegram channels, Discord servers part of CrimeBB, and a pro-ISIS Twitter corpus. URLs, forum ‘banner tags,’ HTML markup, and superfluous whitespace were removed as they represent platform-specific noise or non-linguistic artifacts that could interfere with tokenization and downstream analysis. Emojis were ‘de-emojised’ (e.g., ‘😞’ to ‘:sad_face:’) using the Python emoji library (Kim et al., 2025) to retain their semantic content in textual form, as emojis can convey significant emotional information. The texts were tokenized and lemmatized, and standard English stop-words removed.

The specific pronouns chosen for inclusive (‘we,’ ‘our,’ ‘us,’ ‘together’) and exclusive (‘they,’ ‘them,’ ‘those,’ ‘enemy’) language were selected based on previous research indicating their role in constructing in-group identity and out-group derogation (see, e.g., Rasoulkolamaki et al., 2025; El-Zouka, 2019). These features provide a simple yet potentially indicative measure of group orientation.

Each message was vectorized to determine its emotional content. *pysentimiento* (Pérez et al., 2021) was selected as the primary emotion analysis tool due to its state-of-the-art transformer-based architecture, which allows for more context-aware emotion detection compared to traditional lexicon-based methods. NRCLEX (Mohammad and Turney, 2013) was designated as a robust fallback and used for initial explorations due to its established lexicon and a broader range of emotion categories. However, its inability to disambiguate word senses in context is a known limitation.

Pérez et al. (2021) evaluated *pysentimiento* on four different tasks — sentiment analysis, emotion detection, hate speech detection, and irony detection — showing the viability of the resource. Hence, for the final analyses focusing on core emotions, *pysentimiento* was prioritized. Raw emotion likelihoods or counts were re-normalized so that emotion scores for each message sum to 1, ensuring comparability across messages of different lengths and facilitating subsequent statistical analyses.

The analytical core of this research is structured around three distinct experiment pipelines, corresponding to the three research questions posed in Section 1.

The primary goal of the first pipeline is to quantitatively assess and compare the emotional content expressed within texts produced by different ideological groups. The pipeline begins by loading the preprocessed data, specifically focusing on the probability vectors for six core emotions: fear, anger, surprise, sadness, joy, and disgust. Non-parametric bootstrap confidence intervals (10,000 samples) were employed to quantify the statistical uncertainty around the mean emotion scores for each ideology. This method was chosen as it does not rely on assumptions about the underlying distribution of emotion scores, making it robust for potentially non-normally distributed linguistic data. A difference was deemed noteworthy when 95% CIs did not overlap, a common heuristic for initial comparison. Visualizations consist of faceted bar charts plus an optional yearly trend line plot.

The second pipeline quantifies how closely different ideologies resemble one another in the emotion-probability space. The analysis utilizes the ideology-level mean emotion vectors ($\mu_c \in \mathbb{R}^{10}$), which have already been normalized to sum to 1 during preprocessing. Pairwise cosine *distances* $d_{ij} = 1 - \frac{\mu_i \cdot \mu_j}{\|\mu_i\|_2 \|\mu_j\|_2}$ are computed to form a distance matrix $D \in \mathbb{R}^{C \times C}$, where C is the number of ideologies. Ward’s linkage is applied to D , and the resulting hierarchical tree is rendered as a dendrogram and a heat-map PNG.

To visualize the global structure, Principal Component Analysis (PCA) was performed, retaining two components for visualization purposes. This choice facilitates an intuitive 2D representation of the primary dimensions of variation in the multi-dimensional emotion space, allowing for a global overview of ideological clustering.

A permutation MANOVA using Pingouin (Valat, 2018) based on Pillai’s trace was selected to assess overall differences in emotion profiles. This non-parametric approach was chosen for its robustness to violations of multivariate normality and homogeneity of covariance matrices, which are common with real-world textual data. If the global test is significant, the pipeline runs one-way ANOVAs for each of the six emotion dimensions, followed by Tukey–HSD pairwise comparisons. p -values are corrected for multiple testing with the Benjamini–Hochberg procedure.

The third pipeline evaluates how accurately text snippets can be assigned to the ideology from which they originate. The text is tokenized with HuggingFace tokenizers and sequences are padded and hard-truncated to 256-word pieces.

Two pre-trained encoders are compared: distilbert-base-uncased (Sanh et al., 2019) was chosen for its established balance of strong performance and computational efficiency, representing a lighter yet effective transformer. microsoft/deberta-base (He et al., 2020) was selected because it represents a more powerful architecture with improvements such as disentangled attention, which is expected to offer potentially higher accuracy on this nuanced classification task.

The entire encoder stack of both models is fine-tuned, no layers frozen, and no layer-wise learning-rate decay used. The contextual [CLS] embedding is passed through a two-layer projection MLP with dropout, producing logits over the ideology set. Optimization uses AdamW (Loshchilov and Hut-

Ideology	5k	10k	40k	80k
far-right	44 815	83 924	194 721	314 721
alt-right	15 000	30 000	120 000	240 000
accelerationist	1 025	1 025	1 025	1 025
jihadist	17 410	17 410	17 410	17 410
incel	25 000	46 629	166 629	326 629
MGTOW	10 000	20 000	80 000	160 000
PUA	10 000	20 000	80 000	160 000
religious	5 000	10 000	40 000	66 413
Total	128 250	228 988	699 785	1 286 198

Table 1: Messages per ideology in the four dataset variants (cap of 5k, 10k, 40k and 80k messages per forum, channel or server; Twitter corpus used in full)

ter, 2019) with a default learning rate (2×10^{-5}), $\beta = (0.9, 0.98)$, and $\varepsilon = 10^{-8}$. A linear schedule warms the learning rate over the first 10 % of update steps before decaying to zero. Each model runs for the user-specified number of epochs (default 5). The epoch with the highest macro-F1 on the evaluation fold is retained; as no separate validation split was used, scores are slightly optimistic.

A linear-kernel SVM (Cortes and Vapnik, 1995) with TF-IDF features (10,000 uni- and bigrams) was included as a strong lexical baseline, representing the traditional efficacy of text classification on sparse, high-dimensional data. A Random Forest ensemble (Breiman, 2001) of the eight dense engineered features: six emotion probabilities and inclusive/exclusive pronoun ratios, added a baseline to evaluate the predictive power of these explicitly constructed linguistic and affective features alone.

5 Experiments and Results

To systematically investigate the influence of data volume on the stability of emotional findings and on the performance of classification models, four distinct dataset variants were created by capping the number of messages per forum, channel and server at 5, 000, 10, 000, 40, 000 and 80, 000 (the Twitter corpus is used in full). This approach was chosen to assess model scalability and determine whether observed linguistic patterns and classification accuracies hold robustly under different data availability conditions, simulating scenarios ranging from limited to extensive data collection.

Table 1 summarizes the approximate total number of messages and illustrative class balance for two key ideologies. The actual number of messages per ideology varies based on data availability per source across years and MIN_SIZE filtering.

Ideology	Emotion	Mean	95% CI
alt-right	disgust	0.445	[0.443, 0.447]
far-right	disgust	0.481	[0.480, 0.483]
accelerationist	disgust	0.172	[0.158, 0.188]
incel	disgust	0.406	[0.405, 0.408]
MGTOW	disgust	0.574	[0.573, 0.576]
PUA	disgust	0.436	[0.434, 0.438]
jihadist	disgust	0.308	[0.303, 0.313]
religious	joy	0.300	[0.298, 0.303]

Table 2: Dominant emotion per ideology based on *pysentimento* analysis (Pérez et al., 2021), with mean probability and 95% Confidence Interval (CI)

Bootstrapping was employed to generate 95% confidence intervals for mean emotion probabilities. This non-parametric method was chosen because it avoids assumptions about the underlying distribution, providing robust estimates of uncertainty directly from the data, and thus enabling reliable identification of dominant emotions.

Permutation MANOVA was selected to test overall differences in multivariate emotion profiles due to its robustness to violations of parametric assumptions often encountered with linguistic data. Subsequent ANOVAs with FDR correction and Tukey HSD tests were conducted to identify which specific emotions contributed to these overall differences and to determine which ideology pairs showed significant differences, providing a detailed comparative analysis.

Table 2 reveals primary emotional expressions for eight analyzed extremist ideologies, three far-right political, three manospheres, and two religious. *Disgust* is the dominant emotion for the majority of ideologies, with **MGTOW** showing the highest average probability for it at 0.574, followed by **far-right** at 0.481. The **religious** extremist ideology is an exception, with *joy* being its dominant emotion (mean probability 0.300). Other emotions like *anger*, *fear*, and *sadness* generally show lower mean probabilities across groups but contribute to the overall emotional profile.

The analysis of emotion trends over time reveals dynamic shifts: For **alt-right**, disgust is consistently the most probable emotion from 2013 to 2018, peaking around 0.63 in 2018–2019, before declining; joy generally increases over time (Figure 1). For **far-right**, disgust is overwhelmingly dominant, starting above 0.65 around 2001 and showing a gradual decline (Figure 2). For **jihadist**, the emotion trends over time are more stable (Figure 3).

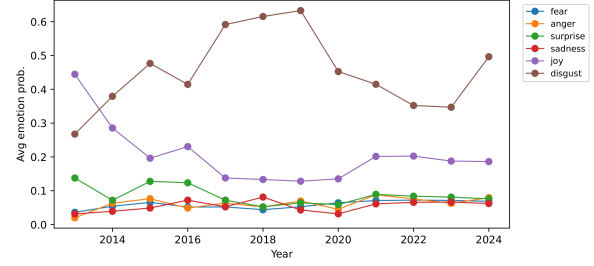


Figure 1: Emotion trends for **alt-right**

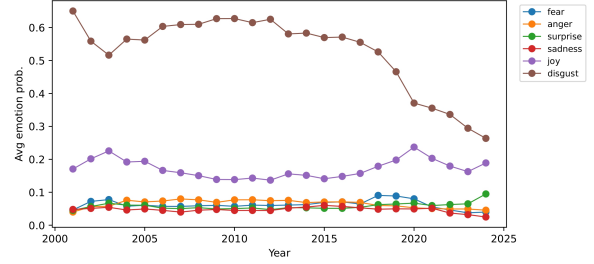


Figure 2: Emotion trends for **far-right**

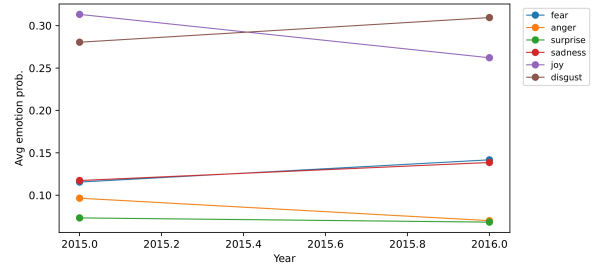


Figure 3: Emotion trends for **jihadist**

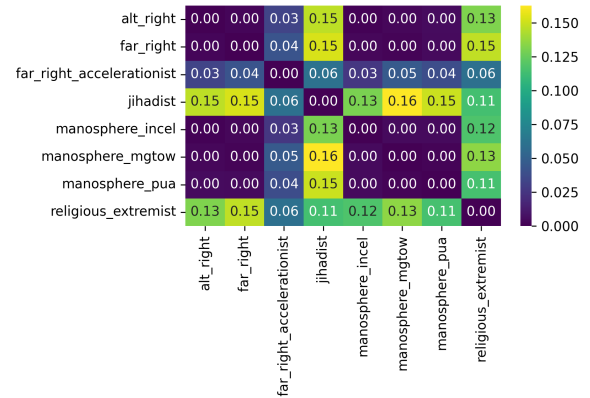


Figure 4: Cosine distance between ideologies in emotion space (80k dataset)

A permutation MANOVA confirmed a statistically significant overall difference in mean emotion profiles across the various ideologies (Pillai's trace = 0.046, $p < 0.001$, 1,000 permutations), indicating that the emotional landscapes of these groups are distinct. To visualize these dissimilarities, a cosine distance heatmap (Figure 4) represents the

Group 1	Group 2	Emotion	Δ
MGTOW	accelerationist	disgust	-0.32
MGTOW	jihadist	disgust	-0.27
MGTOW	religious	disgust	-0.26
far-right	accelerationist	disgust	-0.22
alt-right	accelerationist	disgust	-0.21

Table 3: Top-5 largest significant mean differences between ideology pairs (Tukey HSD, 80k dataset)

dissimilarity ($d_{ij} = 1 - \frac{\mu_i \mu_j}{\|\mu_i\|_2 \|\mu_j\|_2}$) between the mean emotion vectors (μ) of each ideology.

Per-emotion one-way ANOVAs, followed by the Benjamini-Hochberg false discovery rate (FDR) correction, identified specific emotions that differed significantly across ideologies. Pairwise Tukey-HSD tests pinpoint specific inter-ideology mean differences for each emotion (Table 3).

Classification experiments were conducted using traditional machine learning baselines: SVM with TF-IDF features, Random Forest with engineered emotion and rhetorical features, and the transformer-based models DistilBERT and DeBERTa, leveraging libraries such as Scikit-learn (Pedregosa et al., 2011) and Transformers by Hugging Face (Wolf et al., 2020). Table 4 summarizes the mean macro-F1 scores across cross-validation folds for these models and dataset variants. The results consistently demonstrate that transformer-based models significantly outperform traditional machine learning baselines, while the transformers’ performance also generally improved with increasing dataset size. Standard deviations are omitted for brevity in the table, but as can be seen, DeBERTa (He et al., 2020) generally achieved the highest F1 scores (e.g., 0.6182 ± 0.1548 for the 80k model), with DistilBERT (Sanh et al., 2019) not far behind (e.g., with 0.5741 ± 0.1466 for the 80k model).

To gain insight into the linguistic patterns learned by the transformer models, SHapley Additive exPlanations (SHAP) values (Lundberg and Lee, 2017) were computed, to reveal themes and specific phrases indicative of certain ideologies. The analysis indicated that models identify ideologically salient themes and phrases, e.g., misogynistic rhetoric in manosphere texts and specific political grievances or conspiracy theories in far-right texts. This confirms that the classifications are not arbitrary but are based on learned linguistic characteristics relevant to the ideologies.

Model	5k	10k	40k	80k
SVM	0.096	0.096	0.120	0.072
RF	0.194	0.194	0.187	0.183
DistilBERT	0.449	0.496	0.501	0.574
DeBERTa	0.467	0.538	0.547	0.618

Table 4: Mean macro-F1 scores for ideology classification across dataset sizes

6 Discussion

The first research question aimed to identify the dominant emotional and affective cues in extremist discourse. The analysis, leveraging *pysentimiento* for a 6-emotion framework, revealed a more complex affective landscape than often portrayed.

A key finding is the prevalence of *disgust* as the most prominent average emotion across a majority of the analyzed extremist ideologies, including alt-right, far-right, jihadist, and various manosphere subgroups (incel, MGTOW, PUA). Notably, the mean probability of disgust was particularly high for MGTOW (0.574). This mass expression of disgust might be regarded as an essential rhetorical weapon for these groups, perhaps used to dehumanize out-groups, similarly to how ethnic minorities, political adversaries, and, above all, women are silhouetted in manosphere circles to justify violence or discrimination. This aligns with social psychological theories of intergroup conflict and prejudice, which highlight the role of disgust in moral exclusion (Beck and Deffenbacher, 1999).

Disgust can also be used to express profound dissatisfaction with societal norms, political systems, or perceived threats to one’s ideological worldview, thereby reinforcing the “us-versus-them” thinking detailed in the characterization of political radicalization. The use of disgust as a rhetorical tool solidifies in-group boundaries by establishing shared objects of contempt, thereby reinforcing group identity and cohesion, a dynamic particularly evident in the manosphere as discussed by Ging (2019).

The accelerationist far-right group, while also showing disgust as its highest internal average, had a notably lower absolute level of disgust (0.172). This, coupled with generally lower emotion probabilities overall for this group, might indicate a more ideologically driven, perhaps less emotional in a broad sense, or more coded form of communication. This could align with accelerationist aims of systemic disruption, potentially relying on intel-

lectualized justifications or nihilistic apathy rather than widespread emotional appeals.

In stark contrast, the religious extremist group uniquely exhibited *joy* (0.300) as dominant emotion. This could signify an inward focus on communal solidarity, along with spiritual fulfillment, the promise of salvation, or the celebration of shared faith and values, rather than an outward projection of negativity. This suggests that “religious extremism,” as operationalized in this study using data from platforms like the “Traditional Catholicism” Discord server, may employ different primary affective strategies compared to more politically or identity-focused extremist movements, perhaps aiming to attract followers through the positive reinforcement of shared beliefs.

The temporal analyses further revealed the dynamic nature of emotional expressions, potentially reflecting strategic shifts or reactions to socio-political events. However, the interpretation of these trends must be cautious, especially for groups with limited coverage, such as the jihadist data derived solely from a 2014–2016 Twitter corpus.

The second research question compared these groups’ emotional landscapes. The permutation MANOVA results confirmed statistically significant overall differences in mean emotion profiles, supporting the hypothesis that different extremist ideologies cultivate distinct affective environments. Clustering of manosphere sub-groups (MGTO, PUA, incel) alongside alt-right and far-right suggests shared affective underpinnings. These are likely rooted in common grievances against perceived societal changes, anti-feminist sentiments, or similar “othering” tactics that focus on disgust. This observation aligns with studies of the ideological overlaps within the broader “manosphere” and its connections to far-right narratives (Ging, 2019; Van Eerten et al., 2018).

Conversely, religious extremist and jihadist groups occupy more distinct positions in the emotional space, as does accelerationist (right-wing), indicating unique emotional signatures. These distinctions are important in understanding how each ideology composes its worldview and addresses its base, and point away from a monological perspective on ‘extremist emotion.’ Such understanding is necessary for customizing counter-messaging, given that the same strategies for pitting Muslims against one another may not be effective in pushing back against one group’s emotional sales pitch, while the opposite may be true for another.

The third research question assessed the capability of classifiers to differentiate extremist discourse. The findings are compelling. Transformer-based models, particularly DeBERTa (He et al., 2020), significantly outperformed traditional machine learning baselines — SVM (Cortes and Vapnik, 1995) and Random Forest (Breiman, 2001). DeBERTa achieved a macro F1-score of 0.6182 on the largest dataset (capped at 80,000 posts per source; 1.29 million messages), demonstrating a strong capacity to learn linguistic patterns.

The overarching goal of this work was to determine the key linguistic markers that characterize different types of extremist groups’ discourse on social media, providing insights into their persuasive strategies and supporting the development of effective measures to counter radicalization.

The investigation into dominant emotional cues revealed that while specific emotional “fingerprints” vary, emotions like *disgust* are prevalent across several ideologies, potentially serving as a key linguistic marker associated with out-group derogation and in-group bonding. Conversely, the dominance of *joy* in the religious extremist group highlights that not all extremist rhetoric defaults to overtly negative affect, offering nuanced insights into their persuasive strategies, which may focus on positive in-group reinforcement.

The comparative analysis further substantiated that these emotional and linguistic patterns are sufficiently distinct to differentiate ideologies. By establishing a framework to compare these discourses, this work identified both shared affective undercurrents, suggesting common psychological mechanisms and unique ideological markers. This capacity to distinguish between groups based on their emotional expression is a crucial step towards understanding their varied persuasive appeals, and consequently, tailoring countermeasures.

Finally, the successful application of AI-driven classification models demonstrated that the identified linguistic and emotional markers are not only qualitatively observable but also computationally learnable and predictive, particularly with the DeBERTa architecture. The ability of these models to reliably differentiate extremist discourse based solely on linguistic features, with performance improving with dataset size, directly supports the goal of identifying such markers. This offers a realistic technological route to monitoring and analysis tools that could assist in the development of successful counter-extremism measures.

The choice of DistilBERT and DeBERTa was made to balance efficiency with performance for a robust evaluation. The fine-tuning strategy, AdamW optimizer, and learning rate schedule followed established best practices for these models to maximize their potential (Devlin et al., 2019). The use of StratifiedGroupKFold with creator_id was a critical choice to ensure a more realistic and less inflated measure of generalization performance, as it prevents author-specific linguistic styles from leaking between the training and test sets. A limitation is that even with SHAP (Lundberg and Lee, 2017), the full interpretability of why transformers make certain classifications remains challenging, especially at a deep semantic level.

The decision to frame ideology classification as a single-label problem, assigning each text to a single primary ideology, was a simplification for tractability in this initial investigation. The limitation is that extremist ideologies can be fluid and overlapping, and individuals may draw upon multiple ideological frameworks simultaneously. Future improvements could involve exploring multi-label classification approaches, where a text can be associated with multiple relevant ideological categories, or hierarchical classification to capture broader ideological families and more specific subtypes. This would better model the nuanced reality of online extremist expression.

7 Conclusion and Future Work

The study has shown that different emotional profiles are characteristic of various extremist ideologies, with the emotion *disgust* being prominent in some. Still, not all of them, such as religious extremism, focus on *joy*. These patterns provide critical information to compare the affective registers that these groups mobilize.

Moreover, it was found that these emotional and linguistic forms are not only distinct but also systematically relatable, unveiling not only commonalities but also unique ideological “fingerprints”. This comparison extends beyond a monolithic interpretation of extremist speech.

Most critically, transformer models, notably DeBERTa (He et al., 2020), were able to reliably differentiate between extremist ideologies based on learning these linguistic and emotional markers with significant accuracy.

Collectively, these findings underscore the power of computational linguistics to dissect and under-

stand the complex ways extremist groups use language to persuade, polarize, and mobilize. The identified markers and the successful classification hereunder provide a better empirical basis for further research and the development of evidence-informed approaches to address the prevalent issue of online radicalization, leading to safer cyberspaces and more resilient societies. While the path to mitigating online extremism is complex and requires ongoing vigilance, this research provides valuable analytical tools and insights to support this critical endeavor.

A critical area for future work is the collection and analysis of longitudinal data. This would enable researchers to track the evolution of emotional expression and linguistic markers within specific communities or among individual users over time. This may also be useful in connecting these shifts to news events, deplatforming attempts, intervention campaigns, and resources for deradicalization.

Limitations

The data strategy involved aggregating data from institutional sources (ExtremeBB, CrimeBB from the Cambridge Cybercrime Centre) and public repositories. This choice was made to ensure broad coverage of ideologies and allowed for the potential to make comparisons with earlier research, which included public datasets. The per-source, sampling limit employed during preprocessing was an attempt to mitigate temporal biases and prevent the undue dominance of very large sources in the combined corpus. The primary limitation stemming from this choice is that the study relied on data pre-filtered by the collection methodologies and specific timeframes of these sources. Despite efforts to sample evenly, inherent differences in original data availability resulted in varied message counts per ideology (see Table 1) and sparsity for some sources. This means that findings for smaller or less active groups or ideologies primarily represented by older corpora, such as the Jihadist data, might be less robust or generalizable. Specifically, the emotional profiles and classification performance for these less-represented groups should be interpreted with this data constraint in mind.

The decision to aggregate data from diverse platforms was made to capture a broader spectrum of online extremist expression, recognizing that different platforms may host different facets of these ideologies. However, discourse styles and platform

affordances vary significantly, as X's character limits contrast with the long-form nature of forum posts. The limitation is that the current study did not perform an explicit comparative analysis of how these platform-specific characteristics might shape emotional expression or linguistic markers differently within the same broad ideology.

The decision to utilize anonymized data from institutional sources was primarily driven by ethical considerations and data access agreements. This ensured compliance with privacy norms and the terms of data usage. The limitation is that anonymization, while essential, prevents tracking individual radicalization pathways or conducting nuanced, user-level longitudinal analyses across platforms. Furthermore, the pre-collected nature of the datasets limits access to the dynamic, real-time contextual metadata that direct observation might provide.

For emotion analysis, *pysentimiento* (Pérez et al., 2021) was chosen as the primary tool for the final analysis of six core emotions. This decision was based on its state-of-the-art transformer-based architecture, which offers more context-aware emotion detection compared to lexicon-based methods such as NRCLEX (Mohammad and Turney, 2013). This nuanced understanding of affect was deemed crucial for analyzing potentially ambiguous online extremist text.

Another advantage of using *pysentimiento* is that it has been explicitly evaluated on emotion analysis in a hate speech detection context, which contrasts with using other potential transformer-based models such as Mistral-Small-3.1-24B-GoEmotions (Korab, 2026) trained on the GoEmotions dataset (Demszky et al., 2020), or affective instruction-tuned LLMs (Liu et al., 2024). Meinert and Koufakou (2026) question the current utility of the latter strategy, however, with non-conclusive results, leaving the field open for further testing. A negative with *pysentimiento* is its training data only being Twitter posts, its English model is BERTweet-based and was not validated on forum, Telegram or discord text.

Normalizing emotion scores to sum to 1 per message was a choice to ensure comparability across texts of varying lengths. A limitation with using *pysentimiento* is that it may still not fully capture sarcasm, irony, coded language ("dog whistles"), or highly context-specific emotional expressions common in these communities. The six discrete Ekman-style emotion categories might also oversimplify complex affective states, such as contempt.

The choice to use inclusive and exclusive pronoun ratios was a pragmatic one, providing a simple, computationally inexpensive proxy for basic rhetorical stance related to in-group/out-group dynamics. These were selected based on their common role in identity marking. The limitation is that these two features offer only a very superficial view of complex extremist rhetoric. More sophisticated rhetorical or linguistic feature engineering, measures of narrative complexity, argumentation styles, and specific propaganda techniques were considered alternatives but were not pursued to keep the feature set for the Random Forest baseline relatively simple and to focus on the capabilities of transformer models with raw text.

Acknowledgments

We are grateful to Professor Richard Clayton and the Cambridge Cybercrime Centre, UK, for their trust and for granting access to the ExtremeBB and CrimeBB datasets, which were fundamental to the empirical contributions of this work.

All the reported studies were carried out at the Department of Computer Science at the Norwegian University of Science and Technology (NTNU) in Trondheim, Norway, and were supported by the university and department, in particular in terms of computing power and facilities.

To ensure reproducibility, all Python code and data access protocols used in the experiments described in the paper are available on GitHub at

<https://github.com/HasanOma/the-language-of-radicalization>

References

- Izzeldin Abuelaish and Neil Arya. 2017. *Hatred — A Public Health Issue*. *Medicine, Conflict and Survival*, 33:125–130.
- Apache. 2025. *Apache feather: A lightweight data format for efficient storage and processing*. Accessed: 2025-05-25.
- Stephane J. Baele, Katharine Boyd, Travis G. Coan, and Elahe Naserian. 2023. *Lethal words: An integrated model of violent extremists' language*. *Studies in Conflict & Terrorism*.
- Aaron T. Beck and Jerry Deffenbacher. 1999. *Prisoners of hate: The cognitive basis of anger, hostility and violence*. *Journal of Cognitive Psychotherapy*, 14:201–203.

- Leo Breiman. 2001. [Random forests](#). *Machine Learning*, 45(1):5–32.
- Louisa Buckingham and Nusiebah Alali. 2020. [Extreme parallels: A corpus-driven analysis of ISIS and far-right discourse](#). *Kōtuitui: New Zealand Journal of Social Sciences Online*, 15(2):310–331.
- Gian Vittorio Caprara, Michele Vecchione, Shalom H. Schwartz, Harald Schoen, Paul G. Bain, Jo Silvester, Jan Cieciuch, Vassilis Pavlopoulos, Gabriel Bianchi, Hasan Kirmanoglu, and 1 others. 2017. [Basic values, ideological self-placement, and voting: A cross-cultural study](#). *Cross-Cultural Research*, 51(4):388–411.
- Corinna Cortes and Vladimir Vapnik. 1995. [Support-vector networks](#). *Machine Learning*, 20(3):273–297.
- Patricia Alonso del Real and Oscar Araque. 2023. [Contextualization of a radical language detection system through moral values and emotions](#). *IEEE Access*, 11:119634–119646.
- Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. [GoEmotions: A dataset of fine-grained emotions](#). *arXiv preprint arXiv:2005.00547*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 4171–4186.
- Marwa Mohamed Khamis El-Zouka. 2019. [The construction of collective identity in terrorist discourse: A case study of ISIS’ Rumiya magazine](#). *Journal of the Faculty of Arts, Alexandria University*, 69(95):1–53.
- Fifth Tribe. 2020. [How ISIS uses Twitter](#). Kaggle.
- Debbie Ging. 2019. [Alphas, betas, and incels: Theorizing the masculinities of the Manosphere](#). *Men and Masculinities*, 22(4):638–657.
- Jesse Graham, Jonathan Haidt, and Brian A. Nosek. 2009. [Liberals and Conservatives rely on different sets of moral foundations](#). *Journal of Personality and Social Psychology*, 96(5):1029–1046.
- Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. [Deberta: Decoding-enhanced bert with disentangled attention](#). *arXiv*, 2006.03654.
- Sepp Hochreiter and Jürgen Schmidhuber. 1997. [Long Short-Term Memory](#). *Neural Computation*, 9(8):1735–1780.
- M. Luqman Jamil, Sebastião Pais, João Cordeiro, and Gaél Dias. 2022. [Detection of extreme sentiments on social networks with BERT](#). *Social Network Analysis and Mining*, 12(1):1–13.
- John T. Jost, Jack Glaser, Arie W. Kruglanski, and Frank J. Sulloway. 2003. [Political conservatism as motivated social cognition](#). *Psychological Bulletin*, 129(3):339–375.
- Lisa Kaati, Amendra Shrestha, and Katie Cohen. 2016. [Linguistic analysis of lone offender manifestos](#). In *2016 International Conference on CyberCrime and Computer Forensics*, pages 141–143. ICCCF.
- Taehoon Kim, Kevin Wurster, and Tahir Jalilov. 2025. [emoji 2.15.0: Emoji for Python](#).
- Petr Korab. 2026. [Mistral-Small-3.1-24B-GoEmotions: Multilabel emotion recognition for English text](#).
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. [RoBERTa: A robustly optimized BERT pretraining approach](#). *arXiv preprint arXiv:1907.11692*.
- Zhiwei Liu, Kailai Yang, Qianqian Xie, Tianlin Zhang, and Sophia Ananiadou. 2024. [EmoLLMs: A series of emotional large language models and annotation tools for comprehensive affective analysis](#). In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD ’24*, page 5487–5496, Barcelona, Spain. ACM.
- Ilya Loshchilov and Frank Hutter. 2019. [Decoupled weight decay regularization](#). *arXiv preprint arXiv:1711.05101 (Published at ICLR 2019)*.
- Scott M. Lundberg and Su-In Lee. 2017. [A unified approach to interpreting model predictions](#). In *Conference on Neural Information Processing Systems (NIPS)*.
- Timothy Meinert and Anna Koufakou. 2026. [Do LLMs outperform fine-tuned Transformers in emotion classification? a case study of Llama and RoBERTa on an emotion benchmark](#). *The International FLAIRS Conference Proceedings*, 39(1).
- Saif M. Mohammad and Peter D. Turney. 2013. [NRC emotion lexicon](#). Technical Report 21270984, National Research Council of Canada.
- Samuel Oakford, Chris Dehghanpoor, James O’Donnell, and Shane Harris. 2023. [‘problematic pockets’: How Discord became a home for extremists](#). *The Washington Post*, December 12.
- Emma Onraet, Alain Van Hiel, Kristof Dhont, Gordon Hodson, Mark Schittekatte, and Sarah De Pauw. 2015. [The association of cognitive ability with Right-Wing ideological attitudes and prejudice: A meta-analytic review](#). *European Journal of Personality*, 29(6):599–621.
- Sergio Pastrana, Daniel R. Thomas, Alice Hutchings, and Richard Clayton. 2018. [CrimeBB: Enabling cybercrime research on underground forums at scale](#). In *WWW ’18: Proceedings of The Web Conference 2018*, pages 1845–1854, Lyon, France. ACM.

- Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Roman Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. [Scikit-learn: Machine learning in Python](#). *Journal of Machine Learning Research*, 12:2825–2830.
- James W. Pennebaker, Martha E. Francis, and Roger J. Booth. 2001. [Linguistic Inquiry and Word Count: LIWC 2001](#). Erlbaum Publishers.
- Juan Manuel Pérez, David Furman, Laura Alonso Alemany, Lucas Milgrom, Franco Ghignone, and Juan Ormachea. 2021. [pysentimiento: A Python toolkit for sentiment analysis and social NLP tasks](#). In *Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS)*, pages 192–199. Association for Computational Linguistics.
- PostgreSQL Global Development Group. 2025. [PostgreSQL: The world’s most advanced open source database](#).
- Arundarasi Rajendran, Vattikuti Sree Sahithi, Chhavi Gupta, Madhuri Yadav, Swati Ahirrao, Ketan Kotecha, Mayur Gaikwad, Ajith Abraham, Nada Ahmed, and Sarah M. Alhammad. 2022. [Detecting extremism on Twitter during U.S. Capitol Riot using deep learning techniques](#). *IEEE Access*, 10:133052–133065.
- Sahar Rasoulkolamaki, Surinderpal Kaur, Alena Zhdanova, and Noor Aqsa Nabila Mat Isa. 2025. [In-group and out-group identity construction in extremist discourse: a critical multimodal approach](#). *Behavioral Sciences of Terrorism and Political Aggression*, 17(4):377–395.
- Jeff Reback, Wes McKinney, Joris Van Den Bossche, Tom Augspurger, Phillip Cloud, Sinhrks Hawkins, Martin Roeschke, Other Pandas Readers, Pietro Battiston, Kevin Hater, and 1 others. 2020. [pandas-dev/pandas: Pandas](#). Zenodo.
- Richard Rogers. 2020. [Deplatforming: Following extreme Internet celebrities to Telegram and alternative social media](#). *European Journal of Communication*, 35(3):213–229.
- Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. [DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter](#). *arXiv preprint arXiv:1910.01108*.
- Chris G. Sibley, Danny Osborne, and John Duckitt. 2012. [Personality and political orientation: Meta-analysis and test of a threat-constraint model](#). *Journal of Research in Personality*, 46(6):664–677.
- Anton Törnberg and Petter Törnberg. 2025. [White supremacists anonymous: how digital media emotionally energize far-right movements](#). *Journal of Information Technology & Politics*, 22(1):131–148.
- Raphael Vallat. 2018. [Pingouin: statistics in Python](#). *Journal of Open Source Software*, 3(31):1026.
- Jan-Jaap Van Eerten, Bertjan Doosje, Elly A. Konijn, Beatrice A. de Graaf, and Marieke de Goede. 2018. [The appeal of the manosphere: A study into the role of identity, belongingness, and radical expression in online male-separatist communities](#). *New Media & Society*, 20(12):4536–4556.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. [HuggingFace’s Transformers: State-of-the-art natural language processing](#). *arXiv*, 1910.03771.