Early prediction of radicalisation in online extremist communities

Anonymous ACL submission

Abstract

This study investigates early indicators of radicalisation within online extremist communities. Building upon counterterrorism research, we identify and analyse three sociolinguistic markers of radicalisation: hostility, longevity and social connectivity. We develop models to predict the maximum degree of each indicator measured over an individual's lifetime, based on a minimal number of initial interactions. Drawing on data from two diverse extremist communities, our results demonstrate that NLP methods are effective at prioritising at-risk users. This work offers practical insights for intervention and policy development, and highlights an important but under-studied research direction.

1 Introduction

Online extremism is a pressing problem with a proven relation to not only indirect societal harm (Blake et al., 2021; Roberts-Ingleson and McCann, 2023) but also to concrete offline dangers in the form of terrorist activities (Gill et al., 2017; Baele et al., 2023). Though disconcerting, the growth of publicly available online content that espouses extremist views presents an opportunity to use computational methods for detecting, channelling, and combating extremist behaviour.

Despite the significance of language to this issue, there has been limited NLP research on extremism and radicalisation. Existing work has focused on the identification of behaviours related to specific communities. For instance, de Gibert et al. (2018) introduced a dataset of hate speech on a white supremacist forum, and Hartung et al. (2017) develop a method for identifying right-wing extremist Twitter profiles. However, there is a dearth of research on the more general process of radicalisation. Yet relevant resources exist: recent studies in political science (Baele et al., 2023) and cybersecurity (Vu et al., 2021; Ribeiro et al., 2021) have developed large datasets on online extremism. They address the strongly developed in-group language and imagery using surface features such as the lexicon developed by Farrell et al. (2019).

It is widely held that individuals who become radicalised undergo a gradual cognitive shift rather than an instantaneous conversion (Munn, 2019; Beadle, 2017; Winter et al., 2020). This provides an opportunity to identify at-risk individuals for potential deradicalisation initiatives early on in the radicalisation process. In this work, we illustrate the effectiveness of NLP methods for the early prediction of three radicalisation indicators: hostile language usage, long-term engagement on an extremist platform, and connectedness within the social network. Our analysis indicates that these factors provide complementary and compelling perspectives on the radicalisation of individuals.

To formalise the work, we define our task as *predicting the maximum degree of hostility, longevity and inter-group connectivity measured over an individual's lifetime, after observing an initial subset of their interactions within the group*. Our results indicate that it is possible to prioritise at-risk users with an accuracy of 0.70 after 10 posts and 0.68 after 5 posts. Our top-performing approach is a multitask model that jointly predicts the three factors based on a combination of interaction and linguistic inputs. We evaluate our framework on data from 9 platforms from anti-women and white supremacist communities, finding that model performance is improved by integrating out-of-domain data. We further investigate the effect of the number of input posts on prediction accuracy, finding a good tradeoff between early prediction and performance is achieved after 6 posts.

2 Online radicalisation

The exact definitions of extremism and radicalisation are still debated among social science scholars, but there are some common features which

are often recognised. In this work, we follow the definition of radicalisation by Beadle (2017) as "a **process** of **gradually** adopting **extreme views and ideas**, inducing a growing willingness to directly support or engage in violent acts to solve social and political conflicts". Beadle (2017) further states that the internet facilitates radicalisation by providing individuals with **connection to communities** that reaffirm and strengthen extreme beliefs.

From these descriptions, we can identify the following behaviours that relate to online radicalisation at the individual level:

1. Using hostile language originating from a violent extremist ideology (exhibiting the adoption of **extreme views and ideas**),
2. Connecting to a network that espouses these extreme ideas (exhibiting **connection to the community**), and
3. A sustained engagement with its doctrine over time (following a **gradual process**).

Existing research has investigated some of these signals in isolation. For instance, targeted hate speech has been used to identify and sanction the promoters of various extremist ideologies (Hartung et al., 2017; Vidgen and Yasseri, 2020; Alatawi et al., 2021). Community connectedness, as measured through network features, has also been used to identify extremist accounts on Twitter (Gialampoukidis et al., 2017; Ferrara et al., 2016). In research on communities more broadly, connectedness in the social graph and the adoption of shared language have been found to be indicative of a user's social maturity in a community and their likelihood to churn (Danescu-Niculescu-Mizil et al., 2013; Rowe, 2013), as well as the user's loyalty to a particular network (Hamilton et al., 2017).

A lesser-studied component within extremism research is longevity within the community. Most definitions of radicalisation agree that it is a gradual process rather than an immediate conversion; as such, long-term, sustained interaction with an extremist community should also be considered as an indicator when characterising radicalisation. Within research on online communities, the volume of contributions by users is considered an important success metric (Iriberri and Leroy, 2009) and long-term users have been found to be pertinent to the stability of an online community (Wang et al., 2017).

In this work, we investigate the early predictors of these three radicalisation indicators: use of hostile language, connectedness in the social graph, and longevity on the platform. In Section 4, we detail how these factors are quantified. This multi-pronged approach provides a more holistic profiling of a user's behaviour and considers radicalisation as a spectrum, in contrast to the binary classification approaches proposed in other studies (eg. Ferrara et al., 2016). An exception is Hartung et al. (2017), who ranks users along a continuum to identify right-wing extremist users using a similarity-based method. However, their profiling method is specific to right-wing extremism, and they do not pursue the early detection objective. We share the forecasting objective with Ferrara et al. (2016), who predict whether users will interact with extremist accounts, but they frame this as a classification task and they do not consider linguistic features or neural models. Zhang et al. (2018) and Chang and Danescu-Niculescu-Mizil (2019) investigated the task of forecasting toxic language, but focus on the conversation progression rather than the lifecycle of the individual.

We do not consider these indicators to be exhaustive, but believe that they offer diverse and well-justified perspectives.

3 Quantifying radicalisation

We follow Gialampoukidis et al. (2017) in calculating the **betweenness centrality** as a measure for the connectedness of an individual in an extremist community. Betweenness centrality provides a measure for the importance of a node as a function of the number of shortest paths that traverse it, and is often used to identify prominent members of a community (Brandes, 2001). We construct an interaction graph[1] where each node represents a user, and an undirected edge is added between user nodes if they engage in the same conversation thread. The edges are weighted by the number of shared threads. To account for the dynamic nature of the user base, we construct the graph at monthly increments for each community and re-calculate the centrality scores for each user. An objection to this approach may be that the coarseness of aggregation might not capture rapid changes in the network; however, it ensures that our models are not overly sensitive to minor fluctuations.

[1]These forums have no notion of a follower graph, which is often used for calculating centrality.

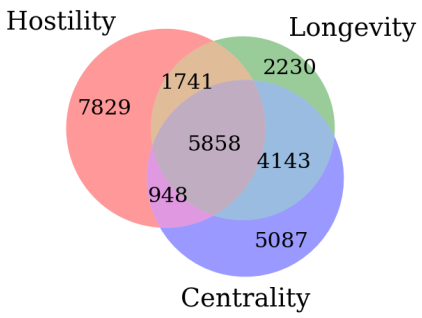

Figure 1: The intersection of the 90th percentile users of longevity, hostility and centrality.

To calculate **hostility**, we use a lexicon of in-group language associated with the violent extremist community. Extremist factions commonly define themselves through the deliberate exclusion of a specific out-group, and consequently, their internal jargon tends to be hostile towards this out-group. An alternative approach could be to consider a broader definition of hostility using pre-trained toxicity models. However, as mentioned in Section 1, these groups have a propensity for using non-standard in-group language which would not be captured by generalised toxicity models. Future work may consider automated approaches for identifying hostile in-group language.

Longevity is calculated based the number of posts produced by a user over their time on the platform, following Danescu-Niculescu-Mizil et al. (2013) and Hamilton et al. (2017), who use the number of posts to measure the maturity and loyalty of a community members, respectively. Time on the platform, in days or months, would also be a possible indicator for longevity and is generally correlated with the volume of posts. However, the latter is considered to be a more robust measure for enduring involvement in a community as it penalises intermittent and sporadic engagement.

4 Analysis

In this section, we investigate the indicators described in Section 2 using a dataset of discussions on 8 extremist anti-women forums by Ribeiro et al. (2021). The dataset consists of 7.4 million posts by 139 090 users ranging from 2005 to 2019. For each post, the author, date, thread ID and text are provided. Ribeiro et al. (2021) used this data to study the evolution of different communities over time, whereas this work focuses on the trajectories of individuals.

The forums in this dataset belong to a larger network of online communities collectively referred to as the "manosphere", which is characterised by sexual objectification of women or endorsements of violence against women. Farrell et al. (2019) and Baele et al. (2023) showed that the language used in manosphere communities is becoming increasingly extreme in nature, and at least 15 acts of real-world terrorism have been connected to this network (Latimore and Coyne, 2023). To measure hostility within this community, we use the lexicon developed by Farrell et al. (2019), consisting of 424 words and phrases. Evaluating the radicalisation indicators on this dataset, a number of conclusions can be drawn.

(i) Longevity, hostility and centrality provide complementary perspectives. Figure 1 illustrates the intersection of the 90th percentile users per indicator. To find these groups, we use the maximum value over each user's lifetime (hereafter referred to as their **eventual** value) and we calculate percentiles for each forum separately. It is evident that the sets intersect to some degree, but there is also substantial non-overlapping components. We further calculate the Spearman correlation between these factors for the full population. The strongest correlation ($\rho = 0.798$) is observed between the eventual longevity and centrality values per user, whereas the weakest correlation is between hostility and centrality ($\rho = 0.469$), and $\rho = 0.613$ for hostility and longevity. Thus, we conclude that these factors interact but that each also offers a distinct perspective, with hostility being the most disjunct.

(ii) Many users churn quickly. There is a steep drop-off in users after relatively few interactions on the forums, which aligns with the proposition by Barrelle (2010) that high turnover is characteristic of extreme groups. Figure 2 shows the survival function (Goel et al., 2010) for the number of posts per user for each forum, which illustrates the fraction of users who have more than N posts for $N \leq 10$. For half of the forums, more than 60% of their users have fewer than 5 posts in their lifetime on the platform. This may be due to users realising after further exposure to the community that the extremeness of the ideology does not resonate with them. The forum with the least churn is Incels, which could be related to the fact that many users migrated to this forum after the *r/incels* subreddit

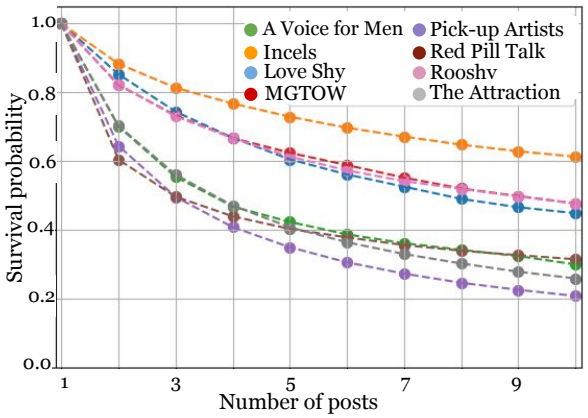

Figure 2: Survival curves for 8 manosphere forums, illustrating the likelihood of a user to continue interacting on the platform after N posts, for $N < 10$.

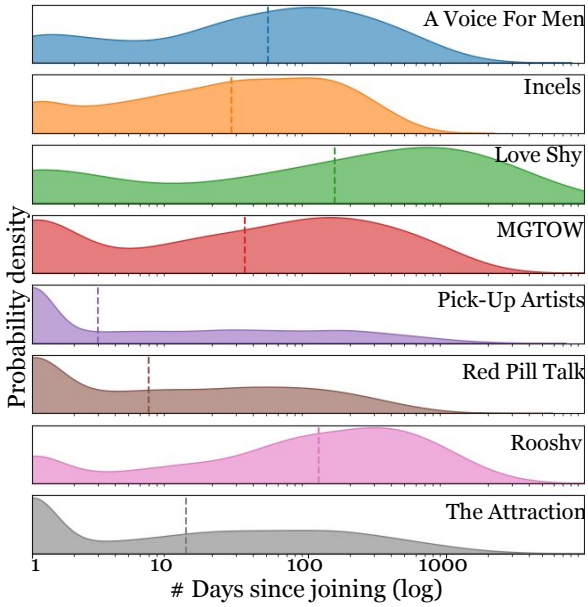

Figure 3: The number of days (logscale) for users to reach the 90th percentile of hostility, per forum.

was banned in 2017 (Hauser, 2017); as such, users would already have been inducted into the ideology before joining.

(iii) Some users start out hostile; others become hostile. The radicalisation factors vary over the course of a user's lifetime on the platform. From the positive correlation between hostility and longevity, we know that that users who are on the platform for longer reach higher levels of hostility, but how quickly does this happen? Figure 3 shows the number of days it takes for users to reach the 90th percentile of hostility. For five of the forums, a bimodal distribution is observed, with an early peak ($<$ 10 days) as well as a later peak between 100 and 1000 days. This indicates that a subset of users already exhibit these behaviours when they join the platform, whereas others develop them over time. The stage in their radicalisation process at which a user joins the platform would likely play a role in this phenomenon. This supports the social science research that states that there is no single, agreed upon pathway to radicalisation, and highlights the importance of considering multiple indicators.

The three platforms that do not exhibit this trend, having only an early peak, also had higher early churn rates (Figure 2). For the longevity and centrality factors, this bimodality is not present: only a later peak (100–1000 days) is observed.

(iv) Early signs of radicalisation. Having noted that the indicator values vary over time, we turn to the question of which early signals are predictive of eventual radicalisation. To do this, we calculate the following features for the first 10 user interactions for users with 10 or more posts:

- **Post length**: the median character count per post,
- **Hostility terms**: the median number of terms used from the Farrell et al. (2019) lexicon per post,
- **Number of threads** in which a user engaged,
- **Time between posts**: the median number of hours between posts, and
- **Days engaged**: number of distinct days on which the user engaged on the platform.

We calculate the Spearman correlation of the eventual indicator values with the above feature values after 10 interactions. The results, in Table 1, show that these early behaviours are correlated to varying degrees with each of the indicators. The strongest correlation to all three indicators is given by the number of distinct days a user engaged on the platform through their first 10 posts. A possible explanation is that a user who comes back repeatedly on separate occasions indicates a higher level of interest and receptiveness, compared to one who posts a larger volume of posts at once, and then disconnects for several days. The largest correlation is to eventual longevity, which aligns with our expectation that longevity is tied to loyalty (Hamilton et al., 2017). Linguistic features (post length and hostility terms) are correlated to eventual hostility, but not to eventual centrality and longevity. The number of threads in which a user engaged is

Feature	Centr.	Host.	Long.
Post length	-0.041	0.380	-0.006
Hostility terms	0.00	0.317	0.023
# threads	0.243	-0.066	0.115
Time between posts	-0.184	-0.014	-0.134
# days engaged	0.470	0.468	0.748

Table 1: The Spearman correlation between features of the first 10 posts by a user and eventual indicator levels.

correlated to eventual centrality, but not to hostility and weakly correlated to longevity. This shows that there are early signs of each of the three indicators that are not correlated to the others, providing further support for our multi-indicator approach. The the time between posts has a slight negative correlation to centrality and longevity, meaning that more frequent engagements are positively correlated to these indicators.

In the remainder of this paper, we investigate how accurately the three indicators can be predicted based on the early behaviour of a user.

5 Early prediction of radicalisation

We define the task of predicting a user's maximum lifetime score on the three radicalisation indicators after observing an initial subset of N posts by that user, with $N \in \{5, 10\}$. We choose these values of N based on the survival curves (Fig. 2), which indicate a substantial drop-off in users with fewer than 5 posts and a stabilisation after $N = 10$. Earlier detection is better, but models do require sufficiently strong signals which may not be present if the information is too limited. Since these indicators take on real-valued numbers, this is a regression task.

5.1 Metrics

We use two metrics to compare performance on this task. Since an aim of this work is to prioritise users for deradicalisation initiatives, the ordering of users is of interest. To measure this, we report the **concordance index** (**CI**, Harrell et al., 1982). A pair of observations i, j is considered concordant if the prediction and the ground truth have the same inequality relation, i.e. $(y_i > y_j, \hat{y}_i > \hat{y}_j)$ or $(y_i < y_j, \hat{y}_i < \hat{y}_j)$. The concordance index is the fraction of concordant pairs in the test set. A random model would achieve a CI of 0.5 and a perfect score is 1. We also report the mean absolute error (**MAE**) for each indicator. MAE is widely used in regression studies as it provides an intuitive measure for numerical accuracy. However, it is susceptible to outliers and could not be compared between factors, since they operate on different numeric scales. Consequently, we rely on the CI for model selection. Significance testing is performed with the two-sided randomised permutation test, using Monte Carlo approximation with $R = 9999$.

5.2 Data

We use the Ribeiro et al. (2021) manosphere dataset, described in Section 4, in this evaluation. We filter entries with missing dates, texts, authors or thread IDs and remove users with fewer than 10 interactions. The resulting dataset contains 7.1 million posts by 39 765 users. The median post length is 33 tokens and the median number of posts per user is 30. The labels are given by the indicator definitions as provided in Section 4 and we release our labels to the community. Since the distributions are heavy-tailed, we truncate the indicator values beyond the 95th percentile of each indicator per forum. We split the data into a training, test and development set with a ratio of 75:15:10.

5.3 Methods

Our objective in these experiments is to develop quantitative methods for the early prediction of radicalisation indicators. We therefore experiment with various input and auxiliary task combinations to determine what type of information is useful to model the these indicators.

Feature-based models We use the features described in Section 4 as a baseline, evaluating models with and without glossary features to investigate the effect of adding linguistic information. For the glossary features, we use the mean and maximum of number of glossary terms per post. The feature and indicator values are normalised using min-max scaling. The model architecture consists of a multilayer perceptron (MLP) with two hidden layers. Three separate models are trained to predict each indicator value independently. Hyperparameters and training details are provided in Appendix A.

Text-based models Models that operate directly upon the post text, as opposed to engineered features, are expected to capture more nuanced features that extend beyond the hostility lexicon and post length. We use the pretrained `all-mpnet-base-v2`[2] sentence transformer (Reimers and Gurevych, 2019) to obtain an

[2] https://huggingface.co/sentence-transformers/all-mpnet-base-v2.

embedding of length 768 for each post. The model architecture consists of an LSTM layer (Hochreiter and Schmidhuber, 1997) followed by two hidden layers. Since the embeddings are produced by a large pretrained language model, we expect that a relatively small number of layers should be sufficient to finetune them to our task.

Mixed-input models A dual-input architecture is used to combine the text-level learning from embeddings with the engineered interaction and glossary-based features. The glossary-based features capture the use of non-standard in-group terms which may not appear in the vocabulary of a pretrained language model; as such, both types of linguistic inputs may be useful. An LSTM layer and two MLP layers are used to process the text and feature inputs in parallel. The outputs are concatenated and two further hidden layers are applied.

Multitask models The analysis in Section 4 indicated that the different indicators interact and correlate to some extent. As such, we expect that parameter sharing might be beneficial, as opposed to training a separate model for each indicator. We keep the same initial architecture as in the mixed input models, but use a separate prediction head with two additional hidden layers for each output.

Our dataset consists of user profiles from 8 platforms, which may have distinct user-level characteristics. To investigate whether there are useful features that are tied to the different platforms, we further experiment with predicting the forum from which the sample originates as an auxiliary task.

Survival regression For time-to-event prediction from text inputs, such as the longevity prediction task, survival regression has been illustrated to outperform traditional regression approaches (De Kock and Vlachos, 2021). This framework has a more explicit treatment of time and events within a standard regression setting, and is particularly effective for modelling real-valued, exponentially-distributed outcomes. We use the logistic hazard model (Gensheimer and Narasimhan, 2019) for the longevity predictions. Using this framework, we can retain the same neural architectures, but modify the objective to predict the probability of churn for an individual within each timestep, given survival up to that point (also known as the hazard). The outputs are transformed into 100 equidistant timesteps, and the loss is the negative log likelihood of the predicted versus actual hazard per timestep.

6 Results

Our results are shown in Table 2. Significance of improvements in CI ($P \leq 0.05$) as compared to the model directly above is indicated by asterisks. The CI scores for the three indicators are in a relatively close range to one another for most models. The top-performing model has a CI of 0.667 for centrality, 0.698 for hostility and 0.681 for longevity (at $N = 10$), constituting a statistically significant improvement over baselines of respectively +1%, +6.3% and +7.9%. For all models and indicators, the performance at $N = 5$ is worse than at $N = 10$. Of the three indicators, centrality has the largest increase in CI between $N = 5$ and $N = 10$. The MAE values generally follow the CIs in terms of direction of improvement.

Adding sources of information or auxiliary tasks tends to improve performance in our experiments. Using glossary-based features in addition to interaction-based features improves CI (significant for 4 out of 6 cases), which supports our central hypothesis that linguistic cues can be helpful at foreshadowing radicalisation. Using only post embeddings outperforms feature-based approaches for hostility and longevity prediction, but reduces the CI for centrality. Combining features and embeddings improves the CI over embedding-only models (significant for 3 out of 6 cases), indicating that the features contain useful information beyond what is captured by the language model. Joint training of the three indicators yields a further improvement, particularly in MAE, which aligns with expectation that the three factors contain mutually informative signals. Marginal improvements, significant in 2 cases, are made by adding the forum prediction auxiliary task. The experiments in the remainder of this section use this model.

The performance of the feature-based centrality model declined when the text embeddings were added, and although the highest score for this indicator was achieved by the multifactor model which uses embeddings, this improvement was smaller than for the other indicators. Considering that the analysis in Table 1 showed no correlation between the early use of hostility terms and eventual centrality, this is perhaps not surprising. We can conclude that the language features and models used in this study are less apt at detecting the early cues that foreshadow centrality, if they are present.

Model	Centrality		Hostility		Longevity	
	CI ↑	MAE ↓	CI ↑	MAE ↓	CI ↑	MAE ↓
$N = 5$						
Interaction features	0.620	0.380	0.616	7.150	0.561	49.43
Interaction + glossary features	0.621	0.388	0.640*	7.258	0.572*	50.46
Transformer embeddings	0.595	0.376	0.658*	7.628	0.647*	46.33
+ all features	0.608*	0.381	0.666	7.754	0.652	46.55
+ multifactor training	**0.622***	0.315	0.672	5.730	0.645	**45.18**
+ forum aux. task	0.621	**0.314**	**0.677**	**5.737**	**0.656***	45.675
$N = 10$						
Interaction features	0.657	0.388	0.635	7.279	0.602	48.15
Interaction + glossary features	0.659	0.390	0.665*	7.341	0.615*	47.59
Transformer embeddings	0.616	0.382	0.679*	7.749	0.654*	45.12
+ all features	0.651*	0.393	0.689	7.956	0.677*	44.40
+ multifactor training	0.666*	**0.287**	0.693	**5.527**	0.672	43.56
+ forum aux. task	**0.667**	0.288	**0.698**	5.538	**0.681***	**43.24**

Table 2: Results for predicting the eventual centrality, hostility and longevity values at $N = 5$ and $N = 10$. Arrows indicate the preferred directions per metric and best models per indicator and metric are shown in bold.

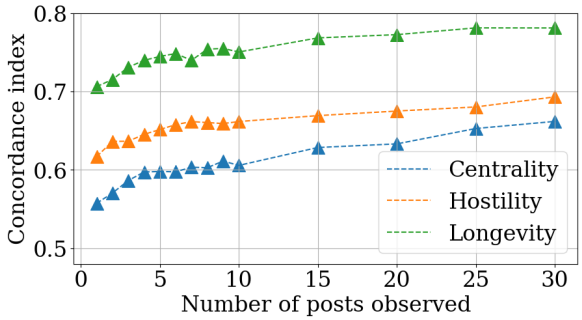

Figure 4: Performance at different N.

	2	3	4	5	6	7	8	9	10
1	**.029**	.037	0	0	0	0	0	0	0
2	-	.994	.325	.093	**.02**	.004	.01	.013	.009
3	-	-	.323	.098	.017	.006	.005	.007	.006
4	-	-	-	.475	.167	.078	.105	.119	.082
5	-	-	-	-	.47	.276	.316	.419	.268
6	-	-	-	-	-	.65	.755	.86	.652
7	-	-	-	-	-	-	.907	.791	.988
8	-	-	-	-	-	-	-	.88	.875
9	-	-	-	-	-	-	-	-	.767
10	-	-	-	-	-	-	-	-	-

Table 3: Significance of performance increases with larger N for the hostility indicator.

6.1 Optimising the number of inputs

Our aim in this work is the early identification of users who are at risk of radicalisation. In this section, we consider *how early* such a prediction might be made. Given the tradeoff between prioritising performance versus earlier prediction, the optimal prediction point will be where improvement starts to saturate as N increases. To find this, we train models with inputs ranging from 1 to 30 posts, sampling more densely at $N < 10$ as larger improvements are expected.

The results are shown in Figure 4. Only users with 30 or more posts are included in this experiment, so the CI values cannot be directly compared to the results in Table 2. For all three indicators, there is an upward trend in CI as N increases, with a steeper increase for $N < 5$ and a more moderate improvement for $5 < N \leq 10$. Beyond $N = 10$, diminishing returns are observed for the longevity and hostility indicators, meaning that delaying the prediction beyond this point is not well-justified. It is worth noting that centrality still improves substantially beyond this point.

We are interested in the minimum improvement in N which would constitute a significant improvement in CI. We use randomised permutation testing to evaluate the significance of the improvement at each step for $N < 10$. The P-values for hostility are shown in Table 3, with significance ($P \leq 0.05$) indicated in green. A significant improvement ($P = 0.029$, shown in bold) is observed between 1 and 2 inputs. From 2, we would need to increase the number of inputs to 6 to obtain a significant improvement ($P = 0.02$). No further significant improvements are possible in the observed range. For centrality and longevity, following a similar procedure yields significant improvements until $N = 8$ and $N = 6$, respectively. As such, we recommend using the initial 6 posts made by a user to predict radicalisation as early as possible with a good tradeoff in accuracy.

6.2 Out-of-domain evaluation

This paper is concerned with radicalisation as a general concept, and not only its specific manifestation in the manosphere. As such, we also evaluate

	Manosphere			Stormfront		
Training data	**Cent**	**Host**	**Long**	**Cent**	**Host**	**Long**
Manosphere	0.666	0.693	0.672	0.592	0.660	0.584
Stormfront	–	–	–	0.635*	0.682*	**0.603***
Combined	0.662	0.689	0.667	0.635	0.705*	0.590
Combined + forum aux. task	**0.668**	**0.699**	**0.675**	**0.640***	**0.721***	0.598

Table 4: Concordance index of multifactor models for the Manosphere and Stormfront datasets.

our framework on the white supremacy platform Stormfront, using the ExtremeBB dataset (Vu et al., 2021).

Applying the same filters as in Section 5.2, we obtain a dataset of posts by 25 895 users. The centrality and longevity indicators are calculated as described in Section 3. The hostility indicator is intended to capture the adoption of extreme ideas from the community in question, which we operationalise using a lexicon. A list of 293 alt-right phrases and symbols was scraped from RationalWiki[3] and is shared with the community. The labels for this dataset cannot be shared under the ExtremeBB data agreement.

We expect to see differences in the numeric values of the indicators as their distributions will differ between the populations. This is accounted for in our framework by *(i)* applying min-max scaling to the indicator values during training, and *(ii)* using the CI metric for evaluation, which is concerned with relative ordering rather than absolute values.

We evaluate a number of different training configurations, with CI values at $N = 10$ shown in Table 4. Using the best model as trained on manosphere data, lower CI values are recorded for all three indicators compared to the original dataset. Training on the Stormfront dataset instead improves the scores for all three indicators on the same data (significant at the $\alpha = 0.05$ level). Training on both datasets increases the CI for the hostility prediction on Stormfront but reduces the CI for all others. However, when the forum prediction auxiliary task is included, there is a statistically significant improvement on the centrality and hostility metrics on the Stormfront data.

In conclusion, a drop in model performance is to be expected if a model trained on data from one extremist community is transferred to a different community without any adjustment. However, joint training on unrelated communities is useful if the platform information is provided in the form of an auxiliary task. Future work may explore training on larger multi-community datasets, or pretraining and finetuning configurations.

[3] https://rationalwiki.org/wiki/Alt-right_glossary

7 Conclusion

We have proposed a framework for quantifying radicalisation based on sociolinguistic indicators. We investigated the interaction of these indicators using a dataset of posts on extremist platforms and identified early signals that correspond to the eventual radicalisation of an individual. We then developed and evaluated models that can preemptively identify users who are at risk of radicalisation.

In contrast to prior work, our approach does not require specialist annotation, which is resource-intensive and susceptible to annotator biases. By framing it as a regression task, we avoid the need to make a binary decision on edge cases, allowing for a framework that more closely resembles the spectrum of behaviours that are observed in these communities. Our approach is not tied to any specific extremist movement as it relies on more general characteristics of networks and groups. By considering multiple indicators, we can elicit a more holistic perspective that captures different types of signals that may indicate radicalisation.

A comprehensive understanding of radicalisation requires inputs from several disciplines to capture the various contributing factors, including the psychological, educational, economic, and social-adjustment parameters of individuals. However, capturing these factors and merging them into a single predictive model is not feasible within the current data landscape. Using language as a proxy for some of these parameters, identifying the features most predictive of radicalisation, and quantifying them using NLP is a promising methodology. We look forward to addressing more of these parameters in work across relevant disciplines.

8 Limitations

We hope that this work will serve as a foundation for further NLP work in this direction, which may address some of the following limitations.

The hostility indicator is reliant on a lexicon. Linguistic resources have been developed for many online extremist communities, but using manually constructed lexicons is sub-optimal as they are bound to have imperfect recall and they are constructed for the community at a particular point in time, which ignores the fact that community language is highly dynamic.

The centrality indicator is intended to capture social connectedness and is a well-established metric for this purpose. However, extremist groups are known to be prone *splintering*, a process whereby the more extreme community members form subgroups with limited interaction with the larger community. This behaviour is highly indicative of radicalisation but is not captured by the centrality indicator.

The longevity metric assumes that users who churn early, do so because they are disengaging from the group. It is also plausible that some users may leave a community to seek out more extreme groups. However, since early churn is commonly observed in all extreme groups (Barrelle, 2010), we assume that the former explanation holds true for the majority of users.

Finally, our work builds on prior research in online communities. More consideration could be devoted to the characteristics that differentiate extreme communities from online communities more broadly.

References

Hind S. Alatawi, Areej M. Alhothali, and Kawthar M. Moria. 2021. Detecting white supremacist hate speech using domain specific word embedding with deep learning and bert. *IEEE Access*, 9:106363–106374.

Stephane Baele, Lewys Brace, and Debbie Ging. 2023. A diachronic cross-platforms analysis of violent extremist language in the incel online ecosystem. *Terrorism and Political Violence*, pages 1–24.

Kate Barrelle. 2010. Disengagement from violent extremism. In *Conference paper. Monash University: Global Terrorism Research Centre and Politics Department*.

S Beadle. 2017. How does the internet facilitate radicalization?, war studies department, king's college london, essay 2.

Khandis R Blake, Siobhan M O'Dean, James Lian, and Thomas F Denson. 2021. Misogynistic tweets correlate with violence against women. *Psychological science*, 32(3):315–325.

Ulrik Brandes. 2001. A faster algorithm for betweenness centrality. *Journal of mathematical sociology*, 25(2):163–177.

Jonathan P Chang and Cristian Danescu-Niculescu-Mizil. 2019. Trouble on the horizon: Forecasting the derailment of online conversations as they develop. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4743–4754.

Cristian Danescu-Niculescu-Mizil, Robert West, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. No country for old members: User lifecycle and linguistic change in online communities. In *Proceedings of the 22nd international conference on World Wide Web*, pages 307–318.

Ona de Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. 2018. Hate speech dataset from a white supremacy forum. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 11–20, Brussels, Belgium. Association for Computational Linguistics.

Christine De Kock and Andreas Vlachos. 2021. Survival text regression for time-to-event prediction in conversations. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1219–1229, Online. Association for Computational Linguistics.

Tracie Farrell, Miriam Fernandez, Jakub Novotny, and Harith Alani. 2019. Exploring misogyny across the manosphere in reddit. In *Proceedings of the 10th ACM conference on web science*, pages 87–96.

Emilio Ferrara, Wen-Qiang Wang, Onur Varol, Alessandro Flammini, and Aram Galstyan. 2016. Predicting online extremism, content adopters, and interaction reciprocity. In *Social Informatics: 8th International Conference, SocInfo 2016, Bellevue, WA, USA, November 11-14, 2016, Proceedings, Part II 8*, pages 22–39. Springer.

Michael F Gensheimer and Balasubramanian Narasimhan. 2019. A scalable discrete-time survival model for neural networks. *PeerJ*, 7:e6257.

Ilias Gialampoukidis, George Kalpakis, Theodora Tsikrika, Symeon Papadopoulos, Stefanos Vrochidis, and Ioannis Kompatsiaris. 2017. Detection of terrorism-related twitter communities using centrality scores. In *Proceedings of the 2nd International Workshop on Multimedia Forensics and Security*, MFSec '17, page 21–25, New York, NY, USA. Association for Computing Machinery.

Paul Gill, Emily Corner, Maura Conway, Amy Thornton, Mia Bloom, and John Horgan. 2017. Terrorist use of the internet by the numbers: Quantifying behaviors, patterns, and processes. *Criminology & Public Policy*, 16(1):99–117.

Manish Kumar Goel, Pardeep Khanna, and Jugal Kishore. 2010. Understanding survival analysis: Kaplan-meier estimate. *International journal of Ayurveda research*, 1(4):274.

William Hamilton, Justine Zhang, Cristian Danescu-Niculescu-Mizil, Dan Jurafsky, and Jure Leskovec. 2017. Loyalty in online communities. In *Proceedings of the International AAAI conference on web and social media*, volume 11, pages 540–543.

Frank E Harrell, Robert M Califf, David B Pryor, Kerry L Lee, and Robert A Rosati. 1982. Evaluating the yield of medical tests. *Jama*, 247(18):2543–2546.

Matthias Hartung, Roman Klinger, Franziska Schmidtke, and Lars Vogel. 2017. Ranking right-wing extremist social media profiles by similarity to democratic and extremist groups. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 24–33, Copenhagen, Denmark. Association for Computational Linguistics.

Christine Hauser. 2017. Reddit bans 'incel' group for inciting violence against women. New York Times. Accessed 10-11-2023.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Alicia Iriberri and Gondy Leroy. 2009. A life-cycle perspective on online community success. *ACM Computing Surveys (CSUR)*, 41(2):1–29.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Jasmine Latimore and John Coyne. 2023. Incels in australia: the ideology, the threat, and a way forward.

Luke Munn. 2019. Alt-right pipeline: Individual journeys to extremism online. *First Monday*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992.

Manoel Horta Ribeiro, Jeremy Blackburn, Barry Bradlyn, Emiliano De Cristofaro, Gianluca Stringhini, Summer Long, Stephanie Greenberg, and Savvas Zannettou. 2021. The evolution of the manosphere across the web. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pages 196–207.

Elise M Roberts-Ingleson and Wesley S McCann. 2023. The link between misinformation and radicalisation. *Perspectives on Terrorism*, 17(1):36–49.

Matthew Rowe. 2013. Mining user lifecycles from online community platforms and their application to churn prediction. In *2013 IEEE 13th International Conference on Data Mining*, pages 637–646.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

Bertie Vidgen and Taha Yasseri. 2020. Detecting weak and strong islamophobic hate speech on social media. *Journal of Information Technology & Politics*, 17(1):66–78.

Anh V Vu, Lydia Wilson, Yi Ting Chua, Ilia Shumailov, and Ross Anderson. 2021. Extremebb: Enabling large-scale research into extremism, the manosphere and their correlation by online forum data. *arXiv preprint arXiv:2111.04479*.

Xi Wang, Kang Zhao, Nick Street, et al. 2017. Analyzing and predicting user participations in online health communities: a social support perspective. *Journal of medical Internet research*, 19(4):e6834.

Charlie Winter, Peter Neumann, Alexander Meleagrou-Hitchens, Magnus Ranstorp, Lorenzo Vidino, and Johanna Fürst. 2020. Online extremism: research trends in internet activism, radicalization, and counter-strategies. *International Journal of Conflict and Violence (IJCV)*, 14:1–20.

Justine Zhang, Jonathan Chang, Cristian Danescu-Niculescu-Mizil, Lucas Dixon, Yiqing Hua, Dario Taraborelli, and Nithum Thain. 2018. Conversations gone awry: Detecting early signs of conversational failure. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1350–1361.

A Training specifications

In all experiments, we use a batch size of 32 and ReLU activation functions between hidden layers. We train with early stopping with a patience of 20 epochs. Models are developed in `PyTorch`. We use a gridsearch to determine the best hyperparameter values, experimenting with hidden layer sizes in $\{32, 64, 128\}$ and dropout (Srivastava et al., 2014) with $p \in \{0.1, 0.2, 0.5\}$. The Adam (Kingma and Ba, 2014) optimiser is used, with $\eta \in \{1e-4, 5e-4, 1e-3\}$. The best value per model are reported in Tables 5.

Model	**Factor**	**Dropout (p)**	**Hidden units per layer**	**Learning rate**
$N = 5$				
Frequency features	Centrality	0.1	32	0.0005
	Hostility	0.2	32	0.0005
	Longevity	0.2	128	0.0005
Frequency + glossary features	Centrality	0.1	64	0.0005
	Hostility	0.1	32	0.0005
	Longevity	0.1	64	0.0005
Embeddings	Centrality	0.1	32	0.0001
	Hostility	0.1	64	0.0001
	Longevity	0.2	128	0.0005
Embeddings + features	Centrality	0.1	64	0.0005
	Hostility	0.1	32	0.0001
	Longevity	0.1	64	0.0005
Multifactor	All	0.1	64	0.0005
+ forum aux.task	All	0.1	128	0.0005
$N = 10$				
Frequency features	Centrality	0.1	128	0.0005
	Hostility	0.1	32	0.0005
	Longevity	0.2	128	0.0005
Frequency + glossary features	Centrality	0.1	32	0.0005
	Hostility	0.1	128	0.0005
	Longevity	0.1	32	0.0005
Embeddings	Centrality	0.1	32	0.0001
	Hostility	0.2	64	0.0001
	Longevity	0.1	128	0.0005
Embeddings + features	Centrality	0.1	32	0.0001
	Hostility	0.2	64	0.0001
	Longevity	0.1	128	0.0005
Multifactor	All	0.1	128	0.0005
+ forum aux.task	All	0.1	128	0.0001

Table 5: Hyperparameters for per-factor models.