Inducing lexicons of in-group language with socio-temporal context

Christine de Kock
University of Melbourne
Computing and Information Systems
christine.dekock@unimelb.edu.au

Abstract

In-group language is an important signifier of group dynamics. This paper proposes a novel method for inducing lexicons of in-group language, which incorporates its socio-temporal context. Existing methods for lexicon induction do not capture the evolving nature of in-group language, nor the social structure of the community. Using dynamic word and user embeddings trained on conversations from online anti-women communities, our approach outperforms prior methods for lexicon induction. We develop a test set for the task of lexicon induction and a new lexicon of manosphere language, validated by human experts, which quantifies the relevance of each term to a specific sub-community at a given point in time. Finally, we present novel insights on in-group language which illustrate the utility of this approach.

1 Introduction

Social groups have a tendency to develop distinctive vernaculars, also known as *in-group language*. According to Drake (1980), in-group language plays two important social roles within groups: first, it can obscure the discussions of the group from out-group onlookers or moderators, and second, it signals cohesion and solidarity with a group by illustrating an awareness and acceptance of their norms. Since this type of informal language evolves rapidly (Stewart and Eisenstein, 2018), the use of recent lexical innovations of the group is a strong signal of group belonging (Danescu-Niculescu-Mizil et al., 2013; Zhang et al., 2017).

Due to its strong relation to implicit social dynamics, lexicons of in-group language are frequently used to study communities (e.g. Baele et al., 2023; Rowe and Saif, 2016). Studies that use hand-compiled lexicons are dependent on the skill of the lexicon builders (Havaldar et al., 2024). In the case of in-group language, there are subtle signals that are challenging to identify as an out-group member (Mendelsohn et al., 2023), especially given the scale of modern text corpora. Since manual lexicon building is expensive, there is a tendency not to update lexicons frequently. For example, Rowe and Saif (2016) include a 7-year-old lexicon in their study on online pro-ISIS language. These factors impact the quality of lexicons, with resulting negative effect on the quality of downstream studies.

Prior work in NLP has proposed computational methods for inducing lexicons of in-group language, often framing the task as identifying relevant words using large text corpora from the group in question. For example, Lucy and Bamman (2021) use statistical properties of word occurrences and contextualised word embeddings. Similarly, Farrell et al. (2020) and Yoder et al. (2023) use approaches based on topic models and static word embeddings. However, existing methods for computationally identifying in-group language tend to focus only on linguistic information, despite the highly dynamic and social nature of groups and their language.

In this work, we propose and evaluate methods for **L**exicon **I**nduction with **S**ocio-**T**emporal **N**uance (**LISTN**) using dynamic word- and user embeddings. This framework provides a natural way to integrate statistical properties of the text with the social structure of the community, by modelling users and words in the same space. Our experiments indicate that LISTN-based methods outperform a range of baselines based on prior work, achieving an average precision score of 0.77 on a newly crafted test set for lexicon induction[1].

Our study centers on *the manosphere*, a growing movement of anti-women online communities which has been described as a violent extremist ideology (Baele et al., 2023) and has been associated several real-world acts of terror against

---

[1]Github: github.com/unimelb-nlp/listn

women (Latimore and Coyne, 2023). These communities are exceptionally linguistically innovative (Bogetić, 2023), making them a particularly appropriate application area for this work, in addition to being a pressing current concern. The manosphere is a highly fractured collective, with well-developed and continuously evolving subgroup structure (Ribeiro et al., 2021), which also highlights the need for incorporating social and temporal dimensions for lexicon induction.

To evaluate our lexicon induction methodology, we use existing lexicons of manosphere language and human expert validation to construct a novel test set. We further release a lexicon of 455 new manosphere terms with scores representing their relevance to specific manosphere subcommunities. Using this lexicon, we find that word embeddings of in-group terms tend to be more static (i.e. showing less variation over time) compared to words of similar frequencies in the general vocabulary. Finally, we look at linguistic specialisation within subgroups of the manosphere, finding that Incels are the most distinctive in their vernacular. These analyses illustrate the utility of our approach for studying in-group language as a phenomenon, and for drawing out latent group dynamics as signalled through language.

2   Related work

A number of prior works have investigated in-group language, its evolution over time, and its relation to social dynamics. Chancellor et al. (2016) investigate linguistic variation and the processes by which existing words are modified to create neologisms in the context of pro-anorexia groups, finding that lexical innovations are associated with increased engagement and more extreme content. Danescu-Niculescu-Mizil et al. (2013) explored the role of linguistic evolution in groups, using snapshot n-gram language models to measure the deviation of a user from the community. They illustrate that the non-adoption of lexical innovations is predictive of user churn. Zhang et al. (2017) characterises how distinctive and temporally dynamic online communities are in their language, finding that communities with more dynamic and distinctive language are likely to have higher retention of members. Stewart and Eisenstein (2018) and Del Tredici and Fernández (2018) investigate factors that influence the dissemination of new lexical innovations, finding that linguistic and social fac-

tors both play a role in this process. These studies illustrate the intimate relationship between social dynamics and language, and the dynamic nature of in-group language and groups. Our work uses a holistic approach for incorporating these different effects to induce lexicons of in-group language.

Language cues have been used to investigate extremist movements; for instance, by Baele et al. (2023) and De Kock and Hovy (2024) in the context of the manosphere, and by Rowe and Saif (2016) to study ISIS radicalisation. However, these works rely on small, manually constructed lexicons. Through this work, we hope to enable future research on this important topic by producing lexicons with a strong empirical justification, grounded in the socio-temporal context of the community being studied.

Our methods leverage dynamic, jointly-learned word and user embeddings (described in Section 3). The representation learning methodology was introduced in our earlier study, De Kock (2024), to forecast what communities a user will interact with, to characterise the level of violent language within sub-groups of the manosphere, and to measure changes in the relevance of temporal concepts (e.g. *#MeToo*) to the community. Here, we reimplement and optimise the methodology (described in Section 3.2 but use the same training data. We then propose and compare 6 different methods for using these embeddings to induce lexicons.

3   Representation learning

The Cerberus architecture (De Kock, 2024) performs a shared temporal matrix factorisation over **content and adjacency** matrices, which are user-content and user-user representations. A **content-only** model variation decomposes user-content matrices over time to yield dynamic word embeddings.

Content-only   At each timestep t, a user-content matrix C_t is constructed using positive pointwise mutual information (PPMI), which quantifies how strongly the presence of user i and word j are associated beyond what would be expected if they were independent (based on a background corpus):

$$\text{PMI}(i, j) = \log \frac{P(j|i)}{P(j)}, \quad (1)$$

$$\text{PPMI}(i, j) = \max\big(\text{PMI}(i, j), 0\big). \quad (2)$$

$P(j)$ calculated is based on word frequencies in a background corpus of 40 million Reddit posts by

Dziri et al. (2019). Matrix factorisation is used to learn user representations U_t and word representations W_t such that $C_t \approx U_t \cdot W_t^T$, for all t.

Content and adjacency To represent interpersonal dynamics, a user-user adjacency matrix A_t is constructed to capture how frequently two users interact in the same thread at time t, with L1 normalisation. This matrix is jointly decomposed with C_t to yield U_t, V_t and W_t such that $C_t \approx U_t \cdot W_t^T$ and $A_T \approx U_t \cdot V_t^T$, where V_t is the context user embedding matrix.

To ensure alignment of embeddings across timesteps, the learning objective includes a regularisation term that penalises large inter-timestep changes. The joint factorisation was found to outperform the content-only approach in De Kock (2024).

3.1 Data

The models are trained on over 4 million utterances from 50 manosphere subreddits over 9 months, using the Reddit portion of the corpus of Ribeiro et al. (2021). The subreddits in the dataset are mapped to five different manosphere sub-communities, referred to as **categories** in this paper: Incels, Men's Rights Activists (**MRA**), Men Going Their Own Way **MGTOW**, Pick-up Artists (**PuA**), and The Red Pill (**TRP**). The dataset further includes data from two anti-manosphere communities (referred to as **Criticism**) and three **Mental Health** related communities, which we use as a control group in this study. Embeddings are produced in monthly windows, representing April-December 2018.

The data is filtered to include only users who are active for 3 or more timesteps and words that are used by more than 20 users, yielding 33,880 users and 44,679 words. The NLTK word tokeniser is used for tokenisation. Words are lowercased and non-alphabetic characters are removed.

3.2 Implementation

The models used in De Kock (2024) were implemented in TensorFlow to match the evaluation setting of antecedent systems, and were not optimised for scaling to large datasets. We re-implement[2] these models in PyTorch, using the generalised matrix factorisation formulation (He et al., 2017) which (1) avoids reading the full sparse matrices into memory, (2) enables batch-wise updates for faster convergence, and (3) allows for increased

---

[2]Github: `github.com/unimelb-nlp/listn`

flexibility to deploy other models for representation learning.

In this formulation, a training step if performed over a batch of users for a given timestep, rather than updating only once all the data has been seen. During the forward pass, interaction vectors are calculated against the full set of context users and words for that timestep. The model parameters are updated after each tuple of (*user-user*, *user-word*) batches. Hyperparameters are tuned based on the validation loss and are shown in Table 4 in Appendix A.

4 LISTN

Using the models described in Section 3, we evaluate a number of approaches for inducing lexicons. We refer to these methods as **LISTN** (**L**exicon **I**nduction with **S**ocio-**T**emporal **N**uance) and distinguish between **LISTN-CA** (jointly training on **C**ontent and **A**djacency matrices) and **LISTN-C** (using only the **C**ontent matrix).

Recall that the user-content matrix C is constructed using PMI to compare how often a user uses a word compared to a large background corpus. We evaluate several methods for lexicon induction that are based on the low-rank reconstruction $\hat{C}_t$. The rationale is that we aim to find words that are used more often than expected (as indicated by PMI), using embeddings to provide a more nuanced view of a word or user compared to word frequencies.

Given a user embedding $u_{i,t}$ and a word embedding $w_{j,t}$, we calculate the *relevance* of word j to user i at time t as $r(i,j,t) = u_{i,t} \cdot w_{j,t}^T$ (as also done in De Kock, 2024). This can be generalised to groups of users by finding the group centroid as the mean of the individual user embeddings, denoted $\overline{u}_t$. Farrell et al. (2020) find that the usage of in-group terms varies between the different manosphere categories. Therefore, we explore several methods that allow for sub-group specialisation. Each of the below methods produce a (scalar) score r representing the relevance of a word j within the manosphere ecosystem at time t.

Community centroid: We calculate word relevance scores as $r_{comm}(j,t) = \overline{u}_t \cdot w_{j,t}^T$, where $\overline{u}_t$ represents the centroid of all users in the training set at time t.

Category: We calculate a centroid $\overline{u}_{k,t}$ for each of the manosphere categories (as described

in Section 3), defined as the mean embedding of all users who contributed to category k at timestep t. For every word, we then find its maximum relevance score across all categories: $r_{cat}(j,t) = \max_k(\overline{u}_{k,t} \cdot w_{j,t}^T)$. We use maximum aggregation (as opposed to mean or median) because a word only needs to be highly relevant to one sub-group to be considered as valid in-group language.

Subreddit: Similar to the category-based method, we find the maximum relevance score for a subreddit p using its centroid: $r_{sub}(j,t) = \max_p(\overline{u}_{p,t} \cdot w_{j,t}^T)$.

Cluster: We use K-means clustering of the user embeddings to find subcommunities. The motivation is that there may be groups that span different categories or subreddits, or sub-groups within a given category. For each cluster n we find $r_{clust}(j,t) = \max_n(\overline{u}_{n,t} \cdot w_{j,t}^T)$. We do this for $K \in \{5, 20, 100\}$. Since the clustering depends on random initialisation, we report the mean over 5 runs.

Bootstrap: Departing from the reconstruction-based approaches, we use a list of confirmed terms from manosphere lexicons (provided in Table 1) and find their nearest neighbours based on cosine similarity of the W embeddings. The score for each neighbouring word is given by its maximum similarity to a lexicon word.

Bias: The factorisation models described in Section 3 include an inferred bias term for each user and word to capture population-level trends. In the case of the word bias, it captures the tendency for a word to have a high PMI across all users as compared to a background corpus. We use this directly as the word score.

Aggregation These approaches produce a score for each word at every timestep. To aggregate, we use the maximum score of a word across all timesteps. This has the benefit of potentially capturing words that move into and out of popularity over the training window.

5 Evaluation

The goal of this work is to induce lexicons of in-group language. For the purpose of evaluation, we frame this as a **binary classification task** over **single-token lexical innovations**. Our rationale for this task definition and our evaluation setting

is provided below. The test set constructed for this task is described in Section 6.

A challenge in evaluating lexicons is that they are not a precise science and often rely on subjective expert judgment (Havaldar et al., 2024), as discussed in Section 1. For example, the lexicons in Table 3 contain words such as *hurt* and *dumb*, which represent borderline cases in deciding whether they are sufficiently relevant for inclusion. A further challenge is that dogwhistles (words that take on a special meaning within an in-group that is different from its meaning to the general population) are difficult for researchers to identify (Lucy and Bamman, 2021; Mendelsohn et al., 2023).

To address these challenges, we focus on **lexical innovations** in our evaluation. Lexical innovations are a form of in-group language which involve non-standard or extragrammatical words (Slotta, 2016); for example, *foid* is a common derogatory term for a woman in the Incel community. Lexical innovations provide a clearer decision boundary as they tend to be more easily attributed to a specific community. Importantly, this is a constraint on the evaluation and not a limitation of the method; all methods used in this work can produce relevance scores for standard and non-standard language alike.

A set of standard words is required in order to identify non-standard lexical innovations. For this purpose, we use the vocabulary of GloVe-400k (Pennington et al., 2014), which is based on news and Wikipedia articles. Words that are not in this vocabulary are considered non-standard. We make an exception to include in-vocabulary words from existing manosphere lexicons (detailed in Section 6) in our evaluation, irrespective of whether they are lexical innovations. Since prior work has deemed these terms to be of particular importance to the manosphere, we consider them to be relevant and useful for evaluation.

Finally, we consider only **single-token words** to limit the vocabulary size. Future work may investigate N-gram models, since multi-word in-group expressions are also observed in the data. Within the manosphere, common multi-word expressions are often made into acronyms; for example, *AWALT* is an acronym of "all women are like that".

5.1 Baselines

We use baselines based on two prior lexicon induction approaches:

Lexicon expansion   Similar to the approaches of Yoder et al. (2023), Farrell et al. (2020) and Havaldar et al. (2024), we train a word2vec model (Mikolov et al., 2013) on the same subset of posts used to train the LISTN models. We then find the nearest neighbours of known manosphere in-group language (based on the lexicons in Table 1) using cosine similarity over the W embeddings.

Statistical measures   Similar to Lucy and Bamman (2021) and Zhang et al. (2017), we evaluate variants of PMI, including PPMI (Eq. 1) and normalised PMI (**NPMI**). Given a word j and a particular setting (e.g. subreddit or community) k:

$$\text{NPMI}(j, k) = \frac{\text{PMI}(j, k)}{-\log P(j, k)}. \qquad (3)$$

To determine $P(j)$ in the PMI calculation, we use the same background corpus used to construct the C matrices. Probabilities are calculated based on word frequencies. Mirroring the LISTN evaluations, we account for sub-group specialisation by evaluating each PMI metric at the level of the full community, subreddit, and category, using the maximum score for a given word across the set of subgroups in each case. We also evaluate the NPMI and PPMI using monthly slices of data and maximum aggregation to mimic the temporally-aware nature of the LISTN methods.

5.2   Metrics

We report the average precision (AP) and area under the receiver-operator characteristic curve (AUROC) for each method. The AP provides a measure of the model's precision over all possible recall values; that is, how well the model avoids false positives. AUROC is a measure of its ability to distinguish between positive and negative cases. Davis and Goadrich (2006) show that a model dominates in AUROC space if and only if it dominates in AP space, whereas the converse is not true; as such, we prioritise the AP. We use the randomised permutation test with Monte Carlo approximation ($R = 9999$ and $\alpha = 0.05$) for significance testing.

6   Test set construction

We perform a preliminary scoring step for the purpose of constructing a test set for our main evaluation. Existing lexicons of manosphere language are used as a reference point.

Source	# words
IncelWiki	189
Farrell et al. (2019)	194
Baele et al. (2023)	122
Lucy and Bamman (2021)	52
Hatebase	29
Total (deduplicated)	**483**

Table 1: Existing lexicons used in the evaluation.

Existing lexicons   The lexicons and the number of words in each, filtered to include only words that are in our corpus, are shown in Table 1. For the Farrell et al. (2019) dataset, we follow Ribeiro et al. (2021) in excluding the lexicon categories that are not specific to the manosphere (e.g. racism). The Hatebase[3] and IncelWiki[4] lexicons were collected by us and will be released. Combined, this provides a set of 483 unique words.

Preliminary scoring   In this step, we evaluate each of the approaches in Sections 4 and 5.1 against the lexicons in Table 1. Terms in the lexicons are assigned a positive label; all others are considered negatives. The average precision and area under the receiver-operator characteristic curve of each method are provided in Appendix B. An issue with this approach, however, is that the lexicons suffer from low recall (a known issue; see Lucy and Bamman, 2021), meaning that true in-group terms identified by the methods are often evaluated as false positives. For this reason, we use this scoring only to identify words to include in our test set.

Manual validation   We retrieve the top-scoring 1000 terms produced by the best baseline and best LISTN approaches from the preliminary scoring step. These terms are manually evaluated by the author of this paper as well as a social psychology PhD student who is a specialist on the manosphere. Annotators are tasked with assigning a binary label to each word, based on whether or not it strongly signals membership of the manosphere. Annotators are also provided with three example usages of each word and the corresponding subreddit name. Following the expert annotator's advice, we also annotate "cultural exports" – terms that originated within the manosphere but have spread to be used more widely. In our evaluation, we treat these terms as positives. We further include the terms from the

---

[3]hatebase.org/
[4]incels.wiki/w/Incel_Glossary

	AP	AUROC
Random	0.52	0.5
Baselines		
word2vec bootstrap	0.5563	0.5427
PPMI	0.6333	0.6709
PPMI-subreddit	0.5910	0.5916
PPMI-category	0.6517	0.6470
PPMI-month	0.6165	0.6071
NPMI	0.6709	**0.7034**
NPMI-subredit	0.6170	0.6187
NPMI-category	**0.6790**	0.6647
NPMI-month	0.6449	0.6386
LISTN-CA		
Community centroid	0.6228	0.5131
Category centroid	0.7231	0.6713
Subreddit centroid	0.5519	0.5043
Cluster-5	**0.7620**	**0.7403**
Cluster-20	0.7069	0.6954
Cluster-100	0.6950	0.6566
Bootstrap	0.5349	0.5206
Bias	0.6190	0.5560
LISTN-C		
Community centroid	0.6297	0.5255
Category centroid	0.7272	0.6809
Subreddit centroid	0.5891	0.5461
Cluster-5	**0.7679**	**0.7363**
Cluster-20	0.7554	0.7289
Cluster-100	0.7040	0.6864
Bootstrap	0.5276	0.5077
Bias	0.6016	0.5383

Table 2: Comparison of different lexicon induction approaches.

existing lexicons in Table 1 as positive samples.

The resulting test set consists of 1803 words with binary labels, with an inter-annotator agreement score of 0.726 using Cohen's Kappa, indicating substantial agreement. In cases of disagreement, we use the label from the expert annotator. The test set is reasonably balanced, consisting of 944 positive and 859 negative samples.

7 Results

Results are shown in Table 2. The best LISTN-CA and LISTN-C methods both outperform the best baseline method (statistically significant, $P << 0.05$). Using LISTN-C produces a higher AP score compared to LISTN-CA for 6 out of the 8 induction methods, including the top-performing method (not statistically significant; $P = 0.723$). This is a sur-

prising result, as it indicates that the induction does not benefit from the added author adjacency information (i.e. shared thread interaction) in LISTN-CA. A benefit of this outcome is that the LISTN-C model has fewer parameters and is more computationally efficient to train. The LISTN-C model does still contain social information, in that the source matrix is a user-content representation, meaning that users are represented as being similar if they use the same words.

LISTN For both LISTN-CA and LISTN-C, the 5-cluster approach performs the best ($AP = 0.75$ and 0.77 respectively). Using a larger number of clusters ($K = 20$ and $K = 100$) leads to worse performance. A possible explanation might be that smaller clusters lead to less meaningful centroids that overemphasise irrelevant words. On the other hand, averaging over the full community performs poorly, indicating that it is necessary to take into account *some* level of sub-group linguistic specialisation. This is corroborated by the relatively low scores of the bias and bootstrap methods, which also do not account for community structure.

Notably, the 5-cluster method performs better than category-level induction, which computes centroids of a similar level of granularity. Subreddit-level induction (a more granular version of the category-level method, with 52 subgroups) performs the worst of the centroid-based approaches for both models and metrics. This is an unexpected result; one would expect this approach to perform on-par with the other centroid-based approaches. The cluster-based methods are intended to capture social structure at the sub-subreddit and cross-subreddit level, and assign each user to one cluster only. The subreddit representations average over all users who have contributed to a given subreddit per timestep, meaning that one user may form a part of many representations. This might include users who are incidental posters, which would add noise to the centroids. This effect would be more pronounced at the subreddit level than the category level, as there is less cross-category activity, which would explain why the category aggregation is better than subreddit aggregation but still worse than the clustering-based methods. Using a user weighting based on post frequency might resolve this issue. However, since the subreddit structure does not translate to other platforms, it is beneficial that the best method does not require it.
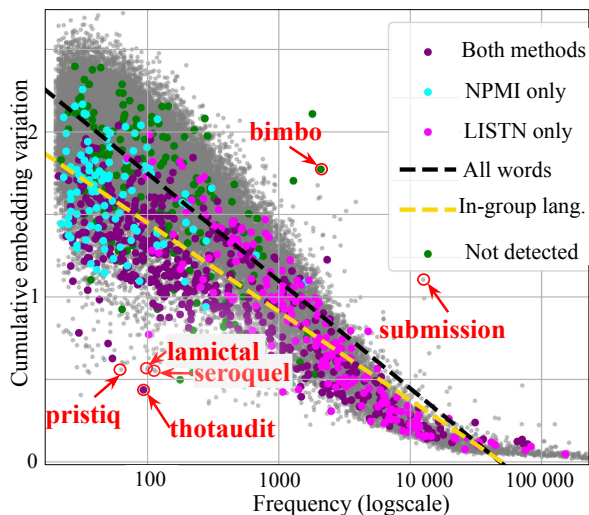
Figure 1: Word frequency versus embedding stability.

Baselines All baseline approaches outperform the random baseline. The worst-performing baseline is the word2vec-based method. For both LISTN models, the lexicon expansion approaches ("bootstrap") perform the worst as well. A possible reason for this might be that the seed lexicons are not varied enough, such that they are missing some area of the embedding space. The NPMI baseline performs the best on AUROC, whereas the category-level NPMI performs the best for the AP. In all cases, NPMI outperforms the PPMI, which corresponds with the results of Lucy and Bamman (2021). Interestingly, using month-level NPMI scoring does not improve over the time-aggregated NPMI, indicating that these superficial methods for accounting for time are not sufficient.

8 Embedding analysis

Beyond the improved performance over prior lexicon induction methods, an advantage of this approach is that it produces dynamic embeddings for words and users, in the same space. In this section, we use these embeddings to investigate (*i*) changes in words over time and (*ii*) sub-group linguistic specialisation in the manosphere.

8.1 Change in word representations

Recall that the training objective includes a regularisation term that penalises large changes between temporally successive word representations, meaning that changes in embeddings are only permitted if they reduce the reconstruction error sufficiently. In other words, word embeddings change as much as is needed, but no more. To determine how much

a word changes between two successive timesteps, we calculate the cosine distance between successive W_t representations. The sum of these distances all timesteps is referred to as the cumulative embedding variation (CEV).

We find that the words with the lowest CEV are stopwords, e.g. *from*, *here*, *there*. This makes sense, given that the embeddings are a function of how a word is used throughout a community, and these words are not expected to be impacted by the social evolution of the group. At the opposite end of the spectrum, showing the most variation, we observe less common words such as *lackeys*, *wooo*, *fragment*, and *booklet*. Noting that there seems to be a relation between word frequency and CEV, we calculate their Spearman correlation, yielding strongly negative value, $\rho = -0.77$ ($P << 0.05$).

Figure 1 shows the relation between the log frequency of a word and CEV, for all words in the training data (grey dots). We also plot the confirmed in-group language along the same axes (colourful dots) in Figure 1. The strong negative correlation only appears to hold for frequencies under approx. 10 000; beyond this point, changes become close to zero. We have two hypotheses for why this might be. Firstly, it may simply be a computational issue, whereby words that are less common do not have high-quality representations due to insufficient training data, and therefore their embeddings are less stable. The second option is that words that are less common have more pliability in their colloquial usage. We expect the reality is a combination of the two.

We further observe that the variance in CEV covaries with frequency, such that a larger range of values is observed for low-frequency words. The confirmed in-group language tend to exhibit lower CEV per frequency, relative to the general vocabulary. Dotted lines indicate straight-line regressions to CEV of the whole vocabulary (black) and to the in-group language lexicon (yellow). We observe that the gradient for the in-group language is indeed less steep. This suggests that in-group language terms are more stable and less pliable in how they are used within this community, relative to what we would expect given their frequency. This could be a result of their special status as a social artefact in the group: there is an incentive to use the term "correctly" to signal group belonging.

The different colours represent the predictions made by the best models in Section 7, showing the top 1000 words predicted by the NPMI-category

Group	Top words
MGTOW	*thot, awalt, **gyow**, thots, simps, simp, beta, mgtow, **cucked**, alpha, cuck, bitches*
Incels	*normie, blackpill, jfl, volcel, suifuel, **braincels**, normies, foids, **incels**, foid, **blackpilled**, stacy*
Mental health	*hurt, failure, personality, killing, cope, harm, kill, hit, tbh, sa, hurt, dumb, beat*
TRP	*trp, smv, rp, ltr, oneitis, alpha, beta, awalt, redpill, bp, **asktrp**, **hamstering***
MRA	*rape, raped, assault, patriarchy, female, abused, **mras**, dumb, hurt, mgtow, raping, fgm*
PuA	*ioi, trp, smv, ltr, oneitis, alpha, **iois**, personality, beta, **seddit**, **daygame**, gf*
Criticism	*trp, **incels**, manosphere, alpha, beta, rp, dumb, personality, asshole, redpill, redpilled, mgtow*

Table 3: Highest relevance lexicon terms per category.

Figure 2: Spearman correlations between in-group language of different manosphere sub-groups.

and `LISTN-C` (cluster-5) methods. The two methods appear to favour different areas in the frequency distribution, with the NPMI-based model being biased to the lower end of the spectrum, and the LISTN-based model to the mid and higher side. Future work may investigate combining these methods for lexicon induction.

Investigating the low-frequency, low-variation words, we find that it includes several mentions of medications (e.g. *lamictal, pristiq, seroquel*). These are concepts with specific technical meanings that are not likely to vary due to socio-temporal changes in the community. We also note the term *thotaudit*, which is the hashtag of a harassment campaign targeting sex workers. On the other end of the spectrum, *submission* and *bimbo* display a large amount of variation relative to their frequency. A possible reason for this might be that these terms occupy different niches for the subcommunities of the manosphere, or that they underwent a socio-semantic change during the training period. While more exploration of this embedding space is warranted, it is evident that it captures some interesting and useful qualities of the data.

8.2 Sub-group linguistic specialisation

The results in Section 7 show that lexicon induction benefits from accounting for social sub-group structure, suggesting that there might be sub-group linguistic specialisation. In this section, we further investigate this idea. The hypothesis is: if manosphere categories share a similar vernacular but express it to varying degrees, we would expect there to be a (nonlinear) correlation between their preference for different words in the lexicon. If the groups favour disjoint parts of the lexicon, there would be no such correlation.
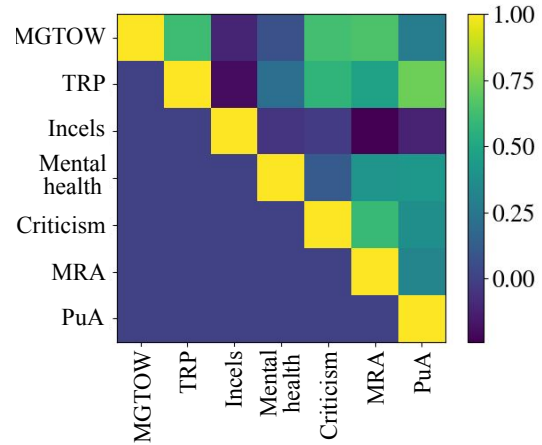
To evaluate this idea, we firstly calculate the relevance of individual words in the lexicon to each of the manosphere categories. Table 3 shows the highest-ranking lexicon words for each grouping. These words align with their known thematic interests; for example, the Pick-up Artists use terms such as *IOI* ("indicators of interest") and *SMV* ("sexual market value"). To enable future research on these communities, these scores are also published.

We then find the Spearman correlation between the per-word relevance scores for each pair of communities, shown in Figure 2. We observe that they correlate to varying degrees, with the largest correlation ($\rho = 0.729$) between the Pick-up Artists (PuA) and The Red Pill (TRP). This has been observed in prior work in sociology: Bachaud (2024) states that PuA terminology is plentiful in the TRP community, as they have a shared focus on seduction. The MGTOWs have a fairly strong correlation with TRP ($\rho = 0.615$) and MRA ($\rho = 0.654$). This aligns with their historical connections: Bachaud (2024) states that MGTOW was created by a group of MRAs and that their ideology is "close to both Red Pillers (critical about female nature) and MRAs (denouncing feminism and structural biases against men)". A much weaker correlation is observed between MGTOW and PuA ($\rho = 0.282$), which can be explained by their conflicting ideologies: PuAs are intent on pursuing women, whereas the main premise of MGTOW is to eschew contact with women. This captures an interesting insight into the relationship between PuAs, MGTOWs and TRPs: non-overlapping parts of the TRP interests are shared by both other groups.

The Incels have negative correlations with all other communities (with a minimum $\rho = -0.240$ compared to the Men's Rights Activists), showing that they have the most specialised in-group language. This, too, is supported by manosphere history: while the other manosphere groups have fairly intertwined origin stories, the Incels emerged from a separate line of new online platforms such as LoveShy (whose administrators openly supported mass killers and femicide) with cross-polination from 4/chan and Reddit (Bachaud, 2024).

The Criticism groups, who are reformed manosphere members, correlate positively with MGTOW, MRA, TRP and PuA. Investigating their most relevant words (Table 3), we note that it mostly lists references to the manosphere itself.

9 Conclusion

In this paper, we introduced a novel framework for lexicon induction based on dynamic word and user embeddings. We compare several approaches based on two different basic models to methods from prior work, finding that the proposed methods achieve better scores on a newly crafted, expert-validated test set. A set of 455 new manosphere terms are to be released based on this analysis, with relevance scores for each manosphere sub-group. Our work provides novel insights into in-group language, including its embedding stability. The investigation into sub-group linguistic specialisation provides insights into the relationships between the manosphere subcommunities, which corroborate claims from prior work in sociology.

To our knowledge, this is the first work to incorporate social and temporal information to induce lexicons of in-group language. As the methodology is not specific to a particular group or language, we hope that it will be applied to create lexical resources for other communities in future.

Ethical considerations

The experiments in this paper make use of data developed in prior work. Ribeiro et al. (2021) state that the ethical standard guidelines of Rivers and Lewis (2014) were followed in the construction of this dataset, which includes not making any attempt to de-anonymize or link users across platforms. Even so, we recognise that the users did not consent to having their data analysed for the purpose of scientific experiments. However, given the real and current threat posed by this group, we view it is an acceptable use of the data. To limit the exposure of individuals, we do not provide any quotations or usernames and all analyses are performed at an aggregate level.

As we are mindful of potential mental health repercussions of this work to collaborators, participants (including annotators) who were exposed to texts originating from manosphere communities were encouraged to attend a support group for researchers who work with extreme content.

Limitations

There are a number of simplifying assumptions that are made in this work. All models evaluated (LISTN and baselines) provide only one score per word, and hence do not account for sense variation or polysemy. We further include only single-word lexical variations in our evaluation and exclude non-alphanumeric characters. Our motivation for these choices is provided in Section 5.

The LISTN and word2vec models are trained on data from 2018. Given the dynamic nature of in-group language, this limits the applicability of the lexicon to current data; however, the evaluations in this work serve to validate our lexicon induction approach. Given recent limitations placed on Reddit data access, the Ribeiro et al. (2021) dataset still constitutes one of the largest and most varied existing manosphere datasets. Future work may look at applying these methods to more modern datasets (possibly in the context of specialised forums like incels.is).

Acknowledgements

References

Louis Bachaud. 2024. *The Manosphere: History, Sociology, and Ideology of Contemporary Manosphere Groups*. Preprint.

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.

Ksenija Bogetić. 2023. Race and the language of incels: Figurative neologisms in an emerging english cryptolect. *English Today*, 39(2):89–99.

Stevie Chancellor, Jessica Annette Pater, Trustin Clear, Eric Gilbert, and Munmun De Choudhury. 2016. # thyghgapp: Instagram content moderation and lexical variation in pro-eating disorder communities. In *Proceedings of the 19th ACM conference on computer-supported cooperative work & social computing*, pages 1201–1213.

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.

Jesse Davis and Mark Goadrich. 2006. The relationship between Precision-Recall and ROC curves. In *Proceedings of the 23rd international conference on Machine learning*, pages 233–240.

Christine De Kock. 2024. Jointly modelling the evolution of community structure and language in online extremist groups. *arXiv preprint arXiv:2409.19257*.

Christine De Kock and Eduard Hovy. 2024. Investigating radicalisation indicators in online extremist communities. In *Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH 2024)*, pages 1–12.

Marco Del Tredici and Raquel Fernández. 2018. The road to success: Assessing the fate of linguistic innovations in online communities. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1591–1603, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Glendon Frank Drake. 1980. The social role of slang. In *Language*, pages 63–70. Elsevier.

Nouha Dziri, Ehsan Kamalloo, Kory Mathewson, and Osmar R Zaiane. 2019. Augmenting neural response generation with context-aware topical attention. In *Proceedings of the First Workshop on NLP for Conversational AI*, pages 18–31.

Tracie Farrell, Oscar Araque, Miriam Fernandez, and Harith Alani. 2020. On the use of Jargon and Word Embeddings to Explore Subculture within the Reddit's Manosphere. In *Proceedings of the 12th ACM conference on Web Science*, pages 221–230.

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.

Shreya Havaldar, Salvatore Giorgi, Sunny Rai, Thomas Talhelm, Sharath Chandra Guntuku, and Lyle Ungar. 2024. Building knowledge-guided lexica to model cultural variation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 211–226.

Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural collaborative filtering. In *Proceedings of the 26th international conference on world wide web*, pages 173–182.

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

Li Lucy and David Bamman. 2021. Characterizing english variation across social media communities with bert. *Transactions of the Association for Computational Linguistics*, 9:538–556.

Julia Mendelsohn, Ronan Le Bras, Yejin Choi, and Maarten Sap. 2023. From dogwhistles to bullhorns: Unveiling coded rhetoric with language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15162–15180.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

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.

Caitlin M Rivers and Bryan L Lewis. 2014. Ethical research standards in a world of big data. *F1000Research*, 3:38.

Matthew Rowe and Hassan Saif. 2016. Mining pro-ISIS radicalisation signals from social media users. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 10, pages 329–338.

James Slotta. 2016. Slang and the semantic sense of identity.

Ian Stewart and Jacob Eisenstein. 2018. Making "fetch" happen: The influence of social and linguistic context on nonstandard word growth and decline. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4360–4370.

Michael Yoder, Chloe Perry, David Brown, Kathleen M Carley, and Meredith Pruden. 2023. Identity construction in a misogynist incels forum. In *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 1–13.

Justine Zhang, William Hamilton, Cristian Danescu-Niculescu-Mizil, Dan Jurafsky, and Jure Leskovec. 2017. Community identity and user engagement in a multi-community landscape. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 377–386.

A   Hyperparameters

Hyperparameters for the representation learning component are shown in Table 4.

Parameter	Value
Early stopping patience	5 epochs
Early stopping tolerance	0.001
K	100
Batch size	80
Weight of A	1
Weight of C	1
Learning rate	0.001
λ_1	1
λ_2	1
c_0 scaler	0.01
Max epochs	100

Table 4: Hyperparameters for training the Cerberus system.

B   Preliminary evaluation

Results for the first step of the evaluation are shown in Table 5.

	AP	AUROC
Baselines		
PPMI	0.2162	0.3576
PPMI-subreddit	0.2564	0.4293
PPMI-category	0.2103	0.3470
NPMI	0.2864	0.4171
NPMI-subredit	**0.3377**	**0.5322**
NPMI-category	0.2748	0.4114
LISTN-CA		
Community centroid	**0.6147**	**0.7344**
Category centroid	0.5577	0.6997
Subreddit centroid	0.3987	0.6287
Cluster-5	0.4302	0.637
Cluster-100	0.3399	0.5554
Bootstrap	0.2811	0.4960
Bias	0.5690	0.7065
LISTN-C		
Community centroid	0.6147	0.7288
Category centroid	**0.6809**	**0.7272**
Subreddit centroid	0.5461	0.5891
Cluster-5	0.491	0.6827
Cluster-100	0.4007	0.6184
Bootstrap	0.5276	0.5077
Bias	0.6015	0.5382

Table 5: Results for preliminary experiments, used to construct the gold label test set.