Topic Modeling and Insights from AITA Subreddit

Henry Arze*
henryarze@vt.edu
Virginia Tech

Paras Goda*
pgoda@vt.edu
Virginia Tech

Figure 1: Banner image from Reddit's r/AmItheAs!*@!e subreddit [8].

1 Abstract

We are investigating how moral reasoning is expressed in the r/AmITheAsshole (AITA) subreddit using transformer-based topic modeling and contextual analysis. We build on prior linguistic insights [7] by using BERTopic [5] with MPNet and RoBERTa to generate semantically coherent clusters and relevant topic pairs. Furthermore, we measure how the meaning of Moral Foundation Dictionary [4] terms shifts across different contexts by comparing topic pair specific embeddings to global baseline embeddings.

Our results demonstrate that moral terms undergo significant semantic shift depending on the social dilemma being discussed. Transformer models, specifically MPNet, capture these contextual nuances with great range. By analyzing the intersection of topic pairs and Moral Foundations, we provide a more granular framework for categorizing moral dilemmas based on the magnitude of these contextual deviations.

1.1 Motivation

Ideal moral dilemmas like a tram that will kill the people on one of two tracks, one with a single person and one with five, [3] have been well deliberated. "Morally Charged" situations [2] expand moral dilemmas to more relatable experiences like "AITA For Firing An Employee After His Parents Died? [7]." Classification algorithms for the AITA exists [1], however moral dilemmas do not have a clear right or wrong answer and are highly sensitive to culture and social norms. While we will not delve into the reasoning behind this, it is a key driver for exploring moral dilemmas. Understanding the behavioral and linguistic patterns is a precursor to creating AGI (artificial general intelligence) that can consider different views, background and individual differences. This is highlighted by the fact that the verdict for the subreddit threads is determined by majority vote, illustrating that the moral consensus is not unanimous.

2 Dataset

For consistency and to build on the exciting prior work of Nguyen et al.[7], the released dataset from the AITA subreddit is selected for analysis. This subreddit consist of common real-life scenarios that are more likely to have engaging conversations from the masses. It should provide a large demographic (assumed to be true by the fact that it is open to the public) as well as a copious dataset as opposed to "lab" style surveys. The mechanism for AITA is an original poster (OP) will create a dilemma. Anyone can post a comment, debate and promote a certain verdict. The verdict/comment with the largest upvotes is deemed the final verdict for the post. This allows a single viewpoint to not dominate the judgments. This crowd-sourced data is perfect for collecting a diverse set of viewpoints that are not tainted or tailored for this analysis. The voting and automatic judgment mechanism of AITA subreddit allows for consistency over mediums like news, twitter or other subreddits.

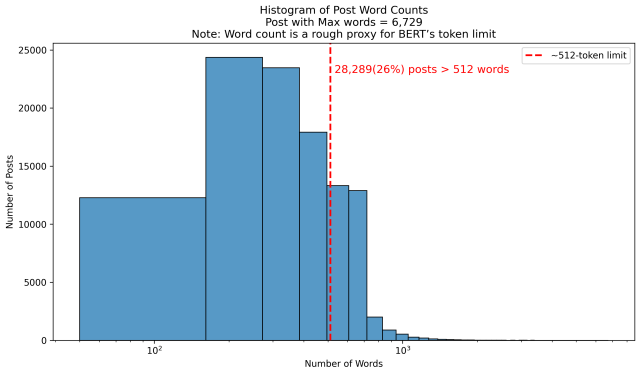

Figure 2: Histogram of Post Word Counts

148,961 posts from 8 June 2013 to 30 April 2020 of the AITA subreddit will be examined. Only posts with verdicts, titles beginning with "AITA" or "WIBTA", have at least 50 words in the body, 10+ comments and a verdict will be used. This ensure we have a breadth of context and comments. Figure 2 shows a histogram of the 108,634 posts that meet the filter criteria. Note that 26% of the posts have more than 512 words. Tokenization will increase the count, but a significant percentage will need to be addressed via chunking for models like BERT. The verdicts that will be used include "you are

*Both authors contributed equally to this research.

the asshole" (YTA) and "everyone sucks here" (ESH) to indicate a negative valance of "Yes Asshole" (YA) class. Verdicts of "not the asshole" (NTA) and "no asshole here" (NAH) will indicate a positive valance or "Not Asshole" (NA) class. Figure 3 shows the breakdown of verdicts. Only 29% of the posts are YA, so we expect more pronounced shifts in that subset.

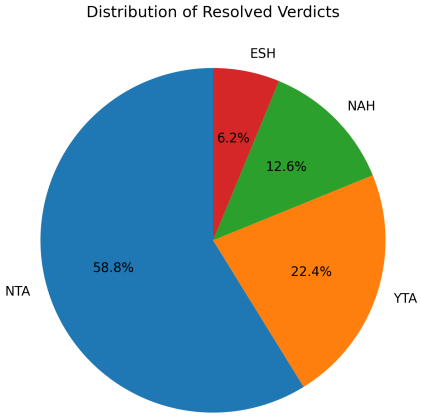

Figure 3: Pie Chart of Resolved Verdicts

2.1 Examples from Dataset

The raw dataset is in JSON and consists of a number of key/value pairs. The relevant ones to our research are listed below.

Key	Meaning
_id	Unique ID of the post
author	Author of the post
title	Title of the post
selftext	Body of the post
num_comments	Total number of comments the post received
created_utc	post create date/time
num_words	Total number of words in the post
resolved_verdict	Acronym of final verdict of the post

Examples of a posts (abridged to fit on the page) follow.

```
{
  _id: '1fy0bx',
  author: 'xxxxxxx',
  title: 'AItA: I like air conditioning and my coworkers like working half-naked
       .',
  selftext: "I work in an office that requires me to wear a suit all the time.
       ...  *Am I the Asshole?*\n"
  num_comments: 1,
  created_utc: ISODate('2013-06-08T20:42:55.000Z'),
  num_words: 242,
  resolved_verdict: 'NTA',
  ...
}
```

The longest post in the dataset is over 6000 words, which is addressed in the models.

```
{
    _id: 'aqc06k',
    author: 'xxxxxx',
    link_flair_text: 'Asshole',
    title: 'AITA for divorcing my wife even though I know we love each other?',
    subreddit: 'AmItheAsshole',
    selftext: '...' +
      'My wife and I love each other, but getting divorced because we love each
           other differently, and our parents have different views on a
           healthy relationship.',
    num_comments: 34,
    created_utc: ISODate('2019-02-13T21:50:51.000Z'),
    num_words: 6729,
    resolved_verdict: 'YTA',
    ...
}
```

3 Approach

The following approaches are used to gain insight on the AITA dataset.

3.1 Transformer-Based Topic Modeling

Nguyen et al. [7] previously used Latent Dirichlet Allocation (LDA) to infer topics from AITA posts. Their approach relied on perplexity to select the optimal number of topics, K, but after review, they manually increased the value.

Modeling with BERTopic, we utilize the transformer-based embeddings to group semantically similar posts without manually specifying the number of topics, K [5]. Leveraging existing Python libraries, the process is:

- **Embedding**: Encode each document with a pre-trained transformer (MPNet, RoBERTa)
- **Dimension Reduction**: Reduce the document embedding into a lower-dimensional space (through UMAP).
- **Clustering**: Use a density-based clustering algorithm (HDBSCAN) to produce K clusters.
- **Topic Representation**: Manually label meaningful topics.

We assess how the underlying encoder affects the quality of the produced topics. Encoders are RoBERTa [6] because of its broad appeal and stable embeddings, and MPNet [9] which can produce slightly better contextual semantics. We ask, are moral foundations expressed differently based on topic or theme of conversation and what are the patterns.

3.2 Embedding Documents

The first necessary step in the topic modeling process is to generate vector representations for each AITA post. Our goal is to use transformer-based embeddings (RoBERTa and MPNet) to embed each document, with the aim of capturing more context than frequency-based methods. One key characteristic of the posts in this subreddit is that they can come in a variety of lengths. More than a quarter of the posts in the dataset exceed 512 words, which is particularly important since the transformer models we use have 512 as the maximum token limit. Our proposed procedure ensures that all posts are represented consistently, regardless of the length.

The input of the embedding pipeline is the *selftext* field, which represents the body of the AITA post. Minimal preprocessing is performed on the input (empty values are removed), as we aim to retain as much context as possible. When sending this input to the encoder, a post that exceeds the maximum token limit will lead to truncation, ignoring any additional content past the limit. To avoid this behavior, we implement a chunking strategy to ensure that longer posts can still be fully represented, rather than being reduced

to only the earlier part of their discussion. Our process essentially tokenizes each post, splits the token sequence into contiguous chunks of at most 512 tokens, and then embeds each chunk. From here we can use the CLS token, which represents a learned summary of the entire chunk, to average the chunk embeddings into one single embedding. Averaging should preserve the entire narrative without adding extra weight to certain parts of a post. The final result is a document-level embedding that reflects the full content of each post through a stable and interpretable vector.

After embedding all documents from the dataset, the output of our implementation is an N x D matrix, where N is the total number of posts and D is the number of dimensions for each vector. (108634, 768)

3.3 Clustering

After generating document-level embeddings for each AITA post, the next step is to group semantically similar posts into topics. We will use BERTopic, which can use our embeddings to form coherent topic clusters. The topic modeling procedure consists of three main parts: dimensionality reduction, clustering, and topic word representation.

The 768 dimension embeddings from MPNet and RoBERTa are difficult to cluster directly, so our goal is to implement dimension reduction to give the clustering algorithm a workable dimension space. UMAP is the default in BERTopic, and will allow us to capture both local and global context in a lower-dimensional space [5]. We tuned UMAP to encourage tighter groupings with greater separation between clusters. Through our implementation, this reduction preserves local semantic relationships while preparing the embeddings for clustering.

Next, the UMAP embeddings are passed to HDBSCAN, BERTopic's default clustering technique. HDBSCAN will be able to capture the varying densities in structure of the dataset, creating more accurate topic representations [5]. Our goal was to use HDBCAN to naturally adapt the structure of the dataset, unlike k-means where the number of clusters (topics) must be explicitly set. We tuned HDBSCAN to produce more topics, which could reflect the diverse scenarios of the AITA subreddit. Furthermore, we wanted to produce results that could be comparable to the work of Nguyen et al. [7], and without proper tuning, we may find fewer or broader clusters. One major difference between HDBSCAN and k-means used in the referenced research paper is HDBSCAN labels low-density points as outliers. This is extremely important because in the next section, we discuss how each post is given a topic label, but outliers are not explicitly given a topic. However, through the computed probability distribution of each cluster (topic) per post, we can reassign the label of an outlier to its most probable topic.

3.4 Topic Modeling

From the results of our clustering, we store several key outputs for each document. These outputs will be essential for the contextual and moral foundation analysis. As mentioned before, each post will be assigned a hard label given by HDBSCAN, which includes outliers. Our code builds on this by also assigning each post their most and second most probable topic, derived form BERTopic's probability distribution across topics per post. With these two topics, we also store a topic pair, an idea proposed by Nguyen et al. [7] when documents are believed to be better represented with two topics rather than a single topic. This topic pair is unordered, emphasizing the relationship between posts with similar topic assignments.

Once the clusters are formed, BERTopic provides a deeper view of how each topic is represented by extracting the most informative words in each cluster. The default method from BERTopic uses c-TF-IDF to identify the top words that distinguish each cluster from the rest of the corpus [5]. We implemented a vectorizer that ensures representative words are useful by excluding English stop words and highly frequent and infrequent words.These (top 10) words will form the basis for our manual interpretation and topic naming.

3.5 Model Comparison

Before we dive deeper into the analysis of the topic modeling, it was important to address which pretrained transformer model would be stronger for our analysis. We directly compared the qualitative and quantitative results of both RoBERTa and MPNet. Below are two figures that illustrate why MPNet outperformed RoBERTA.

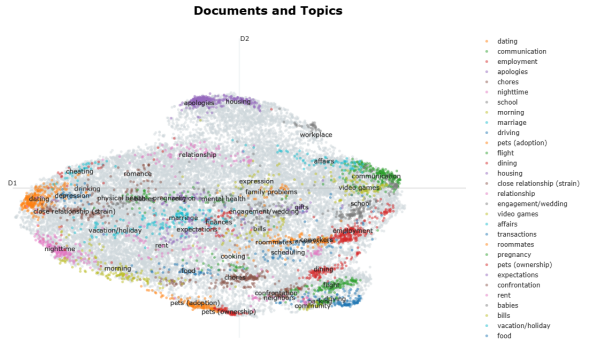

Figure 4: MPNet Clustering (1/4 of dataset)

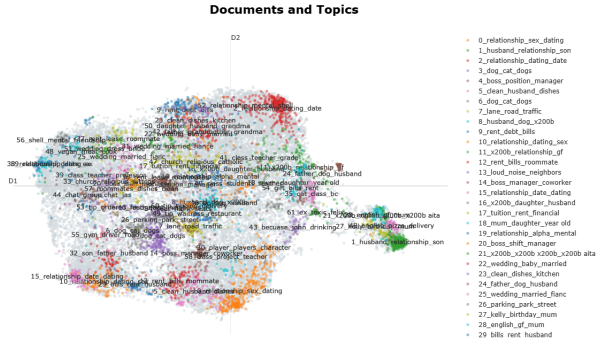

Figure 5: RoBERTa Clustering (1/4 of dataset)

We notice some key differences in the shape between the two scatter plots. Figure 4 shows more separated regions from MPNet

that correspond to interpretable themes, while the clusters formed using RoBERTa in Figure 5 are more blurry with unclear boundaries. The MPNet clusters have a bit more structure, while certain regions from RoBERTa appear more stretched or collapsed together. Looking at Figure 5, you can see much more overlap in the center of the plot, as well as a noisy region isolated to the right.

Beyond visualization, we can look at some metrics that further emphasize this difference in models. We computed topic coherence and topic diversity scores, which can be seen in the table below.

Model	Coherence (c_v)	# Topics
MPNet-base	0.6023	48
RoBERTa-base	0.5466	62

Table 1: Topic quality across transformer models.

The results show that MPNet excels in both metrics. The topic coherence score shows that MPNet produces topics with more semantically meaningful top words. We used CV, which works better in our case compared to a measure like UMass because CV will look at local context and document co-occurrence. The top words are seen more consistently across the documents in each topic. Conversely, RoBERTa produces topics where the top words are less similar, meaning that the topics are less interpretable. This will become a problem, particularly when naming topics. RoBERTa shares many words between topics, which explains the overlapped topics seen in Figure 5. An important thing to note is that the number of topics was aimed toward a count of 47, which comes from the work of Nguyen et al. [7]. Our goal was to try resemble their results through our unique setup, which further validates the selection of MPNet over RoBERTa, as it is very close to the count found in the referred study.

The above comparison is fair and appropriate because both models were evaluated under the same topic modeling pipeline. The process involved the same hyperparameters, meaning that the only variable that was changed was the embedding model. This difference allows for a reflection on the quality of the underlying transformer model, not on how much we can optimize for MPNet and RoBERTa separately.

3.6 Topic Representation

With our chosen MPNet topic model, we could look at the top c-TF-IDF words of each cluster to assign a meaningful topic name. As mentioned previously, BERTopic provides a ranked list of the most representative words for each topic. Using these top words, it was our task to manually name them while referencing several documents for validity. This approach of deciding a name that could represent the top words, then validating them through an inspection of the actual content and word use in the topic's post, would allow us to assign human-interpretable topics. Figure 6 gives a glimpse of how topics are represented by BERTopic, and the results of MPNet made this process easy because of the higher coherence. For example, we look at the words in the third topic, which all deal with words you might use when writing about work. We named this topic "employment", because we felt that the posts assigned to this topic involved AITA users describing scenarios at their job. Appendix A.3 shows the full topic naming convention and the top 10 words for each of the 48 topics.

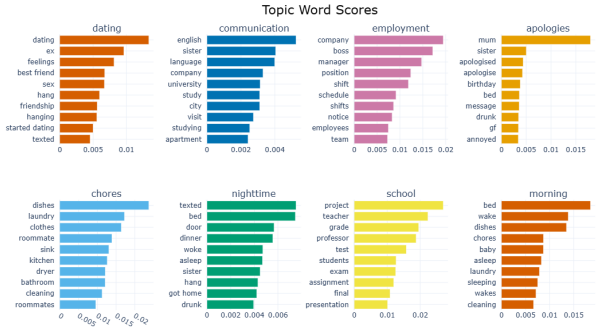

Figure 6: Topic-word distributions for Top 8 topics

3.7 Topic Pair Analysis

Although our BERTopic model produces topic assignments with high confidence, we will still conduct the rest of our analysis using topic pairs. This technique of representing a document by two topics rather than one parallels the work of Nguyen et al. [7]. Although the motivation behind their use emerges from more balanced topic probability distributions, the can still be insightful in providing more context. In the original approach of Nguyen et al. [7], an LDA and k-means setup led to a stronger relationship between two topics. This was due to concept that the top-2 probability was close enough to the top-1 probability that including it would provide more information.

In our transformer-based setup with HDBSCAN we found that the clusters were sharper, resulting in a much larger gap between the top-1 and top-2 probabilities. Due to the nature of density-based clustering, the most probable topic is very definitive. HDBSCAN tends to be very confident in a cluster, so we end up with a high probability for the most probable topic, leaving little room for a similarly probable secondary cluster. In comparison, k-means will split the groups more evenly, which is indicative of the motivation to use topic pairs by Nguyen et al. [7].

Topic Pair	Count
3, 13	5379
15, 44	3965
38, 42	3374
11, 12	2436
13, 36	2368

Table 2: Top 5 topic pairs in the dataset

For our purposes, the top-2 topic still provides a meaningful clue about semantic similarity. Although we cannot verify through something like PMI to measure the association between these two topic probabilities, Table 2 shows that there are topic pairs that occur more frequently than others. They can help identify which

topics are closest in the embedding space and can help isolate contextual ideas where more vocabulary changes. The following section will include our other major approach, which addresses how moral language changes its meaning based on the associated topic pair.

3.8 Moral Foundation Dictionary

The final approach in our study focuses on exploring how moral language changes meaning across topic pairs (context shift). We will use multiple transformer based contextual embeddings to capture the semantic variation of moral language as it relates to the Moral Foundations Dictionary (MFD 2.0) [4]. The implementation of this approach goes as follows:

- **Input**: Posts labeled by topic pair from topic modeling
- **For each topic pair**, per foundation:
  - Extract any word from MFD foundation in all posts
  - Create word-based embedding with a transformer model
  - Average the embeddings at the word-level to get a prototype vector for each word in MFD 2.0
- **Globally for all posts**, per foundation:
  - Create a baseline vector for each word across entire dataset
- **Output**: Calculate the context shift
- **For each topic pair**, per foundation:
  - Calculate the cosine difference (1- cosine similarity) per word between the topic pair prototype vector and global vector
  - Take the max difference per word

The output from this process will yield 2,104 distinct prototype vectors for each topic pair, one for each MFD word. It will also create 10 global baseline vectors, representative of the 5 moral foundations from MFD 2.0 and their vice and virtue splits. Note that we tokenize the posts and return a mean vector for the token span that makes up the matching target word. The posts are normalized by lower-casing all words. We analyze these results through the cosine difference by comparing the topic specific vector to the baseline vector (per foundation, per word). The max cosine difference will highlight the context shift of the foundation within each topic pair. Using the max distance will be indicative of any word where the moral concept shifts in meaning in that particular topic pair. The formula is as follows:

$$\mathbf{G}(\mathbf{w}) = \frac{1}{N}\sum_{i=1}^{N}\mathbf{e}_{w,i} \qquad \mathbf{TP}(\mathbf{w},\mathbf{k}) = \frac{1}{N_k}\sum_{j=1}^{N_k}\mathbf{e}_{w,j}$$

where $\mathbf{G}(\mathbf{w})$ is the global embedding of word w, $\mathbf{T}(\mathbf{w},\mathbf{k})$ is the topic pair specific embedding for pair k, word w, $\mathbf{N}$ is the total number of posts, $\mathbf{N_k}$ is the posts for topic pair k, $\mathbf{e_{w,i}}$ is the word embedding for the ith occurrence of all posts and $\mathbf{e_{w,j}}$ is the word embedding for the jth occurrence of posts in topic pair k. To calculate the heatmap we take

$$\text{Shift}(F,k) = \max_{w\in F}\{1 - \cos_{similarity}(\mathbf{TP}(\mathbf{w},\mathbf{k}),\mathbf{G}(\mathbf{w}))\}$$

where w is all the words in one of the 10 foundations, $\mathbf{F}$.

The two models implemented are *roberta-base* and *mpnet-base* for embeddings generation for four key reasons: 1) Extract precise contextual word embeddings for specific moral terms without "squashing" the entire post's meaning into a single vector. This requires access to pre-pooling hidden states. 2) Transformer style with self-attention mechanism will allow for deeper semantic meaning of the words 3) The bidirectional nature is important to capture the entire post's meaning vs auto-regressive decoder style models. 4) Large corpus including web text training data, better suited for social media informal posts.

To deal with the models 512 token limit, each post is chunked into 512 tokens with a 128 token stride to retain local context. Named entities are removed from the posts as they will distract from the contextual shift via Python's *Spacy* library.

4 Results

The question we ask is, are moral foundations expressed differently based on topic or theme of conversation and what are the patterns. We look into this in a exploratory manner with context shifts. The Moral Foundations context or semantic shift relied on at least one word of the foundations to be in a post. There was a 94% coverage of posts with at least one foundation word, and every topic pair had at least one word from each foundation in it.

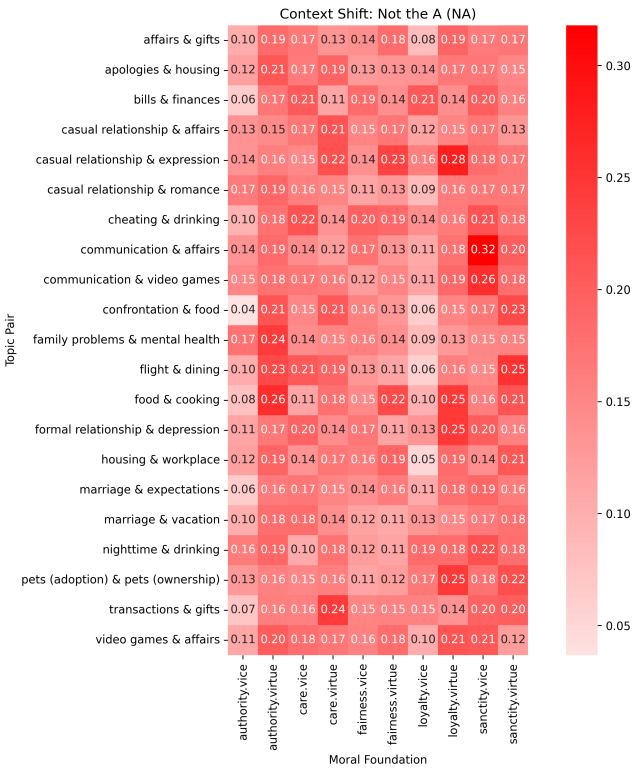

Figure 7: MFD Heatmap - RoBERTa (NA)

Figure 7 and Figure 8 show the top 20 topic pairs and intensity of the context shift. Note the scale, 0 represents perfect cosine similarly while the bright red squares indicate a defined shift in

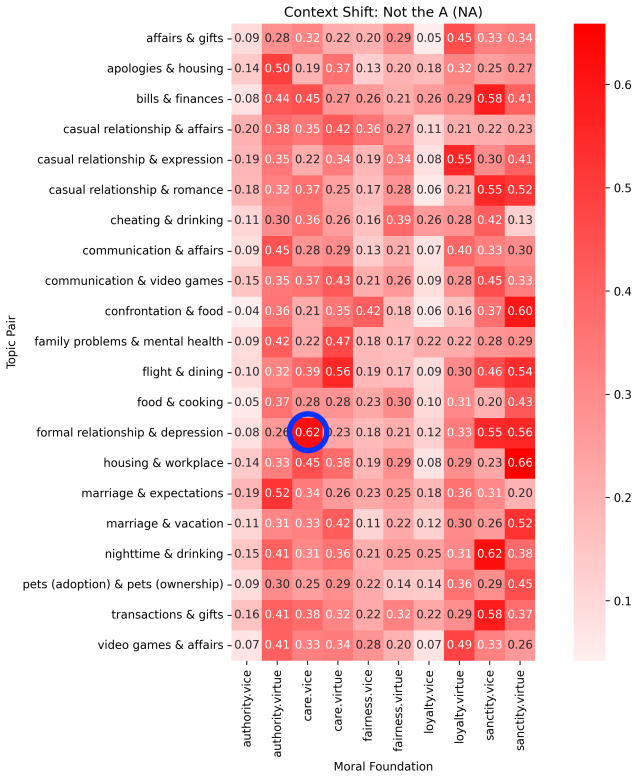

Figure 8: MFD Heatmap - MPNet (NA)

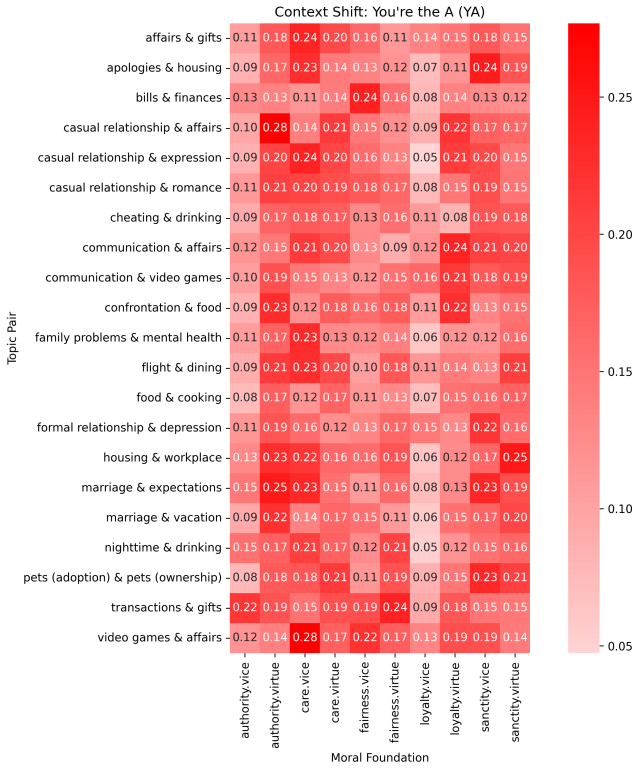

Figure 9: MFD Heatmap - RoBERTa (YA)

the embedding and meaning of the word. Both figure are the NA verdict produced with the RoBERTa and MPNet embedding models respectively. MPNet provides richer contextual representations due to its superior training and model architecture, to detect the drift better. RoBERTa and MPNet both use transformer attention heads allowing them to dynamically understand meaning of words based on surrounding context. Both models are similar in size (0.1B parameters) and bidirectional, however MPNet includes tricks like PLM (permutation language modeling). During training, PLM will create permutation of token order to better learn relationships between words (attention head). This objective fixes the issue with [MASK] token being predicted independently. MPNet predicts masked tokens auto regressively versus independently, to learn dependencies, and its attention head has auxiliary position information available. Similar behavior is also seen in Figure 9 versus Figure 10.

In order to first determine if our context drift values are meaningful, we need to understand a threshold of importance. A context drift value of 0 means the words in the foundation are used similar in the topic pair versus globally. A value of 1 means they are used differently and 2 means they are used with an opposite semantic. A crude test is to partition the dataset randomly into topic (pairs) and run the same algorithm as above. Appendix A.1 shows the results of a random topic pairing, where it is clear the max value of shift is 0.1. This is a crude way to remove noise from our results.

Before diving into the details, we can compare RoBERTA and MPNet models to see a few key points. One is that there are a significant number of squares with values above 0.1, showing the algorithm is detecting meaningful shifts. Over 88% of the cells were above 0.1 for RoBERTa and over 90% for MPNet. However, MPNet demonstrated a significantly wider dynamic range: its maximum cosine shift reached 0.77, compared to only 0.32 for RoBERTa. The larger delta for MPNet shows it is more sensitive to distinct contexts. MPNet highlights certain cells more and others less than RoBERTa, showing not only a richer relationship, but also more nuanced embeddings.

The blue circle in Figure 8 was analyzed for its top context shift words and corresponding post for the MPNet model (See Appendix A.2, Figure 13. Examples A and B below show a literal and hyperbolic meaning of the care foundation word "destroys". One example is from the topic pair while the other is from the global dataset.

Figure 10: MFD Heatmap - MPNet (YA)

Example A: "Destroys" Literally vs. Hyperbolically

Topic: formal relationship & depression

My SO and I have been dating for about 4 1/2 years now yet we barely have any physical intimacy. [...truncated for brevity...]
Aside from that detail she's quite possibly the closest thing to a superhuman I've ever seen. She runs her own business, helps me maintain the home, **destroys** me in smash, makes me laugh until my sides hurt, frequently buys me gifts and listens to me vent about work. [...]

Example B: "Destroys" Literally vs. Hyperbolically

Topic: pet (adoption) & pet (ownership)

So my girlfriend (whom I live with) has always wanted a dog. [...truncated for brevity...] Well fast forward a few months later and here we are with a 60lb 8 month old boxer.... And I hate it. ... When left alone, It chews anything near by, **destroys** blinds, tears up carpet and generally trashes the place. [...]

Example C is from the yellow circle in Figure 10, and its top context words can be seen in Appendix A.2, Figure 14. Looking at the word "ranking", we can see this is a failure case where misspellings created an out-of-place embedding.

Example C: The "Ranking" vs. "Renting" Typo

Topic: casual relationship & affairs

So my father is the sweetest man that you can meet... He is an amazing guy 99% of the time. [...truncated for brevity...] After that he starts talking about why i should listen to him and buy a house instead or **ranking** one.

A third example set shows a more common context shift that is common knowledge but undermines the value of the moral foundations dictionary. Many words have multiple definitions and may not express the moral foundation as the authors intended. Example D represents the shift at the intersection identified in blue in Figure 10, its shift words can be found in the Appendix A.2, Figure 15.

Example D: "Drug" dual meaning

Topic: marriage & expectations

So basic background is I'm 42, two kids age 14 and 12 and signed divorce papers in April after almost 2 years of it being **drug** out trying to get my fair share from him. [...truncated for brevity...]

Example E shows the more common use of the word as a noun that aligns with the moral foundation's intent.

Example E: "Drug" dual meaning

Topic: neighbors & community

Howdy all. I live in a fairly small city/town in central Oklahoma. We have a fairly large transient population due to 1) high **drug** use around here, 2) being on rail and highway corridors near OKC, 3) a relatively poor local economy, and 4) lots of reservations nearby (four actually within or right next to town). [...truncated for brevity...]

Major findings show that they dynamic range in MPNet embeddings allow us to see context shifts like Examples A and B that were not apparent from the RoBERTa heatmap. We also see that moral words across the authority, care and sanctity foundations are likely to be used in hyperbolic ways, where the other foundations are used by their common definition. The mechanisms behind this are outside of the scope of this study. Furthermore, NA posts are more likely to use different meanings of words then YA posts (shown by the higher number of large shifts). Affairs is a likely topic where one would use care vice words in emotional driven ways however less obvious pairs like food & cooking show a large shift for the sanctity vice foundation. Many instances were adjectives and proper nouns mixed up like "united" and "united states" or master's degree and master of a craft. Named entity recognition was able to clear some of these up, but not all.

5 Conclusion

In this study, we applied transformer-based topic modeling to explore how moral language varies between different interpersonal dilemmas from the AITA subreddit. Through a BERTopic pipeline that utilized UMAP and HDBSCAN, we produced semantically coherent topics that aimed to improve on alternative approaches that require a manually inputted number of topics. Although our HDBSCAN produced a highly confident topic assignment, topic pairs remained an area of discussion to characterize how moral language can change across related topics. Future improvements that could improve the results would be to minimize outliers produced by HDBSCAN. Although we could still assign a post to a topic, an "outlier post" did not have the same confidence in a topic assignment as the posts with a true topic label. Reducing the outliers would produce stronger clusters, which would improve many aspects of our analysis.

Moral foundation analysis was successful in showing interesting patterns for how web text of AITA uses words for moral dilemmas. However, the shear number of posts to review by humans is still large. Many of the richer context shifts were dominated by spelling errors or named entities. It is therefore more informative to look at the middle band of the context shift range. Future improvements that are most likely to help include spell checking. A mechanism to use control words could help remove more noise as would comparing the embeddings against ideal care foundation sentences versus global baselines. Some other future experiments could include stride length and larger context window models like longformer. However, the use of max cosine difference proves to be a powerful tool with the addition of topic pairs in order to see patterns of moral context shifts. We note that some foundations show larger shifts than others such as authority and sanctity, and NA posts have richer shifts than YA posts. Emotional topics are likely to see the largest shift as slang words are used hyperbolically, some of our biggest insights.

Overall, this work shows that transformer-based topic modeling and contextual semantic analysis can offer a strong framework for understanding variations in moral language in different real-world narratives. While there remain areas for improvement, our findings emphasize the value of contextually rich embeddings for capturing moral distinctions that traditional models may overlook.

References

[1] Nicholas Botzer, Shawn Gu, and Tim Weninger. 2021. Analysis of Moral Judgement on Reddit. arXiv:2101.07664 [cs.SI] https://arxiv.org/abs/2101.07664

[2] Julia Driver. 1992. The suberogatory. *Australasian Journal of Philosophy* 70, 3 (1992), 286–295. arXiv:https://doi.org/10.1080/00048409212345181 doi:10.1080/00048409212345181

[3] Philippa Foot. 1978. The Problem of Abortion and the Doctrine of the Double Effect. In *Virtues and Vices: And Other Essays in Moral Philosophy*. Basil Blackwell, Oxford, 19. Originally appeared in *Oxford Review*, No. 5 (1967).

[4] Jeremy A. Frimer, Reihane Boghrati, Jonathan Haidt, Jesse Graham, and Morteza Dehghani. 2019. Moral Foundations Dictionary for Linguistic Analyses 2.0. Unpublished manuscript. https://osf.io/ezn37/ Accessed: 2025-11-06.

[5] Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794* (2022). https://arxiv.org/abs/2203.05794

[6] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692* (2019). https://arxiv.org/abs/1907.11692

[7] Tuan Dung Nguyen, Georgiana Lyall, Alasdair Tran, Minjeong Shin, Nicholas George Carroll, Colin Klein, and Lexing Xie. 2022. Mapping Topics in 100,000 Real-Life Moral Dilemmas. *Proceedings of the International AAAI Conference on Web and Social Media* 16, 1 (2022), 699–710.

[8] Reddit, Inc. 2025. *r/AmItheAs****e*. Reddit. https://www.reddit.com/r/AmItheAsshole/ Accessed: 2025-10-31.

[9] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. MPNet: Masked and Permuted Pre-training for Language Understanding. *arXiv preprint arXiv:2004.09297* (2020). https://arxiv.org/abs/2004.09297

A Appendix

A.1 MFD Random Topic Basline

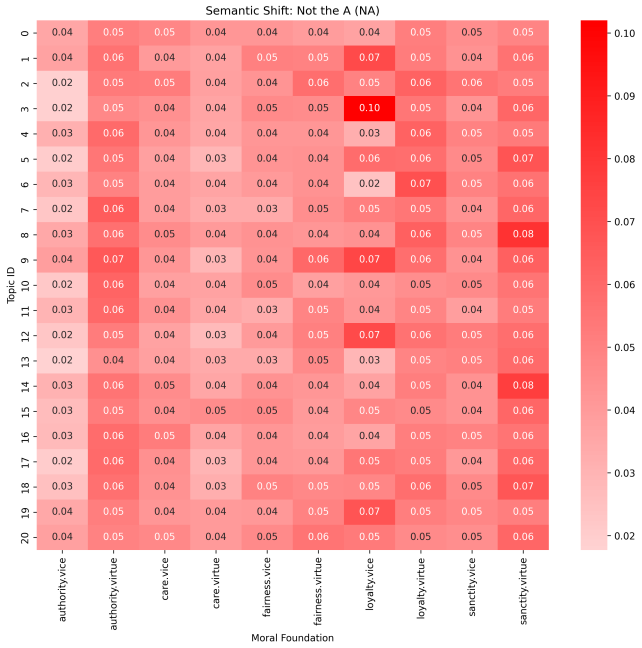

Figure 11: MFD Heatmap Random Topics Baseline - MPNet (NA)

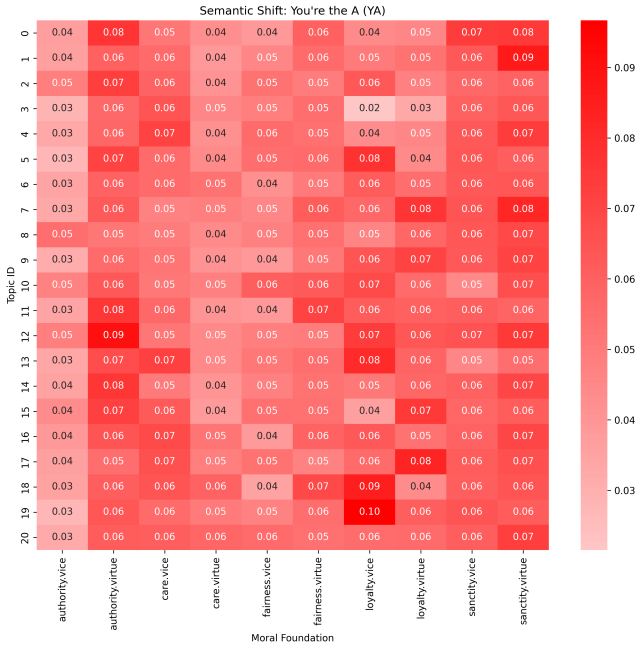

Figure 12: MFD Heatmap Random Topics Baseline - MPNet (YA)

A.2 MFD Top words in a topic pair/foundation

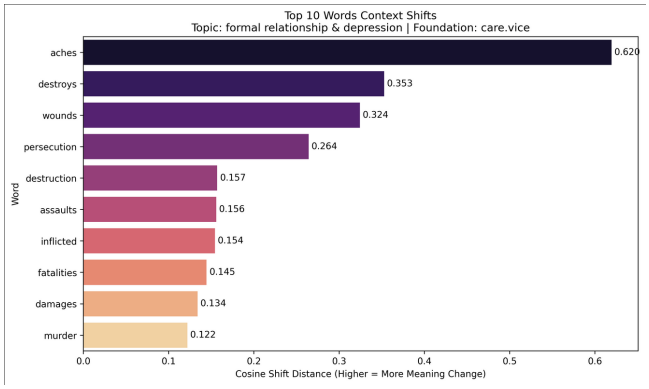

Figure 13: MFD Top 10 Words for the care.vice foundation in the formal relationship & depression topic pair - MPNet (YA)

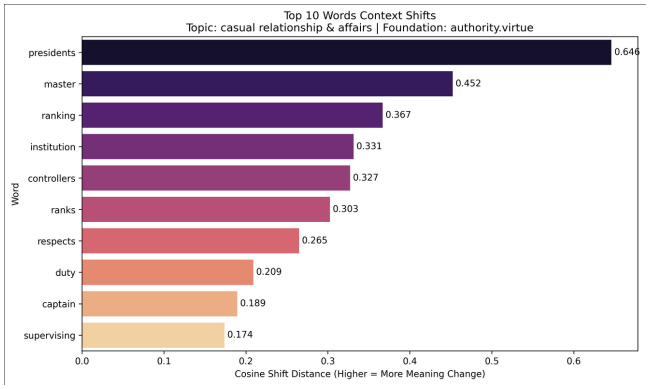

Figure 14: MFD Top 10 Words for the authority.virtue foundation in the casual relationship & affairs topic pair - MPNet (YA)

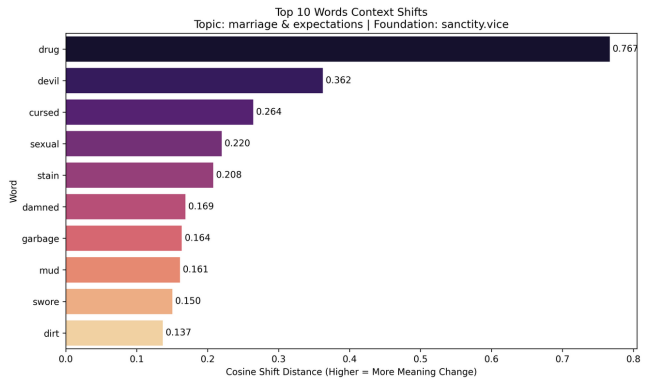

Figure 15: MFD Top 10 Words for the sanctity.vice foundation in the marriage & expectations topic pair - MPNet (YA)

A.3 Topic Labels and Top Words (MPNet Model)

Topic 0: dating dating, ex, feelings, best friend, sex, hang, friendship, hanging, started dating, texted
Topic 1: communication english, sister, language, company, university, study, city, visit, studying, apartment
Topic 2: employment company, boss, manager, position, shift, schedule, shifts, notice, employees, team
Topic 3: apologies mum, sister, apologised, apologise, birthday, bed, message, drunk, gf, annoyed
*__Topic 4: chores__ dishes, laundry, clothes, roommate, sink, kitchen, dryer, bathroom, cleaning, roommates
Topic 5: nighttime texted, bed, door, dinner, woke, asleep, sister, hang, got home, drunk
Topic 6: school project, teacher, grade, professor, test, students, exam, assignment, final, presentation
Topic 7: morning bed, wake, dishes, chores, baby, asleep, laundry, sleeping, wakes, cleaning
*__Topic 8: marriage__ wedding, married, sister, invited, invite, getting married, fiance, fianc, engaged, ceremony
Topic 9: driving lane, road, traffic, driving, speed, cars, light, lanes, driver, left lane
Topic 10: pets (adoption) dog, cat, dogs, puppy, cats, vet, kitten, pet, shelter, adopted
Topic 11: flight seat, seats, flight, bus, aisle, row, lady, window, plane, attendant
Topic 12: dining tip, restaurant, ordered, waitress, service, table, tipping, waiter, server, cashier
Topic 13: housing mum, rent, sister, flat, uni, bills, university, savings, afford, holiday
*__Topic 14: close relationship (strain)__ baby, sister, pregnant, married, marriage, pregnancy, ex, divorce, abortion, biological
*__Topic 15: relationship__ ex, dating, gf, feelings, best friend, crush, girls, meet, message, hang
Topic 16: engagment/wedding wedding, married, bride, fianc, shower, birthday, reception, getting married, gift, dress
Topic 17: video games game, play, server, character, player, players, campaign, dm, discord, team
*__Topic 18: affairs__ english, gf, meet, sad, friendship, joke, best friend, language, sister, feelings
Topic 19: transactions item, price, seller, shipping, refund, selling, sold, items, ebay, sell
Topic 20: roommates rent, lease, roommate, roommates, apartment, utilities, deposit, landlord, split, bedroom
Topic 21: pregnancy pregnant, baby, sister, pregnancy, grandmother, married, birth, abortion, donor, marriage
Topic 22: pets (ownership) dog, cat, dogs, cats, puppy, shelter, vet, pet, foster, animal
*__Topic 23: expectations__ gift, gifts, birthday, presents, dress, clothes, expensive, sister, wear, list
Topic 24: confrontation lady, walked, seat, store, walking, stall, cashier, door, bus, bathroom
Topic 25: rent rent, lease, apartment, roommate, roommates, bedroom, afford, pay rent, landlord, split
Topic 26: babies names, baby, named, naming, pregnant, nickname, changing, boy, married, michael
Topic 27: bills bills, debt, rent, account, credit, savings, expenses, income, insurance, loan
Topic 28: vacation/holiday vacation, disney, birthday, thanksgiving, sister, wedding, baby, visit, spending, holiday
Topic 29: food eat, pizza, eating, ate, dinner, cook, cream, eaten, cheese, hungry
Topic 30: coworkers boss, company, manager, coworkers, position, office, coworker, department, quit, hr
Topic 31: parking parking, park, spot, street, driveway, spots, parked, cars, neighbor, parking spot
Topic 32: finances savings, income, debt, financial, account, rent, bills, married, financially, marriage
Topic 33: religion church, catholic, religious, religion, christian, god, atheist, mass, baptized, going church
Topic 34: physical health weight, fat, body, eating, pounds, gym, lose weight, eat, diet, healthy
*__Topic 35: neighbors__ neighbors, noise, apartment, loud, building, smoke, music, smell, smoking, door
Topic 36: workplace manager, company, staff, team, boss, role, colleague, colleagues, office, interview
Topic 37: expression women, political, gender, dating, sex, mental, sharing, transgender, female, age
Topic 38: cheating dating, ex, best friend, gf, sex, cheating, feelings, blocked, bc, friendship
*__Topic 39: depression__ fat, depression, hang, depressed, dating, shell, sex, makes feel, feels like, weight
Topic 40: family problems sister, siblings, ex, married, divorce, birth, therapy, abuse, mental, ps
*__Topic 41: cooking__ vegan, meat, eat, cook, dinner, vegetarian, gluten, cooking, meal, eating
Topic 42: drinking drunk, mike, kiss, dude, sex, meet, sam, yeah, dance, ashley
Topic 43: gifts gift, gifts, presents, birthday, santa, secret santa, buying, secret, item, card
Topic 44: romance dating, ex, friendship, feelings, relationships, se, greg, laura, mary, ef
Topic 45: scheduling shift, boss, manager, schedule, shifts, scheduled, saturday, cover, coworker, friday
Topic 46: mental health sister, cancer, abusive, siblings, married, mental, mentally, grandma, grandmother, fil
*__Topic 47: community__ property, fence, yard, neighbors, trash, neighbor, land, neighborhood, cans, driveway

A.4 Top 4 Topic Top Words (RoBERTa Model)

Topic 0: relationship, sex, dating, bf, gay, friendship, ex, x200b, hed, date
Topic 1: husband, relationship, son, wedding, x200b, dog, baby, father, ex, married
Topic 2: relationship, dating, date, ex, sex, friendship, gf, best friend, x200b, mutual
Topic 3: dog, cat, dogs, cats, puppy, vet, pet, shelter, animal, kitten

A.5 Topic Representation

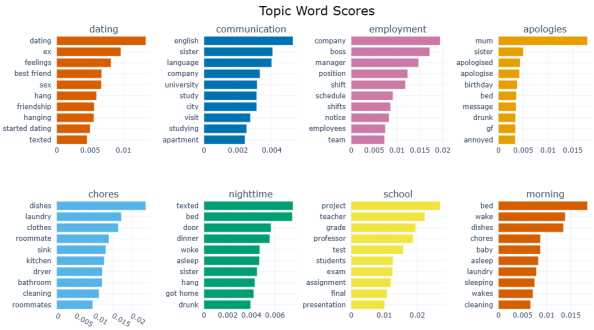

Figure 16: MPNet Top 8 Topics and Top Representative Words

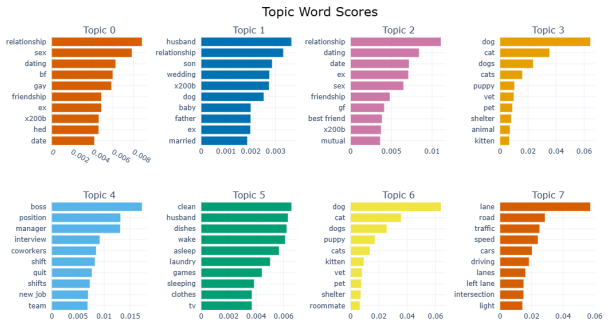

Figure 17: RoBERTa Top 8 Topics and Top Representative Words