Hertie School

Who Argues about What? Topic Modeling and Psychological Profiling on r/ChangeMyView

Research note for GRAD-E1282: Natural Language Processing course

Sofiya Berdiyeva, MDS 2024 (246934)
23/12/2025

Introduction

The internet is a giant machine for disagreement, and the subreddit r/ChangeMyView (CMV) is one of its most organized parts, within which users engage in "good faith" debates with the explicit goal of having their minds changed. 2The original study by (Al Khatib et al., 2020) modeled debaters' prior beliefs and personality traits based on their Reddit history to predict if they would be persuasive or, conversely, resistant to persuasion. Their analysis included allocation of posts to the certain topics, but it was performed through a vaguely described technique of 'identification of Wikipedia entities via entity linking' within the texts (Al Khatib et al., 2020). The current research note will extend the analysis of this study by introducing Latent Dirichlet Allocation (LDA) as an approach that would have higher relevance to initial data and reveal more 'authentic' and comprehensible thematic areas than simple Wikipedia categories. Besides that, we will perform a small downstream task of inter-topic comparison of psychological and linguistic characteristics of the debaters that initially wanted to be persuaded in each thread of CMV subreddit.

Respectively, this leads us to two research questions, with the first one having a higher priority:

1. What are the most prevalent issues and trends discussed in r/ChangeMyView (CMV) subreddit?[1]
2. How do the psychological and linguistic traits of debaters differ across topics they discuss?

Dataset Description

This research relies on the collection of interrelated datasets 'Webis-CMV-20' provided by (Kolyada et al., 2020), which include over 65,169 discussion trees and 3,449,917 posts (discussion nodes) from the foundation of the r/ChangeMyView subreddit in 2005 until September 2017. Later, authors of the dataset sampled a balanced subset of delta-earning and -unearning comments paired based on their length, and extracted information on debators' psycholinguistic traits.

Surprisingly, the data promised in the documentation were significantly different from the ones observed in the files, but they were still sufficient for further analysis. For instance, the main file with personality traits features only contained authors that posted from the period of March 31st, 2019 to April 2nd, 2019, which is a substantial limitation for the scope of both analytical research imagination and results interpretation. For each submission within this period, we merged data on the initial post, several subsequent comments, and psycholinguistic traits of the authors of initial posts (we'd prefer to call them original post (OP) submitters for brevity). Then, we concatenated the posts

[1] With a methodological subquestion: 'How can we derive them in the most optimal way?'

within one discussion tree, starting with the initial post, and got the corpus of 5990 documents for further topic modeling. The step of concatenation was crucial for LDA model performance, since on the shorter version of documents it resulted in poor results in the form of low optimal number of topics (k=2). Even though a focus of discussion might have been slightly shifted from comment to comment, we still assumed and hoped that they would mostly stay within one area of thought.

It is also worth noticing that some of the OP submitters within analyzed corpus were repeated, and number of unique post authors equals to 4895.

LIWC - Linguistic Inquiry and Word Count

The authors of the original paper used LIWC (Linguistic Inquiry and Word Count) (Pennebaker et al., 2015) tool to ‘the first 1000 words extracted from all Reddit posts made by a debater in temporal order’ (Al Khatib et al., 2020). LIWC is a gold-standard text analysis tool in computational linguistics that counts words in specific categories to determine the emotional, cognitive, and structural components of a text. Here, these scores are used to model the personality traits and linguistic style of OP submitters. The dataset provided numerous indices produced by the tool, grouped by the following categories: summary language variables, linguistic dimensions and psychological processes. For our analysis, we have chosen 6 dimensions measured for the OP submitters – three summary variables and three low-abstraction-level psychological affective processes:

1. Analytical thinking: higher scores indicate more formal, logical, and hierarchical thinking in contrast to narrative, personal, or intuitive one. Besides that, higher scores are associated (at the linguistic level) with more complex, abstract, and structured expression; lower scores tend to reflect more informal, here-and-now, story-like language that is closer to spoken conversation.
2. Authentic: how honest, personal, and disclosing the text is, high scores correlate with authenticity and low scores with more guarded or "distanced" language;
3. Emotional tone: a high score (above 50) indicates a more positive or upbeat tone, a low score (below 50) means a more negative or anxious tone. Here positive and negative emotion usage is combined into a single scale;
4. Anxiety, having 116 words in dictionary indicating it (e.g. ‘worried’, ‘fearful’);
5. Anger, with dictionary of 230 words like ‘hate’, ‘kill’, ‘annoyed’;
6. Sadness, identified by set of 136 words including ‘crying’, ‘grief’, ‘sad’.

The first three features are roughly measured on the scale from 1 to 99, being standardized percentile scores, and last three are expressed through simple percentages of total words that belong to a category, which makes their amplitude relatively small.

Methods

Text preprocessing

Text data were preprocessed through a custom pipeline to produce a bag-of-words representation suitable for Latent Dirichlet Allocation (LDA), reflecting the model's assumption that documents can be represented as unordered collections of words rather than syntactically structured text. Since topic modeling does not require detailed grammatical or semantic annotation and benefits from computational efficiency at scale, English language resources were loaded from spaCy ("en_core_web_sm") with all syntactic and semantic components disabled to retain only lightweight tokenization, while NLTK was used to supply English stopwords and a WordNet-based lemmatizer.

To ensure that word co-occurrence statistics reflect semantic content rather than orthographic variation, all documents were lowercased and tokenized with spaCy, after which tokens consisting of punctuation, whitespace, or numbers were removed. Given that function words and contractions do not contribute meaningfully to latent topics and can skew frequency distributions, we further filtered tokens by excluding standard English stopwords, a custom list of contraction remnants (e.g., "'s", "n't"), and any tokens with leading or trailing apostrophes. To reduce vocabulary noise and sparsity, which can destabilize topic inference, only alphabetic tokens were retained.

Because LDA is sensitive to vocabulary size and benefits from collapsing inflectional variants that share semantic meaning, remaining tokens were lemmatized using the WordNet lemmatizers for verbs, nouns, adjectives and adverbs. While this strategy may incorrectly transform or fail to normalize nouns and adjectives, resulting in uneven lemmatization quality across grammatical categories, we consider it sufficient for the current project. Finally, to emphasize robust cross-document word distributions while suppressing uninformative extremes, the resulting tokens were used to construct a document-term matrix via scikit-learn's 'CountVectorizer', applying document frequency thresholds to remove extremely common terms (max_df = 0.90) and very rare terms (min_df = 0.005). This process yielded a sparse bag-of-words matrix used as input for the further LDA topic modeling.

LDA

To identify the core themes of the debates, we used Latent Dirichlet Allocation (LDA). This method was chosen for its higher interpretability over Latent Semantic Analysis and lower computational demands than more advanced techniques like BERTopic. Rather than just guessing the number of topics, we conducted a grid search across several possibilities (see Appendix I) ranging from 3 to 12 topics, and different combinations of alpha and beta parameters. We evaluated the models using Perplexity (how well the model predicts the sample) and Coherence scores (how well the words in a topic actually make sense together). Based on these metrics, we settled on 9 topics as the optimal balance between detail and interpretability (see the table in Appendix II), as it has higher coherence than its neighbors 8 and 10, and lower perplexity than cases when number of topics is lower than 6. The alpha parameter was equal to 0.1, while beta was set to 0.2.

Results

Topic modeling

During the analysis, we came up with the following list of topics (Figure 1).

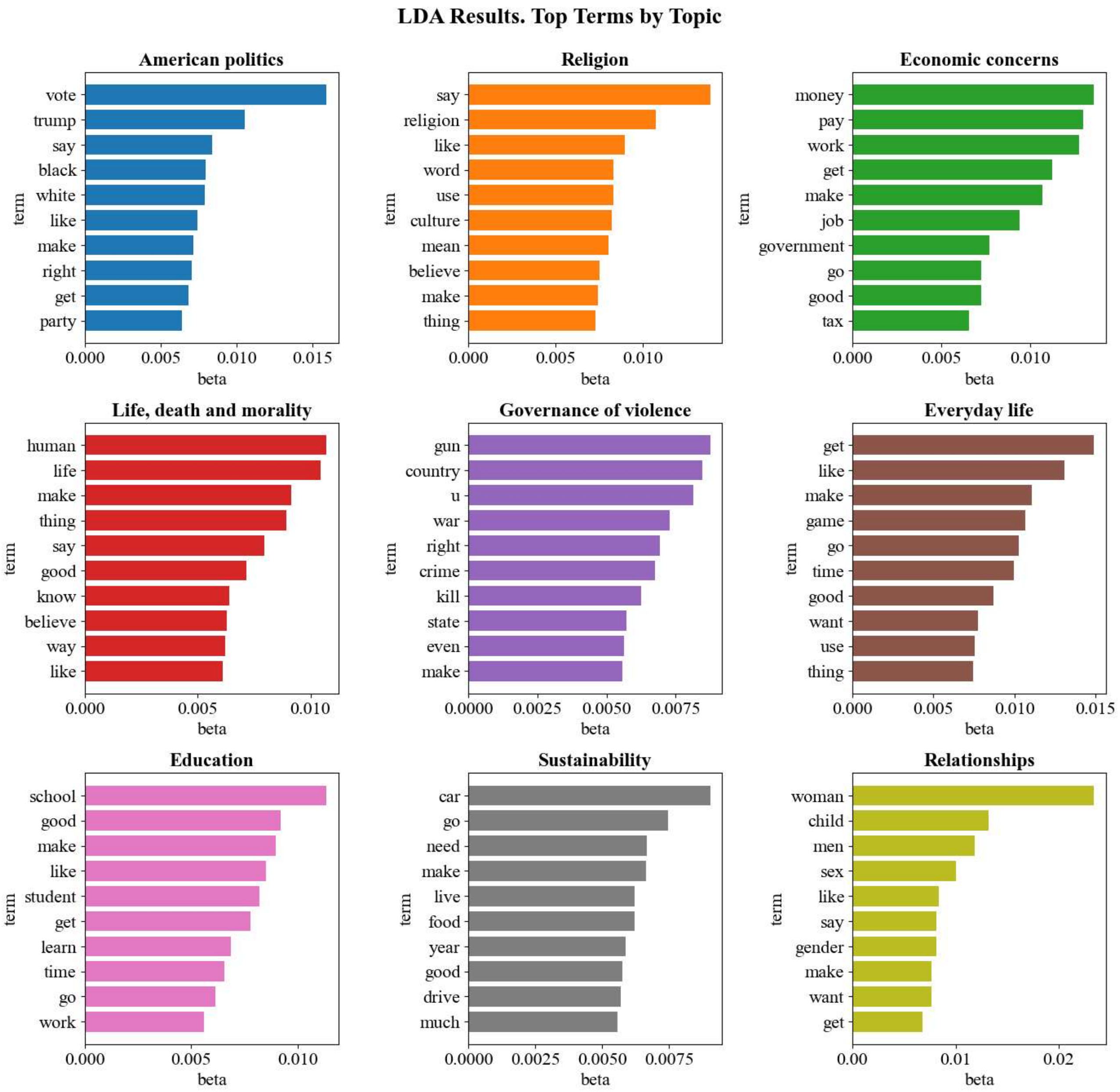


Figure 1. Words with the highest probabilities of belonging to topics (betas (β)) across topics.

The first topic was present in at least 760 posts from the corpus[2]. It is mainly dedicated to American politics: elections, Trump, parties and partially one of the most frequently discussed themes in this domain – racism.

The second topic, religion, included about 530 texts. Out of all religions, only 'Christian' appeared in the top 50 words, which partially represents bias present in the data due to the Reddit audience demographics and backgrounds.

Topic #3 (ca 681 post), included discussions of governmental regulations of taxes, monetary system, companies and businesses, and their impact on society. One slightly strange outlier in the dictionary for the topic was the word 'drug', which we leave for the reader's interpretation.

The fourth topic, comprising 771 post, turned out to be more philosophical (but not in academic sense), primarily focusing on the issues of humanism, morality, ethics and deontology.

Topic number five (617 posts) was one of the most intense and filled with the domain-specific terms: gun, war, right, crime, kill, state, law, police, government, military, violence, power, fight, force, case, shoot, criminal, death, victim, and so on. Taking into account the presence of the words 'government', 'state' and 'law' and broadness of scope, we decided to frame the topic as referring to violence regulation.

The sixth topic (816 texts) was the hardest to name, as it didn't quite have any specificity and concentration on a certain area. The only common traits of the words within it were their neutrality and reference to simple everyday life concepts (game, time, drink, friend, play), so we ended up naming it respectively.

Topic #7 (615 posts) was relatively easy to identify as education-oriented. The vocabulary is dominated by words referring to learning environments and practices, such as school, student, class, teacher, college, test, and study. At the same time, the topic extends beyond institutional education, incorporating broader cultural and developmental aspects through terms like art, music, sport, movie, character, and skill.

Topic eight with 329 posts can be interpreted as centered on sustainability and long-term societal challenges. Although smaller in size than most other topics, it shows a clear thematic coherence around environmental, technological, and demographic issues. Key terms such as climate, resource, water, population, global, technology, and system point to discussions of ecological limits and systemic change.

Ninth topic #9 (871 posts) was the largest of the remaining topics and focused on relationships and gender-related issues. The word list strongly emphasizes social roles and identities, with frequent

[2] For each text, LDA provided the topic vectors with scores for allocation. Even though some of the texts had relatively high scores for the multiple topics at once, we calculated final text-topic allocation based on the highest score in the distribution for a given text, so that total number of documents across all topics is equal to number of documents in the corpus (5990).

references to woman, men, gender, male, female, parent, and child. Some terms indicated that discussions often revolved around sexuality, reproduction, and contemporary debates on gender identity.

Topics and psycholinguistic traits

To examine whether the psychological and linguistic traits of original post submitters differed across discussion topics, we analyzed six LIWC-based dimensions capturing analytical thinking, authenticity, emotional tone, and low-level affective processes (anxiety, anger, sadness). For all six variables, Shapiro-Wilk tests indicated that the majority of topic-specific distributions significantly deviated from normality. In addition, Levene's tests consistently rejected the assumption of homogeneity of variances. Consequently, non-parametric Kruskal-Wallis tests were used throughout (Figure 2). In all cases, the omnibus tests revealed statistically significant differences between topics (all $p < .001$), indicating systematic variation in what kind of people were discussing different topics.

Analytical thinking

Analytical thinking differed strongly across topics ($H = 282.17$, $p < .001$). Topics related to governance of violence and sustainability exhibited the highest mean analytical thinking scores (both around 57-58), suggesting a more formal, logical, and structured linguistic style of writing of their authors. Economic concerns also scored relatively high, aligning with its policy- and system-oriented focus. In contrast, relationships and life, death, and morality showed the lowest analytical thinking scores (both around 49), reflecting slightly more narrative, experiential, and less hierarchical language usage by submitters.

Authenticity

Authenticity scores also varied significantly by topic ($H = 145.42$, $p < .001$), with none of the topic-specific distributions approximating normality. The most authentic language was found in everyday life, life, death, and morality, and education (means = 35-37), suggesting higher levels of self-disclosure and personal engagement of respective text creators. In contrast, American politics, economic concerns, and governance of violence exhibited noticeably lower authenticity scores (= 30-32), indicating more distanced, guarded, or impersonal language used by the authors.

Emotional tone

Emotional tone showed some of the strongest between-topic differences ($H = 331.99$, $p < .001$). Overall, most topics fell below the neutral midpoint of 50, indicating a generally negative or tense tone across the corpus authors. Education and economic concerns were closest to a neutral-to-positive tone (means = 49 and 47, respectively) language of submitters, while governance of violence stood out as written by the most negatively inclined creators by a large margin (mean = 32).

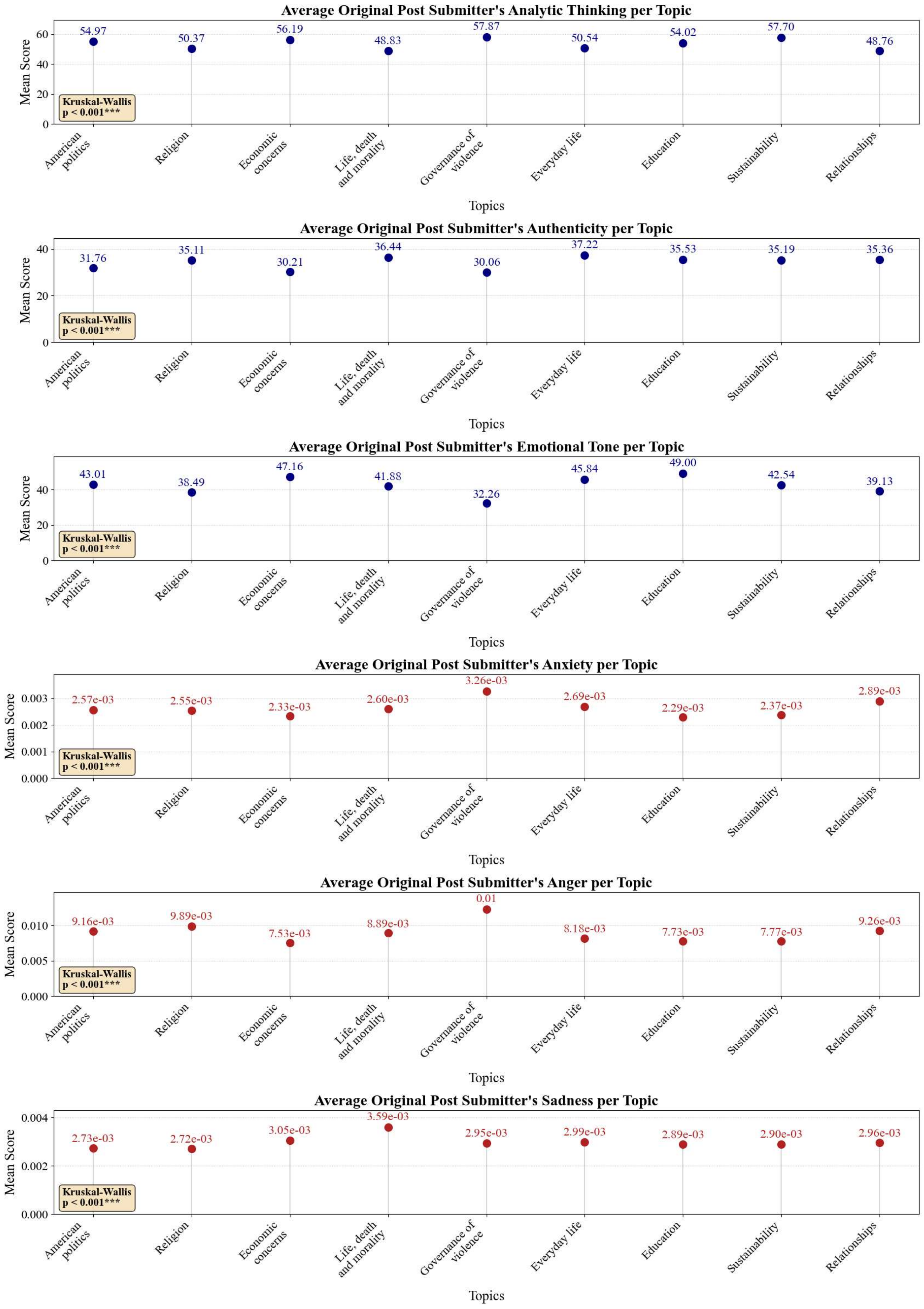


Figure 2. LIWC features of OP submitters across primary topics of posts

Anxiety

All topic distributions for anxiety were markedly non-normal, with small absolute values reflecting the low base rate of anxiety-related words. Nonetheless, differences were statistically significant (H = 120.17, p < .001). Governance of violence displayed the highest anxiety levels among debates creators (mean = 0.0033), followed by relationships and everyday life. By contrast, people discussing education and economic concerns showed the lowest anxiety scores (= 0.0023).

Anger

Anger varied strongly across topics (H = 328.98, p < .001), again with all distributions deviating from normality. As expected, governance of violence exhibited the highest anger scores (mean = 0.012) of the texts authors, substantially exceeding all other topics. Religion, relationships, and American politics formed a second tier with elevated anger levels of Redditors, whereas people expressing opinions on economic concerns, education, and sustainability showed the lowest anger scores. This distribution highlights the affective intensity of morally and politically polarized domains compared to more technocratic discussions.

Sadness

Even though effect sizes were smaller than for anger or emotional tone, sadness also differed significantly between topics (H = 55.78, p < .001). Life, death, and morality was clearly associated with the highest sadness levels (mean = 0.0036) among people discussing them, consistent with its existential focus. Economic concerns, everyday life, and relationships showed moderately elevated sadness of debators, while American politics and religion exhibited slightly lower values.

Taken together, these results demonstrate that Reddit discussion topics are associated with distinct psychological and linguistic profiles of OP submitters. Abstract, system-oriented topics tend to be discussed by people with more analytical but less authentic language, while personal and existential topics are accompanied by greater self-disclosure and emotional complexity. In addition, debaters on violence and conflict stand out as particularly negatively inclined, angry, and anxious.

Discussion

Our findings both support and refine the original study by (Al Khatib et al., 2020). While they proved that debaters' characteristics matter for persuasiveness, we show how they might potentially be related to debater's engagement with specific topic discussions. The original paper focused on the overall "persuadability" of OP submitters; however, our analysis of the 9 topics reveals that certain psychological traits could be "niche-specific", or vice versa.

The original paper noted that modeling psycholinguistic traits enhances persuasiveness prediction accuracy by several percentage points. Our extension suggests this is because persuasive debaters "self-select" into topics that match their psychological strengths. Future work should perhaps consider not just the author’s traits, but the congruence between the author’s personality and the topic’s demands.

Our LDA model could later be used for updating a regression model with a “topic” predictor, to take into account the domain specificities of persuasion. We would also consider applying sentiment analysis not only to the initial post to reveal the stance of the persuaded person (as was performed in the main referenced paper), but also on persuasive comments in order to have a look at relationships of sentiment scores and persuasion outcomes. And, we would have been curious about average sentiment scores within the topics revealed in the current project, as we expect some of them (regarding death, morality and violence) to be severely negative, and others more neutral.

References

Al Khatib, K., Völske, M., Syed, S., Kolyada, N., & Stein, B. (2020). Exploiting Personal Characteristics of Debaters for Predicting Persuasiveness. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 7067–7072). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.632

Kolyada, N., Al-Khatib, K., Völske, M., Syed, S., & Stein, B. (2020). *Webis ChangeMyView Corpus 2020 (Webis-CMV-20)* [Dataset]. Zenodo. https://doi.org/10.5281/zenodo.3778298

Pennebaker, J., Booth, R., Boyd, R., & Francis, M. (2015). *Linguistic Inquiry and Word Count: LIWC2015*.

Files related to the current work are located here:
https://github.com/sophiyaberdiyeva/nlp_hertie_final_project

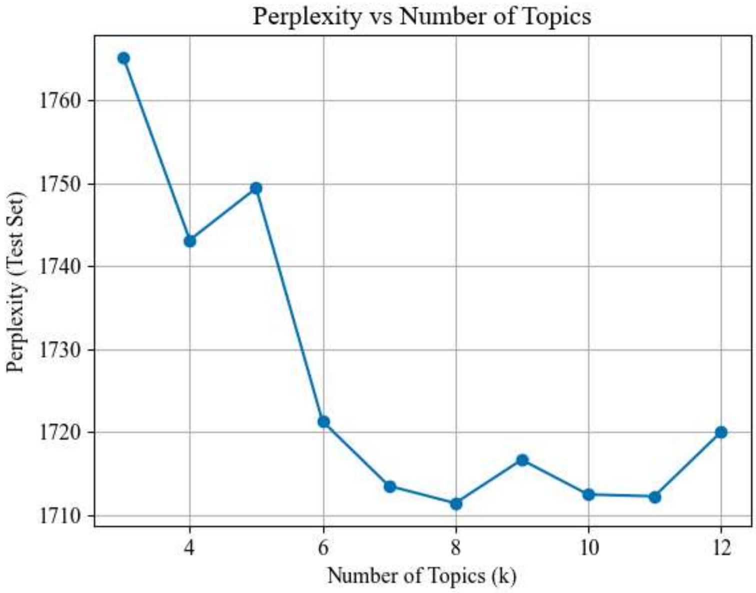
Code was partially taken from the labs materials and is authored by Sascha Goebel and

Luis Fernando Ramirez Ruiz.

AI tools like Claude, ChatGPT and Gemini were used during code creation and debugging, as well as during report preparation for language improvements. NotebookLM was used for faster information retrieval from the main reference paper.

Appendix I

Perplexities across numbers of topics



Appendix II

Hyperparameter tuning results for top-10 best models

k_values = [3, 4, 5, 6, 7, 8, 9, 10, 11, 12]

alpha_values = [None, 0.1]

beta_values = [None, 0.15, 0.2]

Number of Topics	Alpha	Beta	Perplexity	Coherence
3	0.1	0.2	1765.14	0.65
4	0.1	0.2	1743.12	0.53
5	0.1	0.2	1749.37	0.45
6	0.1	0.2	1721.27	0.43
7	0.1	0.2	1713.47	0.41
8	0.1	0.2	1711.39	0.42
9	0.1	0.2	1716.62	0.44
10	0.1	0.2	1712.44	0.40
11	0.1	0.15	1712.21	0.43
12	0.1	0.15	1719.95	0.37