# Psycholinguistic and Structural Features in Manifestos of Violent Offenders

Ema Kahr
*Institute for Complex Networks*
*Vienna University of Economics and Business*
Vienna, Austria
ema.kahr@wu.ac.at

Mark Strembeck
*Vienna University of Economics and Business*
*Secure Business Austria (SBA)*
*Complexity Science Hub (CSH)*
Vienna, Austria
mark.strembeck@wu.ac.at

***Abstract*—In this paper, we analyze personal writings of lone-actor terrorists and mass shooters to uncover how violent intent is linguistically expressed. Moving beyond traditional text analysis, we integrate psycholinguistic theory with network-based deep learning to generate interpretable insights into violent language. Each document is represented as a node in a semantic proximity graph enriched with LIWC, Empath, and grievance dictionary features. Using GraphSAGE embeddings and community detection, we found that violent texts form dense, highly cohesive, and isolated communities, distinguished by pervasive expressions of grievance, hate, and weapon-related language. These clusters reveal stable emotional, behavioral, and moral markers that robustly separate violent writings from non-violent controls, highlighting the dual role of content and cohesion in extremist discourse.**

***Index Terms*—attribute-based community detection, GraphSAGE, psycholinguistics, terrorist manifesto, violent language**

## I. Introduction

Mass shootings and lone-actor terrorist attacks have gained increasing visibility in recent years, prompting urgent inquiries into the psychological and ideological conditions that precede such violence. While a substantial body of scientific studies has focused on the rhetoric of political leaders and organized extremist movements, comparatively little attention has been directed toward the personal writings – such as manifestos, blog entries, or suicide notes – authored by individuals who ultimately engaged in mass violence. Prior research has examined the thematic structures and ideological foundations of texts authored by lone-wolf perpetrators (see, e.g., [1], [2], [3]), yet significant gaps remain in understanding how such texts differ linguistically and psychologically from those written by non-violent individuals. Studying these differences offers critical insight into the cognitive, emotional, and ideological mechanisms that may underlie violent intent. As a medium of thought and affect, language encodes social positioning and psychological states. As shown in [2] and [4], a systematic analysis can uncover latent indicators of grievance, distress, and moral disengagement that often precede acts of targeted violence.

Research shows that both mass shooters and individuals involved in terrorism often experience overlapping psychological and social vulnerabilities, such as social alienation, identity crises, and an unmet need for belonging. Central to their pathways to violence is a deeply held sense of grievance and subjective injury, which can become powerful motivational forces [5]. These emotional and ideological wounds are frequently encoded in their writings—often through heightened expressions of anger, moral absolutism, or existential despair. Therefore, understanding how such grievances manifest linguistically can aid in distinguishing violent intent from non-violent discontent.

In particular, this study focuses on texts written by individuals who acted alone or in loosely connected ideological networks – excluding prominent figures such as Adolf Hitler or Osama bin Laden, whose texts function more as propaganda than personal expression. High-risk texts examined in this study include terrorist manifestos, perpetrators' online blog entries, and abusive or hate-laden messages calling for violence or self-harm (including sacrificial suicide). These texts pose significant threats not only due to their content but also due to their potential to inspire imitation. For example, a manifesto written by a Norwegian neo-Nazi (Anders Behring Breivik) influenced an Australian anti-immigrant extremist (Brenton Tarrant), whose writings later inspired a Texan white supremacist (Patrick Crusius) (see [6], [7], [8]).

Despite advances in computational threat assessment, existing methods often face a trade-off between interpretability and predictive performance. Psycholinguistic analysis, for instance, yields interpretable insights into emotional tone, cognitive distortions, and moral values but often lacks the power of modern deep learning techniques. Conversely, state-of-the-art classification models, particularly those relying on deep neural networks, perform well but operate as black boxes, raising ethical concerns in high-stakes domains where transparency is crucial. To address these challenges, this study bridges computational social science, psycholinguistics, and moral psychology by treating language not merely as content but as structured data embedded with signals of cognition, emotion, moral values, and perceived grievances.

This work presents two main contributions. First, we integrate psycholinguistic theory with deep graph-based learning to represent documents as nodes enriched with multidimensional linguistic features (capturing cognition, emotion, moral values, and grievance signals), while leveraging the network structure of linguistic similarity. Second, we go beyond binary

detection by applying community detection on GraphSAGE embeddings, revealing latent psychological and ideological groupings that distinguish violent from non-violent authors. This approach advances the state of the art by combining interpretability of psycholinguistic features with the representational power of deep learning.

The remainder of this paper is structured as follows: Section II reviews related work. Section III outlines our research methodology. Section IV presents the results, which are further discussed in Section V. We conclude our study in Section VI, offering directions for future research.

## II. Related work

Motivated by a previous study which showed that in-group commitment in radical and extremist groups leads to an elimination of a moral concern about cruel behavior towards fellow humans [9], Vergani et al. [10] focused on authoritarianism and religiousness in texts published in the Dabiq and Inspire magazines. Vergani et al. found that these two features have a profound influence on mobilization by ISIS in the Western world.

Furthermore, Conway and Conway [3] argued that timely analysis of the terrorist rhetoric could be a key element in assessing the potential level of danger that a terrorist group actually poses in the real world. The researchers found that the typical language styles used by terrorists became more and more exaggerated ahead of a planned terrorist attack. According to [3], the rhetoric of terrorists is different from that of other humans. For one, terrorists use simpler rhetoric. Second, affiliation and the sense of community play an essential role in terrorist rhetoric. For the most part, this sense of affiliation is designed to be directed towards the interpersonal relationships of people within the terrorist group which, again, serves as a powerful binding tool towards recruits. Power is another category which is commonly observed in the writings of terrorist groups.

Kaati et al. [2] analyzed the use of personal pronouns in terrorist manifestos, where the use of "they" or "them" effectively suggested that a large part of the group's identity comes from the existence of another group with opposing beliefs. Another common feature of such texts is the use of words that highly associate with power (e.g., words such as "superior" or "weak").

Nouh et al. analyzed linguistic features of the Dabiq magazine to derive linguistic and psychological features that ISIS uses to recruit new followers [11]. These features were subsequently used to distinguish between radical and non-radical tweets. In particular, the authors found that the use of personal pronouns, violent words, and long words were discriminating features for radical tweets.

Ebner [1] conducted an ethnographic content analysis of extremist manifestos and found that certain narratives are more commonly used by violent extremists who went on to commit terrorist attacks. Such dominant narratives include the portrayal of an existential threat, the normalization of violence within group norms, and the dehumanization of perceived opponents.

The political philosophy of Anders Breivik's manifesto was analyzed in [12], identifying key themes such as Islamophobia, culturally conservative nationalism, antifeminism, White Power ideology, and far-right evangelical theology. Building on this, Berntzen and Sandberg [13] conducted a comparative analysis between Breivik's manifesto and the rhetoric of the Norwegian anti-Islamic movement that influenced his actions. The study highlighted shared narratives, including the portrayal of Europe as undergoing Islamization, and framing the conflict as a struggle between the political elite and the common people. Berntzen and Sandberg concluded that lone wolf terrorists can be radicalized through rhetoric disseminated within broader social movements.

Ware et al. [6] argued that lone wolf terrorists often inspire one another through publicly shared manifestos, frequently citing or replicating each other's ideas. However, their influences also extend to non-violent ideologues and extremists. An analysis of six far-right terrorist manifestos [6] revealed recurring themes such as a perceived clash of cultures based on race and religion, dissatisfaction with the current political climate, the framing of terrorism as an act of self-defense or a last resort, and calls for further violence.

## III. Research method

### A. Research Questions

This paper examines whether psycholinguistic and thematic cues can effectively distinguish violent texts authored by terrorists and mass shooters from those written by non-violent authors. We explore how these linguistic markers, combined with structural and contextual information, reveal distinct narrative patterns unique to extremist rhetoric. To do this, we analyze document similarities using graph-based embeddings to identify clusters of related texts, testing if violent texts form distinct groups compared to our control corpora.

Our study is guided by three main research questions:

**RQ1:** Do violent texts cluster separately from other document types, indicating distinct narrative communities?

**RQ2:** Which psycholinguistic and semantic features most distinguish texts authored by violent offenders from those authored by non-violent authors?

**RQ3:** How does removing psycholinguistic markers affect the stability and structure of document communities?

In the scope of our study are:

1) Personal, self-expressive texts written by individuals associated with violent acts: manifestos, blog posts, social media posts, or letters addressed to the public or specific targets.
2) Focus is on Anglophone texts (or texts that have been translated with care) with confirmed authorship.

We excluded:

1) High-profile political leaders, whose texts are often ghost-written or written with strategic propaganda intent at a mass scale.
2) Formal statements by terrorist organizations.

### B. Data Acquisition

To collect the data, we first consulted corresponding Wikipedia pages[1] and other online sources[2] to compile a list of mass shooting events or lone-terrorist attacks and identify the associated perpetrators. Using these names, we conducted targeted searches to determine whether any manifestos or other writings have been authored by the perpetrators. Our search strategy included reviewing reputable news media sources such as The New York Times, BBC, or CNN, alongside other online references, including archived websites and discussion forums. When available, we retrieved the texts in the form in which they were originally published online, ensuring fidelity to the original content. This process resulted in a dataset of 156 texts written by violent actors, including manifestos, letters, blog/diary entries, notes, and transcripts of speeches made by the violent actors.

In order to compare the psycholinguistic style of violent actors to non-violent authors, we sampled a control corpus in a 1:2:2 ratio, where for each violent text, we included two texts from a high-emotion/argumentative control corpus (control corpus A) and two texts from an everyday language corpus (control corpus B). This oversampling was designed to capture subtle psycholinguistic patterns and ensure a broad coverage of thematic and stylistic variation. The selection of texts for the control corpus A was driven by several factors:

1) We included texts with strong emotional or moral appeals similar to violent texts, reflecting emotionally charged but non-violent discourse.
2) We included diverse control sources (including academic essays, Reddit posts on depression and suicide, NGO manifestos, 8kan posts, feminism books, political speeches, religious sermons, court documents, and customer complaints) to avoid overfitting to a single register.
3) We selected texts that are contemporaneous with the violent offender corpus to account for language evolution.

The control corpus B included texts that are thematically, tone-wise, and format-wise in contrast to the texts authored by the violent perpetrators, including jokes, children's and romantic song lyrics, scientific articles, youth novels, IMDB/Amazon/Goodreads reviews, horoscopes, weather reports, Wikipedia articles, game instructions, recipes, LLM-generated comedic stories, celebrity interviews, programming tutorials, travel blogs, university application essays, wedding vow transcripts, and product and pet care manuals. These texts represent everyday language and informational content, as well as entertainment content.

In total, both control corpora included 624 texts (312 texts each).

Manifestos naturally vary greatly in length, and we opted not to alter them to preserve their authentic discourse structure. Control corpora were sampled to match the manifestos' word count distribution as closely as possible, particularly in terms of median and range, see Table I.

[1]https://en.wikipedia.org/wiki/List_of_manifestos_of_mass_killers

[2]https://schoolshooters.info/

TABLE I
WORDCOUNT SUMMARY

| Dataset | Mean (stdev) | Range | Median |
|---|---|---|---|
| Manifesto | 12351.05±73,319.94 | 16 - 898,066 | 834.5 |
| Control corpus A | 9607.83±53,959.36 | 15 - 900,000 | 900.5 |
| Control corpus B | 11182.41±56,401.73 | 13 - 900,000 | 919.5 |

### C. Document pre-processing

Since texts in our corpus originated from diverse formats (PDFs, MS Word, and handwritten scans), we automated text conversion to a `.txt` file format (ASCII) using a custom Python pipeline. For PDFs, the script first attempted extraction with `pdfplumber` and if unsuccessful (e.g., for scanned documents), it applied OCR using Tesseract, with images generated by `pdf2image` and OCR language set to English. Extracted texts were cleaned by removing extraneous whitespace, non-ASCII characters, and irrelevant sections (e.g., headers, footers, metadata).

**Feature Extraction:** We extracted psycholinguistic features using LIWC-2022 (122 features) [14], including its Moral Foundations (33 features) and Grievances (37 features) add-on dictionaries. In addition, we used Empath [15] to obtain scores for the 192 pre-defined semantic categories (e.g., "violence", "aggression", "politeness"). In total, this resulted in 384 features. To ensure comparability, we normalized features by min-max normalization. We also removed duplicate LIWC features (found in its add-on dictionaries) and LIWC features which were deemed uninformative with respect to document analysis, including the wordcount, emojis, punctuation count, article counts, and parent categories of fine-grained features (e.g., "Tone" was removed, while we kept individual emotion scores for "anger", "anxiety", etc.).

### D. Network Construction

From the initial feature set, we identified the most relevant node-level attributes using a `RandomForestClassifier`. The classifier was trained to predict the document class (violent author corpus, control corpus A, or control corpus B) based on the full set of features. We then extracted feature importance scores to determine which features contributed most to distinguishing between the three classes. Second, we calculated the Pearson correlation between each feature and the target variable, selecting those with an absolute correlation above a threshold of 0.15. This process resulted in 55 selected features (see Table II).

Next, we generated edges by calculating the cosine similarity between LIWC-based features, specifically syntactic categories, punctuation, and writing style. Edges were created only if the similarity exceeded the 90th percentile threshold (cut-off value = 0.9641), thereby capturing meaningful linguistic connections. This resulted in a network with a density of 0.1026. The network counted 780 nodes and 31,161 undirected edges.

**Generating node-attribute embeddings:** We trained a GraphSAGE model [16] in a supervised setting using the

TABLE II
SUMMARY OF FEATURES USED IN NETWORK MODELING

| Feature Category | Features |
|---|---|
| **Emotional / Tone** | negative emotion, anger, hate, envy |
| **Cognitive Processes** | cogproc, discrep, focusfuture |
| **Social** | Social, socbehav, dominant_heirarchical, leader |
| **Behavioral** | conflict, aggression, dispute, deception, fight |
| **Moral** | moral, Care_Vice, Authority_Virtue |
| **Topical / Thematic** | art, beach, crime, death, domestic_work, government, home, leisure, military, ocean, pain, politics, power, prison, religion, sailing, shape_and_size, stealing, suffering, swimming, terrorism, torment, traveling, vacation, violence, war, water, weakness, weapon, weather |
| **Grievances / Threats** | Grievance, Murder, Soldier, Suicide, Threat, Weaponry |

cross-entropy loss function. During training, the loss consistently decreased, from 1.5183 at epoch 0 to 0.2838 at epoch 100, indicating effective convergence. Evaluation loss also improved, dropping from 0.2020 to 0.1033, further confirming enhanced classification performance over time.

The model's objective was to classify documents into three categories: (0) violent language corpus, (1) control corpus A, and (2) control corpus B. Each document was represented as a node within the graph, and the model was trained to predict document labels based on graph-structured information.

The dataset was divided into training, validation, and test sets using a stratified random split to preserve label proportions, with the resulting label distributions summarized in Table III. On the held-out test set, the model achieved an overall accuracy of 97.44%. Class-wise F1 scores were 0.97 for class 0, 0.97 for class 1, and 0.98 for class 2, demonstrating consistently strong performance across all classes.

To tune the GraphSAGE hyperparameters (including the number of hidden channels, number of layers, dropout rate, and learning rate), we utilized Optuna [17], with an objective function to minimize the training loss over 50 epochs. Hyperparameters used were `hidden_channels`: 65, `num_layers`: 3, `dropout`: 0.5942, and `lr`: 0.00897.

TABLE III
LABEL DISTRIBUTION ACROSS DATASETS

| Dataset | Manifesto | Control Corpus A | Control Corpus B |
|---|---|---|---|
| Train | 110 | 222 | 214 |
| Validation | 29 | 39 | 49 |
| Test | 17 | 51 | 49 |

After training, we generated node embeddings by extracting the output of the final GraphSAGE layer in evaluation mode without gradient updates. These embeddings capture rich semantic and relational information about each document, integrating both its psycholinguistic features and structural context from the linguistic similarity graph.

**Semantic Proximity Graph:** After obtaining the node embeddings from the trained GraphSAGE model, we constructed a new graph to better capture relationships between documents based on these learned representations. The embeddings for each document node were normalized to unit length to ensure that the cosine similarity effectively reflects angular distances. We computed a similarity matrix using cosine similarity between normalized embeddings. For each node, we identified its top $k = 15$ nearest neighbors (excluding itself). To ensure robustness, edges were included only if the neighbor relationship was mutual (i.e., both nodes include each other among their top $k$ neighbors). This resulted in a symmetric mutual k-NN adjacency matrix representing strong similarity connections. Finally, from this adjacency matrix, we created an undirected graph using the `igraph` library[3]. The resulting graph contained $780$ vertices and $4310$ edges with a density of $0.0142$.

### E. Community Detection

To detect communities, we relied on the Leiden algorithm [18] with the Constant Potts Model (CPM) objective function, which avoids the resolution limit of modularity by directly controlling edge density within communities. We used Optuna to tune the CPM resolution parameter over 50 trials in the range of 0.1 to 2.0, selecting the value that yielded the most coherent and interpretable community structure. The optimal parameter was 0.10503, resulting in 57 detected communities.

### F. Evaluation

To assess the robustness of the detected document communities, we conducted a systematic ablation study by progressively excluding feature categories from the input. Specifically, we iteratively removed one category, then all possible pairs, triplets, and so forth, until only a single feature category remained. For each ablation setting, we computed the Normalized Mutual Information (NMI) between the community assignments from the original full-feature model and those obtained after feature removal to quantify the stability of the community structure. We also analyzed the distribution of document classes within communities, recording the number and type of mixed-class communities as well as the number of singleton communities.

To ensure the reliability of the results, we repeated the entire ablation procedure under a range of hyperparameter configurations, using optimal parameter values for feature selection (via Random Forest), supervised GraphSAGE, and community detection. All other parameters were kept the same as in the full-feature model.

## IV. RESULTS

### A. General statistics

In total, our network counted 3542 intra-community edges and 768 inter-community edges, clearly reflecting a modular

[3]https://igraph.org/

organization within the network. The network was partitioned into 57 distinct communities (see Figure 1) varying significantly in size and connectivity. As shown in Figure 2 community sizes range from a single node to 30 nodes.

Five documents formed single-node communities. The identified singleton communities include 2 documents belonging to control corpus A (court documents), 2 documents from control corpus B (two reviews), and 1 from the violent language corpus. Their isolation suggests that these nodes have distinctive features that separate them from the larger community structures, highlighting the diversity and heterogeneity within the documents.

We also observed that the average degree within communities varied widely. A maximum observed average degree was 11.93 (community #1, including control corpus A documents) and the average maximum degree reached 15 in 14 communities. Degree centrality measures corroborated these trends, reflecting a higher concentration of connectivity in smaller, denser communities.

TABLE IV
SUMMARY STATISTICS OF NETWORK COMMUNITIES

| Statistic | Min | Max | Mean | Median |
|---|---|---|---|---|
| Number of nodes | 1 | 30 | 13.68 | 15 |
| Number of edges | 0 | 179 | 62.14 | 63 |
| Density | 0 | 1.0 | 0.57 | 0.58 |
| Average degree | 0 | 11.93 | 7.01 | 8.44 |
| Max degree | 0 | 15 | 9.77 | 12 |
| Average degree centrality | 0 | 1.0 | 0.57 | 0.58 |
| Max degree centrality | 0 | 1.17 | 0.76 | 0.82 |

The community structure reveals that 12 communities are dominated by violent language texts, while 18 and 20 communities are predominantly composed of control corpus A and control corpus B texts, respectively. As visualized in Figure 3, many communities are compositionally pure and contain documents exclusively from a single class. For example, community 0 consists entirely of violent language texts, while community 1 and 2 of control corpus A documents. A minority of communities exhibit class mixing. Community 33, for example, consists of 10 control corpus A documents, and 3 control corpus B documents. Community 19 includes 2 violent author texts and 16 control corpus A documents.

It is worth noting that mixing occurs rarely, suggesting that the violent language texts and control corpora form distinct and isolated communities, likely due to their unique narrative or structural features. Overall, the distribution answers our first research question: *document classes tend to cluster into cohesive communities, with minimal crossover from the violent author class.*

As seen in Table V, there are 11 inter-class edges, all of which connect violent language texts and control corpus A documents. All 11 edges belong to community 19.

### B. Feature analysis

To answer the second research question, we relied on the Kruskal-Wallis test in order to assess whether the distributions

TABLE V
EDGE CLASS DISTRIBUTION IN THE DOCUMENT SIMILARITY GRAPH. ABBREVIATIONS CCA AND CCB STAND FOR CONTROL CORPUS A AND CONTROL CORPUS B, RESPECTIVELY.

| Edge Type | Class Pair | Number of Edges |
|---|---|---|
| Intra-class | (ccA, ccA) | 1,740 |
| Intra-class | (ccB, ccB) | 1,755 |
| Intra-class | (violent author, violent author) | 782 |
| Inter-class | (ccA, ccB) | 22 |
| Inter-class | (ccA, violent author) | 11 |

of various linguistic, emotional, and thematic features differ significantly between violent offenders' texts and both control corpora, which drive the formation of communities.

Table VI presents the top 10 features exhibiting the strongest distinctions across documents based on the Kruskal-Wallis $H$ statistic. The results highlight the prominent role of emotional tone (hate), behavioral cues (conflict, threat, aggression), moral values (care, power), and topical themes (death, murder, weaponry, kill) in differentiating documents according to their psycholinguistic and thematic characteristics.

Furthermore, we conducted a Dunn post hoc analysis to reveal pairs of document classes which exhibit statistically significant differences. Violent language texts clearly separate from control corpus B (see very small p-values in Table VI, column D(0 vs 2)). The differences between violent language texts and control corpus A are statistically significant (except for 'power'), albeit the differences are lower compared to the control corpus B.

TABLE VI
STATISTICAL TEST RESULTS FOR TOP 10 STATISTICALLY SIGNIFICANT FEATURES: KRUSKAL-WALLIS (KW) AND DUNN POST-HOC COMPARISONS (D). ALL RESULTS ARE REPORTED FOR P-VALUE $< 0.05$. CORPORA ARE LABELED AS FOLLOWS: 0 - VIOLENT AUTHOR TEXTS, 1 - CONTROL CORPUS A, 2 - CONTROL CORPUS B.

| Feature | KW $H$ | D(0 vs 1) | D(0 vs 2) | D(1 vs 2) |
|---|---|---|---|---|
| **Emotional/tone** | | | | |
| Hate | 123.95 | 2.82e-09 | 1.01e-27 | 6.22e-09 |
| **Behavioral** | | | | |
| Conflict | 163.704 | 6.72e-07 | 1.30e-33 | 3.20e-17 |
| Threat | 148.283 | 1.12-05 | 5.00e-30 | 1.41e-16 |
| Aggression | 112.852 | 4.33e-05 | 1.36e-23 | 4.27e-12 |
| **Moral** | | | | |
| Care (Vice) | 183.135 | 2.24e-07 | 2.86e-37 | 1.88e-19 |
| Power | 115.435 | 0.135 | 1.03e-19 | 3.53e-18 |
| **Topical/Thematic** | | | | |
| Death | 146.618 | 4.64e-11 | 1.18e-32 | 3.88e-10 |
| Murder | 183.889 | 3.23e-11 | 9.42e-40 | 6.17e-15 |
| Weaponry | 153.371 | 1.31e-12 | 2.22e-34 | 1.56e-09 |
| Kill | 150.816 | 9.68e-11 | 2.31e-33 | 5.3e-11 |

### C. Linguistic line between manifestos and control corpora

We next turn to the analysis of the bridging documents (i.e., those connecting violent language texts and the control

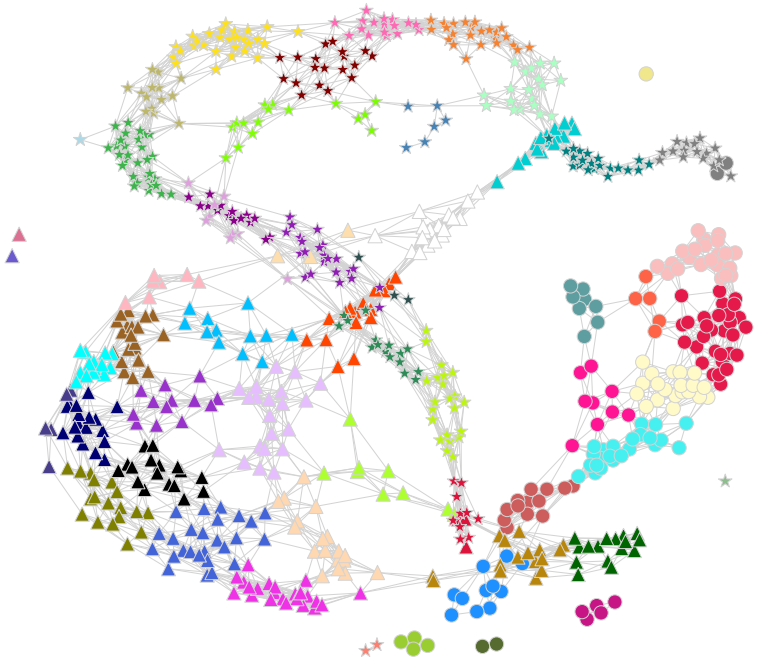

Fig. 1. Communities with Fruchterman-Reingold layout. Circles represent violent author corpus, stars control corpus A, and triangles control corpus B. Communities are plotted each in an own color.

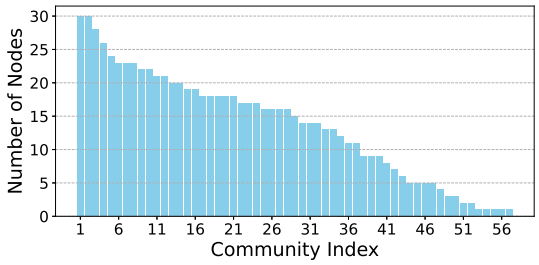


Fig. 2. Sizes of detected communities in the graph.

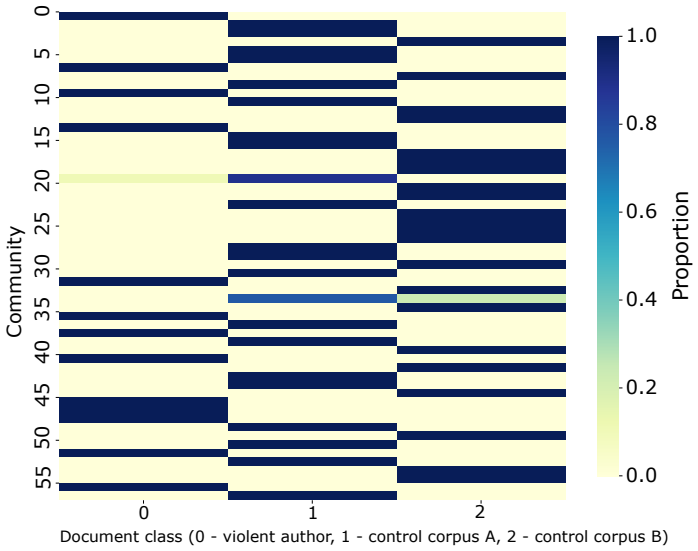


Fig. 3. Normalized community assignment per document class.

corpora), as they may contain clues about the transition between non-violent and violent intent. We detected 11 such bridging edges: two between violent offenders' texts and nine connecting to documents from control corpus A. Based on psycholinguistic features, while disregarding the topological structure, the two violent offenders' texts correlate positively but weakly (0.124288) and show a negative weak to moderate correlation with control corpus A documents (ranging from $-0.06$ to $-0.39$). Thus, from a purely psycholinguistic similarity perspective, violent offenders' texts and control corpora are distinct.

However, since these documents belong to the same cluster (#19), they are embedded together due to graph structure and supervision. This suggests they share higher-level relationships (e.g., topological or contextual proximity) that are not captured by psycholinguistic features alone.

This distinction is further illustrated by the Uniform Manifold Approximation and Projection (UMAP) projections. The projection of raw linguistic feature vectors, as shown in Figure 4, reveals a largely overlapping and scattered distribution of violent offenders' texts and control corpus documents, highlighting the subtle and noisy linguistic differences between these text types. This diffuse pattern suggests that purely psycholinguistic or lexical features alone provide limited discriminative power.

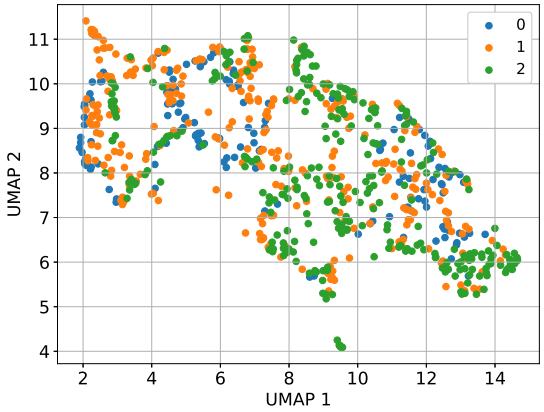


Fig. 4. UMAP of linguistic features.

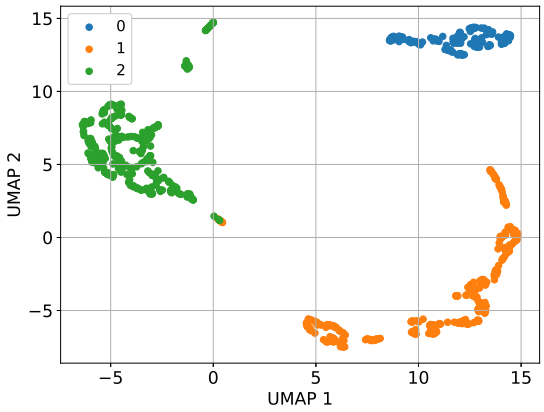


Fig. 5. UMAP of GraphSAGE embeddings.

In contrast, the UMAP visualization of GraphSAGE embeddings, which integrate both linguistic attributes and structural graph information through supervised learning, displays distinct and well-separated clusters (see Figure 5). This separation indicates that the graph-based embeddings successfully capture latent contextual and relational signals beyond surface linguistic cues, thereby revealing non-obvious patterns that distinguish violent offenders' texts in the document network.

### *D. Robustness testing: Ablation study*

As part of our ablation study, we systematically removed feature categories to assess their individual and joint contributions to the clustering performance.

Removing a single feature category resulted in clustering outcomes that were comparable to the full-feature model. Specifically, the Normalized Mutual Information (NMI) remained stable (0.57–0.58), and the number of communities and singleton communities remained within a similar range. A slight increase in community mixing was observed when removing the cognitive category, where violent language texts co-occurred with control corpus A in two communities, and control corpus A appeared with control corpus B in two additional communities. These findings suggest that no single feature category (e.g., emotional, moral, behavioral) dominates the clustering process or serves as the sole driver of signal strength. Instead, each feature category contributes complementary information.

Removing two feature categories had minimal impact on clustering: NMI ranged from 0.54 to 0.60, with 54–61 communities and 2–7 singleton nodes—similar to the original setup.

With three categories removed, NMI remained between 0.5434 and 0.6014, though variability increased (51–70 communities, 4–11 singletons), showing moderate resilience.

Excluding four categories led to more fluctuation: NMI dropped to 0.5192–0.5697, with up to 11 isolated nodes.

When five categories were removed, NMI ranged from 0.5299 to 0.5858, with 54–73 communities and up to 21 isolated nodes. This indicates that while feature redundancy helps maintain structure, heavy dimensional loss increases fragmentation.

The largest deviations from the original clustering emerged when only a single feature category was retained. While the NMI remained high (0.55–0.60), we observed a substantial increase in the number of communities (73–89) and a substantial rise in isolated node communities (22–38). This fragmentation suggests that the retained feature category captured distinct patterns but lacked the generalizability required to form cohesive, interpretable clusters. For instance, when isolated, the emotional, moral, and behavioral features result in over-segmentation, likely reflecting unshared signals.

Across multiple ablation settings, especially when more features were removed, the majority of isolated node communities consisted of documents from control corpus B.

## V. Discussion

We found that violent language texts and control texts are linguistically distinct, though psycholinguistic features alone provide only subtle, noisy differentiation. While violent texts cluster separately from controls in graph-structured embeddings, raw linguistic features show considerable overlap. This indicates that psycholinguistic signals alone offer weak distinctions, but integrating structural and contextual information via graph embeddings reveals deeper latent relationships, consistent with [3], who emphasized that terrorist rhetoric differs not only in lexical choice but also in the community and relational contexts it creates.

Ablation tests revealed that removing individual psycholinguistic feature categories, such as cognitive processes, does not substantially affect cluster purity, but it does increase fragmentation and the number of singleton communities. This suggests that these features primarily support the internal cohesion of clusters rather than serving as the main discriminators between classes. Cognitive features, in particular, may act as cohesion-enhancing elements, analogous to the role of *coreferentiality* and *lexical cohesion* in maintaining thematic consistency within texts. Redundancy across feature categories further reinforces this interpretation. Even under severe dimensional reduction, clustering performance remains

robust, indicating that overlapping information helps preserve community structure. These findings resonate with prior work on terrorist manifestos. For example, [19] showed that grievance-related linguistic markers can distinguish violent from non-violent texts, [20] highlighted how offensive, demonizing, and dehumanizing language contributes to the structure of multiple manifestos, and [21] demonstrated characteristic language profiles that define lone-actor terrorist writings. In this context, cohesion-enhancing psycholinguistic features help sustain ideological consistency, ensuring that references to key concepts like "justice", "oppression", or "enemy" are aligned throughout the text. These features also contribute to the coherent narrative structure and facilitate the identification of thematic patterns unique to violent writings.

Ethnographic observations complement these results. [1] described normalized violence and dehumanization narratives as central and relatively self-contained, while [13] noted isolated social narratives. This explains why document communities in our study are largely homogeneous, with only rare bridging edges, mainly between violent texts and control corpus A, reflecting tightly knit, class-specific discourse clusters.

### A. Limitations

Our results depend strongly on the corpora used. Due to limited public access to terrorist manifestos, diaries, and blog entries, the dataset remains incomplete. The control corpora were systematically gathered, but alternative non-violent texts could produce different results. We also note limitations in the tools: while LIWC is widely used, some feature inaccuracies may persist. Additionally, GraphSAGE lacks native support for edge weights and may miss subtle psycholinguistic nuances.

## VI. Conclusion

Our study showed that texts by violent offenders form linguistically and structurally distinct communities from control corpora, with high class purity and strong internal cohesion. Ablation analyses revealed that no single feature category drives clustering – rather, features primarily support the internal cohesion of clusters rather than acting as the main discriminators between classes. Cognitive features, in particular, may serve as cohesion-enhancing elements, while redundancy across emotional, moral, cognitive, social, grievance-related, and behavioral cues further preserved community structure. Our study showed that graph-based embeddings captured deeper relational patterns that separate violent texts from other writings, and the few bridging documents (mainly between manifestos and control corpus A) highlighted potential transitional links.

Future work will explore multiplex modeling of psycholinguistic dimensions and temporal splits to track language shifts over time.

## References


[1] J. Ebner, C. Kavanagh, and H. W. and, "Is There a Language of Terrorists? A Comparative Manifesto Analysis," *Studies in Conflict & Terrorism*, vol. 0, no. 0, pp. 1–27, 2022.

[2] L. Kaati, A. Shrestha, and K. Cohen, "Linguistic analysis of lone offender manifestos," in *2016 IEEE International Conference on Cybercrime and Computer Forensic (ICCCF)*, 2016, pp. 1–8.

[3] L. G. Conway and K. R. Conway, "The terrorist rhetorical style and its consequences for understanding terrorist violence," *Dynamics of Asymmetric Conflict*, vol. 4, no. 2, pp. 175–192, 2011.

[4] N. Brooks and J. Barry-Walsh, "Understanding the role of grievance and fixation in lone actor violence," *Frontiers in psychology*, vol. 13, no. 1045694, 2022.

[5] L. Y. Hunter, M. H. Ginn, S. Storyllewellyn, and J. R. and, "Are mass shootings acts of terror? Applying key criteria in definitions of terrorism to mass shootings in the United States from 1982 to 2018," *Behavioral Sciences of Terrorism and Political Aggression*, vol. 13, no. 4, pp. 265–294, 2021.

[6] J. Ware, "Testament to murder: The violent far-right's increasing use of terrorist manifestos," International Centre for Counter-Terrorism, Tech. Rep., 2020.

[7] M. Harwood, "Living Death: Imagined History and the Tarrant Manifesto," *Emotions: History, Culture, Society*, vol. 5, no. 1, pp. 25 – 50, 2021.

[8] J. Kupper, T. K. Christensen, D. Wing, M. Hurt, M. Schumacher, and R. Meloy, "The contagion and copycat effect in transnational far-right terrorism: An analysis of language evidence," *Perspectives on Terrorism*, vol. 16, no. 4, pp. pp. 4–26, 2022.

[9] M. Li, B. Leidner, and E. Castano, "Toward a comprehensive taxonomy of dehumanization: Integrating two senses of humanness, mind perception theory, and stereotype content model." *TPM: Testing, Psychometrics, Methodology in Applied Psychology*, vol. 21, no. 3, 2014.

[10] M. Vergani and A.-M. Bliuc, "The Language of New Terrorism: Differences in Psychological Dimensions of Communication in Dabiq and Inspire," *Journal of Language and Social Psychology*, vol. 37, no. 5, pp. 523–540, 2018.

[11] M. Nouh, J. R. Nurse, and M. Goldsmith, "Understanding the Radical Mind: Identifying Signals to Detect Extremist Content on Twitter," in *2019 IEEE International Conference on Intelligence and Security Informatics (ISI)*, 2019, pp. 98–103.

[12] M. G. and, "Crusader Dreams: Oslo 22/7, Islamophobia, and the Quest for a Monocultural Europe," *Terrorism and Political Violence*, vol. 26, no. 1, pp. 129–155, 2014.

[13] L. E. Berntzen and S. S. and, "The Collective Nature of Lone Wolf Terrorism: Anders Behring Breivik and the Anti-Islamic Social Movement," *Terrorism and Political Violence*, vol. 26, no. 5, pp. 759–779, 2014.

[14] Y. R. Tausczik and J. W. Pennebaker, "The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods," *Journal of Language and Social Psychology*, vol. 29, no. 1, pp. 24–54, 2010. [Online]. Available: http://jls.sagepub.com

[15] E. Fast, B. Chen, and M. S. Bernstein, "Empath: Understanding topic signals in large-scale text," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, ser. CHI '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 4647–4657.

[16] W. L. Hamilton, R. Ying, and J. Leskovec, "Inductive representation learning on large graphs," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 1025–1035.

[17] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '19)*. ACM, 2019, pp. 2623–2631.

[18] V. A. Traag, L. Waltman, and N. J. van Eck, "From Louvain to Leiden: guaranteeing well-connected communities," *Scientific Reports*, vol. 9, no. 1, 2019.

[19] I. van der Vegt, M. Mozes, B. Kleinberg *et al.*, "The grievance dictionary: Understanding threatening language use," *Behavior Research Methods*, vol. 53, pp. 2105–2119, 2021. [Online]. Available: https://doi.org/10.3758/s13428-021-01536-2

[20] J. Ebner, "Measuring socio-psychological drivers of extreme violence in online terrorist manifestos: An alternative linguistic risk assessment model," *Journal of Policing, Intelligence and Counter Terrorism*, vol. 19, no. 2, pp. 125–143, 2023.

[21] A. Siggery, D. Hunt, and C. Tzani, "Language profile of lone actor terrorist manifestos: A mixed methods analysis," *Behavioral Sciences of Terrorism and Political Aggression*, vol. 15, no. 3, pp. 390–408, 2023.