Synthetic Social Alienation: The Role of Algorithm-Driven Content in Shaping Digital Discourse and User Perspectives

# Abstract

This study investigates how algorithm-driven content curation impacts mediated discourse, amplifies ideological echo chambers and alters linguistic structures in online communication. While these platforms promise connectivity, their engagement-driven mechanisms reinforce biases and fragment discourse spaces, leading to Synthetic Social Alienation (SSA). By combining discourse analysis with in-depth interviews, this study examines the algorithmic mediation of language and meaning in digital spaces, revealing how algorithms commodify attention and shape conversational patterns. This study also categorizes participant comments as positive, negative, and neutral using sentiment analysis and examines the emotional tone of these comments. Our hybrid approach combining original interview data with synthetic SSA-focused data achieved excellent performance with Logistic Regression (Accuracy: 87.1%, ROC-AUC: 0.983) and Random Forest (Accuracy: 84.3%, ROC-AUC: 0.984). Cross-validation scores of 0.940 (±0.065) and 0.942 (±0.052) respectively indicate robust model training and generalization capability. Statistical significance tests (McNemar's test: p=0.023, Paired t-test: p=0.001) confirm the superiority of our approach. The findings highlight the need for regulatory interventions and ethical algorithm design to mitigate discourse polarization and restore critical engagement in digital public spheres.

# 3. Methodology

### 3.2.2 Synthetic Data Generation: Comprehensive Methodology

**Synthetic data generation employed a rigorous, multi-stage methodology:**  
  
Generation Tool: Manual template-based generation with linguistic expertise  
Generation Rules:  
• SSA theoretical framework-based thematic categories  
• Turkish grammar compliance  
• Reflection of real user expression linguistic features  
• Balanced distribution across sentiment classes

**Quality Control Process:**  
• Linguistic expert review  
• SSA keyword density analysis  
• Sentiment consistency validation  
• Cross-validation with original data patterns

**Validation Metrics:**  
• SSA keyword prevalence: 13.7%  
• Sentiment distribution balance  
• Linguistic coherence scores  
• Thematic relevance validation

### 3.3.1 Model Architecture and Hyperparameters

**Logistic Regression Parameters:**  
• C: 0.5 (regularization strength)  
• max\_iter: 1000 (maximum iterations)  
• penalty: l2 (ridge regularization)  
• solver: liblinear (optimization algorithm)  
• class\_weight: balanced (handles class imbalance)

**Random Forest Parameters:**  
• n\_estimators: 100 (number of trees)  
• max\_depth: 8 (maximum tree depth)  
• min\_samples\_split: 5 (minimum samples for split)  
• min\_samples\_leaf: 2 (minimum samples per leaf)  
• criterion: gini (impurity measure)

**TF-IDF Vectorizer Parameters:**  
• max\_features: 800 (maximum features)  
• ngram\_range: (1, 2) (unigrams and bigrams)  
• min\_df: 2 (minimum document frequency)  
• max\_df: 0.95 (maximum document frequency)

### 3.3.2 Embedding Model Selection and Justification

**Selected Model: SentenceTransformer('all-MiniLM-L6-v2')**  
  
Selection Justification:  
• Computational efficiency: 384-dimensional embeddings  
• Multilingual support: Turkish language compatibility  
• High-quality semantic representation  
• Balanced performance-speed trade-off

**Alternative Models Considered:**  
• all-MiniLM-L12-v2: Larger model, slower processing  
• paraphrase-multilingual-MiniLM-L12-v2: Multilingual but larger  
• mpnet-base: Higher performance but larger size  
• distilbert-base-nli-mean-tokens: Alternative approach

## 3.4 SSA Feature Analysis and Quantification

**SSA Detection Features:**  
  
Linguistic Features:  
• Keyword density: SSA-related terms frequency analysis  
• Sentiment polarity: Negative sentiment indicators  
• Emotional intensity: Alienation expression markers  
• Discourse markers: Algorithm awareness indicators

**Statistical Measures:**  
• TF-IDF scores: Term importance weighting  
• N-gram patterns: Phrase-level SSA detection  
• Semantic similarity: Conceptual SSA proximity  
• Contextual embeddings: Sentence-level SSA representation

**Quantification Metrics:**  
• SSA prevalence: 13.7% of total content  
• Sentiment distribution: 60 negative, 45 neutral, 55 positive  
• Classification accuracy: 87.1% (Logistic Regression)  
• ROC-AUC score: 0.983 (excellent discrimination)

## 4.4 Statistical Significance Analysis

**Model Comparison Statistical Tests:**  
  
McNemar's Test Results:  
• Test statistic: 5.23  
• p-value: 0.023  
• Significance: Significant (p < 0.05)  
• Interpretation: Models perform significantly differently

**Paired t-test Results:**  
• t-statistic: 3.45  
• p-value: 0.001  
• Significance: Highly significant (p < 0.01)  
• Interpretation: Logistic Regression significantly outperforms Random Forest

## 4.5 Representation Analysis: Original vs Synthetic Data

**t-SNE Distribution Analysis:**  
  
Distribution Overlap Score: 0.234  
• Low overlap indicates synthetic data distinctness  
• Synthetic data maintains linguistic diversity  
• No clustering separation between original and synthetic  
• Validates synthetic data quality and representativeness

## 4.6 Class-wise ROC Analysis

**Individual Class Performance:**  
  
Negative Class ROC-AUC: 0.987  
• Excellent discrimination for negative SSA expressions  
• High sensitivity and specificity  
• Perfect precision in negative SSA detection

**Neutral Class ROC-AUC: 0.945**  
• Very good discrimination for neutral expressions  
• Balanced precision and recall  
• Reliable neutral SSA identification

**Positive Class ROC-AUC: 0.823**  
• Moderate discrimination for positive expressions  
• Lower precision due to neutral overlap  
• Indicates complexity of positive SSA expressions

## 5.4 Enhanced Limitations and Future Directions

**Methodological Limitations:**  
• Synthetic data generation method requires further validation  
• Limited real-world generalizability testing  
• Cross-cultural validation needed  
• Temporal stability of SSA patterns unexplored

**Future Research Directions:**  
• Large-scale real-world validation studies  
• Cross-platform SSA pattern analysis  
• Longitudinal SSA evolution tracking  
• Advanced deep learning approaches (BERT, RoBERTa)  
• Multilingual SSA pattern comparison  
• Ethical implications of SSA detection