COMPUTATIONAL ANALYSIS OF REDDIT DISCOURSE ON UYGHUR TOPICS: SENTIMENT, TOPICS, AND COMMUNITY PERSPECTIVES

Executive Summary

This project investigates the discourse surrounding Uyghur topics on Reddit through comprehensive data mining and natural language processing techniques. A corpus of 16,967 items (1,200 posts and 15,767 comments) was collected from multiple subreddits using terms related to Uyghurs, Xinjiang, and related topics. The analysis reveals distinct topic clusters, sentiment distributions, and linguistic patterns across different communities. The findings demonstrate how social media discourse about the Uyghur situation varies by community, content type, and time period, highlighting the polarized nature of the discussion and the predominance of political framing in these conversations.

1. Introduction

1.1 Background and Motivation

The Uyghur situation in Xinjiang, China has become a significant geopolitical and human rights topic in global discourse. Social media platforms like Reddit provide a valuable lens through which to examine public perspectives, sentiment patterns, and discourse framing around this sensitive issue. This analysis seeks to understand how Reddit users discuss, frame, and respond to topics related to Uyghurs across different online communities.

1.2 Project Evolution

This project began with a different approach to sentiment analysis. The initial proposal focused on analyzing an Uyghur-to-English dataset from a GitHub repository for sentiment analysis of the Uyghur diaspora. However, preliminary analysis revealed significant limitations in the dataset - virtually every word in the corpus was a noun, severely restricting the potential for meaningful sentiment analysis which typically relies on adjectives, adverbs, and verbs to detect emotional valence.

Faced with this methodological challenge, the project pivoted to a more robust approach: scraping and analyzing Reddit posts and comments related to Uyghur topics. This adaptation allowed for a more comprehensive sentiment analysis with richer linguistic features, capturing how different online communities discuss and frame Uyghur-related topics. The pivot demonstrates the importance of flexibility in data mining projects and the need to critically evaluate data sources before proceeding with complex analyses.

1.3 Project Objectives

The primary objectives of this analysis were to:

1. Collect and analyze a comprehensive dataset of Reddit content related to Uyghur topics 2.

Identify key topics and themes in the discourse using topic modeling techniques 3. Analyze sentiment patterns across different subreddits, content types, and time periods 4. Visualize and interpret the relationship between topics, sentiment, and community dynamics
5. Extract actionable insights about how this issue is discussed in online spaces

1.4 Research Questions

The analysis was guided by the following research questions:

1. What are the main topics and themes in Reddit discussions about Uyghurs? 2. How does sentiment vary across different subreddits and discussion topics? 3. Are there distinct linguistic patterns associated with different Reddit communities? 4. How has the discourse around Uyghur topics evolved over time?

2. Methodology

2.1 Data Collection

Data was collected from Reddit using the PRAW (Python Reddit API Wrapper) library. The collection targeted multiple subreddits, including general news/politics subreddits (worldnews, news, politics), region-specific subreddits (China, Sino, Xinjiang), and topically relevant communities (Uyghur, FreeTheUyghurs, fucktheccp, islam).

The search queries included terms such as:

- "Uyghur" and "Uighur" (common spellings)
- "Xinjiang" and "East Turkestan" (geographical references)
- Specific topic terms: "Uyghur genocide," "Uyghur camps," "re-education camp"
- Related phrases: "forced labor Uyghur," "Uyghur persecution"

The complete dataset comprises **16,967 items**, including:

- 1,200 posts (7.07%)
- 15,767 comments (92.93%)

2.2 Data Preprocessing

Text preprocessing included:

- Conversion to lowercase
- Removal of URLs, Reddit-specific formatting, and special characters
- Tokenization using NLTK's word_tokenize
- Stopword removal (English stopwords plus Reddit-specific terms)
- Lemmatization using WordNetLemmatizer

2.3 Analytical Approaches

Exploratory Data Analysis:

- Distribution analysis of content across subreddits
- Word frequency analysis and visualization
- Content type and length analysis

Sentiment Analysis:

- TextBlob for basic sentiment polarity and subjectivity scoring
- Transformer model (DistilBERT) for advanced sentiment classification
- Sentiment categorization (positive, negative, neutral)

Topic Modeling:

- TF-IDF vectorization (1000 features)
- Latent Dirichlet Allocation (LDA) with 5 topics
- Non-negative Matrix Factorization (NMF) for comparison
- Topic interpretation and labeling

Dimensionality Reduction and Visualization:

- t-SNE for document embedding visualization
- Document clustering by topic, sentiment, and content type

3. Exploratory Data Analysis

3.1 Dataset Composition

The data came from a diverse set of subreddits, with the highest representation from:

1. r/China (2,478 items)
2. r/news (2,278 items)
3. r/worldnews (2,087 items)
4. r/Sino (1,835 items)
5. r/geopolitics (1,805 items)

This distribution demonstrates that discussions about Uyghur topics occur primarily in geopolitical and news subreddits, but also in specific China-focused communities.

3.2 Content Characteristics

Analysis of content length revealed significant differences between posts and comments:

- Comments: Average length of 260.9 characters (median 119)
- Posts: Average length of 475.7 characters (median 95)

While posts have a higher mean length, this is skewed by some very long posts (max 38,564 characters), whereas comments typically have more consistent lengths. This suggests that while initial posts sometimes contain in-depth information, most engagement occurs through relatively brief comments.

3.3 Word Frequency Analysis

The most common words across the entire dataset were:

1. "china" (appearing approximately 8,000 times)
2. "people" (approximately 3,500 times)
3. "chinese" (approximately 2,500 times)
4. "u" (user references)
5. "like"
6. "xinjiang"
7. "country"

This indicates that discussions primarily frame the topic in terms of China as a nation and its relationship to the Uyghur people, with a focus on national and ethnic identity.

Word frequency analysis by subreddit revealed distinct linguistic patterns:

- r/China: "china," "people," "chinese," "xinjiang," "genocide"
- r/news: "china," "people," "chinese," "like," "u"
- r/worldnews: "china," "people," "u," "country," "like"

3.4 Temporal Patterns

The temporal analysis showed significant increases in Uyghur-related content during certain periods. Notably, there was an initial uptick in posts and comments beginning around 2017, which corresponds with early international reporting on the situation in Xinjiang. Activity then increased substantially during 2020-2022, with peak activity in 2021. This timeline correlates with increased international attention to the situation, including policy decisions and diplomatic statements from various countries. The 2017 initial surge is particularly significant as it coincides with the earliest reports of mass detentions in Xinjiang, suggesting that Reddit discussions were responsive to emerging news about the situation.

4. Topic Modeling Results

4.1 LDA Topic Model

The LDA analysis identified five distinct topics, which were labeled based on their most representative terms:

1. **Political and Religious Context**
    - ○ Key terms: another, free, allah, may, party, china, tibet, chinese, video, communist ○ This topic appears to focus on the religious aspects of the situation and the broader

political context.

2. **International Relations and Politics**
    - Key terms: china, ccp, country, fuck, would, trump, people, america, war, really
    - This topic centers on international political dimensions and U.S.-China relations.
3. **Casual Discussion and Opinions**
    - Key terms: like, lol, know, people, look, good, didnt, never, dont, theyre
    - Represents more informal discussion and personal opinions.
4. **Core Uyghur Crisis Discussion**
    - Key terms: china, nan, xinjiang, genocide, uighur, uyghur, camp, chinese, right, muslim
    - Directly addresses the alleged human rights abuses and situation in Xinjiang.
5. **Meta-Discussion and Moderation**
    - Key terms: please, question, bot, concern, comment, action, rule, let, contact, pooh
    - Reflects Reddit's self-moderation and meta-commentary about the discussion itself.

The dominant topic in the dataset was Topic 4 (Core Uyghur Crisis Discussion) with 6,861 documents, followed by Topic 2 (International Relations) with 4,182 documents.

4.2 NMF Topic Model

The NMF model was applied for comparison and revealed slightly different topic clusters:

1. **Political and Religious Context**
    - Key terms: people, like, chinese, country, dont, muslim, would, one, even, think
2. **International Relations and Politics**
    - Key terms: nan, xinjiang, uighur, uyghur, camp, chinese, uyghurs, forced, muslim, labor
3. **Casual Discussion and Opinions**
    - Key terms: please, question, bot, moderator, contact, automatically, performed, concern, action, rule
4. **Core Uyghur Crisis Discussion**
    - Key terms: china, fuck, world, country, taiwan, un, war, uyghurs, state, part
5. **Meta-Discussion and Moderation**
    - Key terms: genocide, uyghur, cultural, israel, committing, call, group, evidence, stop, committed

Comparison of LDA and NMF results showed significant overlap but also some distinctions in how the models clustered the content, with LDA generally providing more interpretable topics for this dataset.

4.3 Comparison of Topic Models and Methodological Insights

The comparison between LDA and NMF topic models revealed interesting methodological insights:

1. **Divergent Content Classification**: The heatmap comparison of topic assignments shows significant disagreement between the models, particularly for Topics 1, 2, and 4, which were classified differently by LDA and NMF. This suggests that the discourse contains overlapping themes that can be interpreted differently depending on the analytical approach.

2. **Algorithmic Differences**: LDA performed better at identifying cohesive topics related to the core Uyghur situation, while NMF more effectively isolated meta-discussion and moderation content. This demonstrates how different algorithms capture different aspects of textual patterns.
3. **Topic Stability Analysis**: The cross-model comparison provides a form of validation for topics that were consistently identified by both approaches (particularly aspects of the Core Uyghur Crisis Discussion), suggesting these represent more stable and distinct discourse patterns.

4. **Methodological Complementarity**: The differences between models highlight the value of applying multiple topic modeling approaches when analyzing complex, multifaceted discourse on politically sensitive topics.# Computational Analysis of Reddit Discourse on Uyghur Topics

5. Sentiment Analysis

5.1 Overall Sentiment Distribution

The sentiment analysis revealed a mixed but predominantly neutral sentiment across the dataset:

- Neutral: 8,927 items (52.6%)
- Positive: 5,095 items (30.0%)
- Negative: 2,945 items (17.4%)

This distribution suggests that while many discussions maintain a neutral, informational tone, there is a slightly higher tendency toward positive framing than negative framing. This finding is somewhat counterintuitive given the subject matter's gravity, suggesting that discourse framing and community norms significantly influence how emotionally charged topics are discussed on Reddit.

5.2 Sentiment by Content Type

Comparison of sentiment between posts and comments showed interesting patterns:

- Comments had a higher percentage of positive sentiment (31.7%) compared to posts (24.2%)
- Posts had a higher percentage of neutral content (65.7%) compared to comments (49.4%) ● Negative sentiment was relatively consistent between posts (17.5%) and comments (18.9%)

This suggests that while initial posts tend to be more neutrally framed, the ensuing discussion in comments often develops more polarized sentiment.

5.3 Sentiment by Subreddit

Significant variations in sentiment were observed across different subreddits:

- r/Sino showed the most positive sentiment (average polarity of +0.089)
- r/geopolitics and r/China had relatively neutral sentiment (average polarity near +0.05)

- r/humanrights had the most negative sentiment (average polarity of +0.017)

The generally positive sentiment in r/Sino (a subreddit often characterized as pro-Chinese government) compared to more negative sentiment in human rights focused subreddits suggests divergent framing of the same events across different communities.

5.4 Sentiment by Topic

Analysis of sentiment across topics revealed that:

- Topic 3 (Casual Discussion) had the most positive sentiment (34.3% positive) ● Topic 2 (International Relations) had the most negative sentiment (23.4% negative) ● Topic 5 (Meta-Discussion) had the highest proportion of neutral content (62.8% neutral)

This indicates that more casual discussions tend to have more positive framing, while discussions of international politics and relations tend to be more negative.

6. Advanced Visualization Results

6.1 t-SNE Document Clustering

The t-SNE visualization revealed several patterns:

- **Topic Clustering**: Documents clustered according to topic, with distinct regions visible for Core Uyghur Crisis Discussion and Meta-Discussion topics.

- **Sentiment Distribution**: No clear spatial clustering by sentiment was observed, suggesting that sentiment is not strongly correlated with specific vocabulary patterns but rather cuts across different topical discussions.

- **Content Type Differentiation**: Clear separation between posts and comments was visible, with posts appearing as a distinct cluster within the broader comment distribution. This suggests fundamental linguistic differences between initial posts and responsive comments.

- **Vocabulary-Based Clustering**: The visualization shows that vocabulary usage, rather than sentiment or subreddit origin, is the strongest determinant of document similarity. This indicates that particular terminology and framing devices are consistently used across different communities when discussing specific aspects of the Uyghur situation.

6.2 Sentiment-Topic-Subreddit Relationships

Heatmap analysis of positive sentiment by topic and content type showed that:

- Comments discussing Topic 3 (Casual Discussion) had the highest positive sentiment (34.3%)
- Posts discussing Topic 4 (Core Uyghur Crisis) had the lowest positive sentiment (15.4%)

The topic distribution by subreddit visualization demonstrated that:

- r/Sino primarily featured Topic 1 (Political and Religious Context)
- r/worldnews and r/news had more balanced topic distributions
- r/Uyghur was dominated by Topic 4 (Core Uyghur Crisis Discussion)

7. Findings and Insights

7.1 Key Findings

1. **Polarized Community Perspectives**: Different Reddit communities frame the Uyghur situation in distinctly different ways, with some communities (like r/Sino) maintaining more positive sentiment while others (like r/humanrights) skew negative.

2. **Dominance of Political Framing**: The discourse is primarily framed in political terms, with a strong focus on nation-states, international relations, and governmental actions rather than individual experiences or cultural aspects.

3. **Temporal Evolution**: Discussion volume showed an initial uptick in 2017, corresponding with early reports of mass detentions, followed by a major surge during 2020-2022 with peak activity in 2021. This pattern suggests that Reddit discourse closely follows major media reports and policy developments, with the community responding quickly to emerging news.

4. **Topic-Sentiment Relationship**: Casual discussions tend to have more positive sentiment, while discussions of international relations and politics trend negative, highlighting how framing affects emotional valence.

5. **Content Type Differences**: Posts tend to be more neutral and information-focused, while comments show greater sentiment polarization, demonstrating how community engagement affects discourse framing.

6. **LDA-NMF Topic Divergence**: The comparison between LDA and NMF topic models showed significant disagreement in document classification (visible in the heatmap comparison), suggesting that the discourse contains overlapping themes that can be interpreted differently depending on the analytical approach. This methodological insight highlights the complexity of the discourse.

7. **Subreddit-Topic Correlations**: Certain topics show strong associations with specific subreddits - for example, r/Sino predominantly features Political and Religious Context topics, while r/Uyghur focuses on Core Uyghur Crisis Discussion. This demonstrates how community focus shapes the framing of discussions.

7.2 Implications

These findings have several implications for understanding online discourse on sensitive geopolitical topics:

1. **Echo Chamber Effects**: The pronounced differences in sentiment between subreddits suggest the presence of echo chambers where particular viewpoints are reinforced.
2. **Political Polarization**: The dominance of politically-framed topics indicates that the human dimension of the situation may be overshadowed by geopolitical considerations.

3. **Media Influence**: The correlation between temporal patterns and international policy events suggests strong influence of mainstream media coverage on Reddit discourse.

4. **Linguistic Patterns**: The distinct vocabulary and sentiment patterns across communities could be useful for identifying and characterizing different discourse communities in future research.

8. Limitations

This analysis has several limitations that should be acknowledged:

1. **Data Representativeness**: Reddit demographics are not representative of the general population, skewing younger, more male, and more Western.

2. **Language Limitations**: The analysis was limited to English-language content, excluding perspectives expressed in other languages.

3. **Sentiment Analysis Accuracy**: Sentiment analysis tools have limitations in capturing nuance, sarcasm, and complex emotional expressions, particularly for politically charged topics. The comparison between TextBlob and transformer-based sentiment analysis showed only 26% agreement, highlighting the challenges in sentiment classification for complex geopolitical discussions.

4. **Time Period**: The data primarily covers 2017-2024, with concentration in 2020-2022, and may not capture longer-term discourse patterns.

5. **API Limitations**: Reddit's API restrictions limit the comprehensiveness of data collection, potentially missing some relevant content.

6. **Topic Model Stability**: The divergence between LDA and NMF models suggests that topic boundaries are not clearly defined in this discourse, indicating potential instability in the topic modeling results.

9. Conclusions and Recommendations

9.1 Conclusions

This analysis demonstrates the complex and multifaceted nature of online discourse surrounding the Uyghur situation. The findings reveal that discussions are primarily framed in political terms rather than humanitarian ones, with distinct communities developing markedly different perspectives and sentiment patterns. The temporal patterns suggest that Reddit discourse is highly responsive to external media and political events, serving as a barometer of public attention to this issue.

The project's evolution from a limited Uyghur-English dictionary dataset to a rich social media corpus highlights an important methodological insight: the quality and characteristics of the source data fundamentally shape the potential depth and validity of sentiment analysis. The pivot to Reddit content allowed for a more nuanced examination of sentiment, capturing emotional valence across different contexts and communities that would have been impossible with the original dataset.

The topic modeling successfully identified five key discussion themes, highlighting how the conversation ranges from direct discussion of alleged human rights abuses to broader geopolitical considerations and religious contexts. The sentiment analysis revealed generally more neutral-to-positive content than might be expected given the subject matter, suggesting complex framing effects across different communities.

9.2 Recommendations for Future Research

Based on this analysis, several avenues for future research are recommended:

1. **Cross-Platform Comparison**: Extend the analysis to other social media platforms to compare discourse patterns across different online environments.

2. **Longitudinal Analysis**: Conduct more detailed time-series analysis to track how specific events influence discourse patterns and sentiment.

3. **Network Analysis**: Implement user interaction network analysis to better understand information flow and influence patterns in these discussions.

4. **Multilingual Approach**: Incorporate content in multiple languages, particularly Chinese, to capture a more diverse range of perspectives.

5. **Narrative Analysis**: Apply more sophisticated natural language processing techniques to identify and characterize specific narrative frames around this topic.

9.3 Practical Applications

The insights from this analysis could be applied in several contexts:

1. **Media Literacy**: Helping users understand how different communities frame the same issues

differently.

2. **Policy Communication**: Assisting policymakers in understanding public discourse around sensitive international issues.
3. **Content Moderation**: Informing platform policies around politically sensitive content and potential misinformation.

4. **Academic Research**: Contributing to broader understanding of how social media shapes discourse around global human rights issues.

Appendix: Technical Implementation

The analysis was implemented in Python using the following key libraries:

- **Data Collection**: PRAW
- **Data Processing**: Pandas, NumPy
- **NLP Processing**: NLTK, Spacy, TextBlob
- **Topic Modeling**: Scikit-learn (TfidfVectorizer, LatentDirichletAllocation, NMF)
- **Sentiment Analysis**: TextBlob, Hugging Face Transformers
- **Visualization**: Matplotlib, Seaborn, WordCloud
- **Dimensionality Reduction**: t-SNE from Scikit-learn

The implementation followed these primary steps:

1. Data collection via Reddit API
2. Text preprocessing and feature extraction
3. Topic modeling with LDA and NMF
4. Sentiment analysis with multiple approaches
5. Dimensionality reduction for visualization
6. Statistical analysis of relationships between topics, sentiment, subreddits, and content types

The complete code implementation is available in the accompanying Jupyter notebook.