Identifying Patterns and Conventions with Clustering and Dimension Reduction

GitHub Link: https://github.com/Pritam0705/MACS-40123/tree/main/ITR2

Part 1: Analysis of Peer-Reviewed Papers

1. **Analysis of Twitter Data Using Evolutionary Clustering during the COVID-19 Pandemic**
   Author: Ibrahim Arpaci, Shadi Alshehabi, Mostafa Al-Emran3, Mahmoud Khasawneh, Ibrahim Mahariq, Thabet Abdeljawad, and Aboul Ella Hassanien
   Published: Computers, Materials & Continua, 2020

- **Summary:** This article presents a study on the analysis of Twitter data during the early stages of the COVID-19 pandemic using evolutionary clustering techniques. The research aimed to understand public attention trends and sentiments related to the pandemic by examining a large volume of tweets. In evolutionary clustering, the data stream for a given time period should be incorporated into the clustering results from the previous time period, allowing the clustering to be adjusted based on the current data stream. Unigram terms (single words) were found to trend more frequently than bigram and trigram terms (two and three-word phrases). There was a large volume of COVID-19-related tweets, indicating widespread public attention to the pandemic. High-frequency words such as "death", "test", "spread", and "lockdown" suggested public fear of infection and mortality. The study suggests that social media posts can significantly affect human psychology and behavior during a crisis.

2. **Using Fuzzy Clustering with Deep Learning Models for Detection of COVID-19 Disinformation**
   Author: MU-YEN CHEN, YI-WEI LAI
   Published: ACM Transaction Asian Low-Resource Language Information Processing

1. **Summary:** This article presents a novel approach to detecting COVID-19 disinformation using a combination of fuzzy clustering and deep learning models. The study analyzes Chinese and English misinformation related to the COVID-19 pandemic. FCM is a soft clustering algorithm where each data point can belong to multiple clusters with different degrees of membership, rather than belonging to only one cluster. BiLSTM achieved the highest accuracy for COVID-19 misinformation detection, with 99% in English and 86% in Chinese. Integrating fuzzy clustering with English data preserved 99% accuracy while reducing detection time by 10%. It reduced detection time by 15% for Chinese data but lowered accuracy by 8%.

3. **An evaluation of document clustering and topic modeling in two online social networks: Twitter and Reddit**
   Author: Stephan A. Curiskis, Barry Drake, Thomas R. Osborn, Paul J. Kennedy
   Published: Information Processing and Management, 2020

- **Summary:** This article presents an evaluation of document clustering and topic modeling techniques applied to two popular online social networks: Twitter and Reddit. The study aims to compare the performance of various algorithms in extracting meaningful topics and clusters from short-text social media data. Four classic clustering algorithms i.e., K-means, Mini-batch K-means, Agglomerative Clustering, Non-negative Matrix Factorization, and one topic modeling algorithm LDA were used. The study compared four different feature representations i.e., TF-IDF, Word2Vec, Word2Vec weighted with the top 1,000 TF-IDF scores and Doc2Vec. The K-means algorithm using Doc2Vec representation performs best on both Twitter and Reddit datasets.

4. **COVID-19 fake news analytics from social media using topic modeling and clustering**
   Author: Sherrylin Anak John and Pantea Keikhosrokiani
   Published: Big Data Analytics for Healthcare, Academic Press, 2022
   Summary: This article presents a study analyzing COVID-19 fake news on social media using topic modeling and clustering techniques. The research focuses on understanding the spread of misinformation during the pandemic and developing methods to detect and categorize fake news. The study employed the topic modeling technique, latent Dirichlet Allocation (LDA), to identify common themes and subjects in the fake news content. A clustering algorithm was used to group similar pieces of misinformation, helping to identify patterns and categories of fake news.

Clustering and dimensionality reduction techniques can be valuable in analyzing patterns and themes within misinformation data. Clustering algorithms like **K-Means**, **DBSCAN**, or **Hierarchical Clustering** can group claims based on shared themes or linguistic similarities. For instance, one cluster might contain misinformation about elections, another about health, and so on. Reducing the dimensionality of the data can improve the performance and interpretability of clustering algorithms. High-dimensional data often has noise, which can interfere with clustering. Techniques like PCA can help remove noise and extract the most relevant features, leading to clearer clustering results.

Part 2: Interpretation

Principal Component Analysis (PCA):

- **Dimensionality Reduction**: PCA was used to reduce the high-dimensional text data into a lower-dimensional space while retaining the most important variance. This step simplifies the data, allowing for easier clustering without losing essential patterns or themes within the misinformation claims.
- **Noise Reduction**: By reducing the data to the principal components, PCA filters out less significant features, helping to focus on the core aspects of misinformation narratives. This improves the clarity of clusters, ensuring that the main themes are represented without interference from noise or irrelevant details.

K-means Clustering:

- **Identifying Major Themes**: After dimensionality reduction, K-means clustering groups the misinformation data into distinct clusters based on content similarity. Each cluster reveals a specific misinformation theme, such as communal tensions, political narratives, healthcare misinformation, or fraudulent schemes.
- **Visual Insights**: The word clouds generated for each cluster summarize the dominant terms within each theme. This makes it easy to interpret the primary focus of each group, helping to understand how misinformation targets sensitive areas such as religion, politics, and social welfare.

Latent Dirichlet Allocation (LDA):

- **Refining Topics within Clusters**: After clustering, LDA is applied within each group to identify sub-topics or specific themes within the broader misinformation categories. This allows for a more granular view of the misinformation narratives within each cluster.

Summary

By combining PCA, K-means, and LDA, we gain a comprehensive understanding of misinformation patterns and the nuanced narratives within each theme. This approach highlights how misinformation is crafted to exploit sensitive social, political, and communal issues, offering valuable insights for identifying and addressing specific misinformation strategies.

1. **What worked:**
    - PCA effectively reduced dimensionality, highlighting core patterns in the data.

- K-means clustering successfully grouped misinformation into distinct themes, revealing targeted narratives.
- LDA refined these clusters, providing deeper insights into specific sub-topics within each theme.

2. **What didn't work:**
   - Some clusters were broad and required further refinement, as overlapping themes led to mixed content.
   - LDA sometimes produced less coherent topics within clusters, especially with sparse or ambiguous data.
3. **Future revisions:**
   - Experiment with alternative clustering methods (e.g., hierarchical clustering) to improve topic specificity.
   - Test additional dimensionality reduction techniques (like UMAP) to enhance the separation of themes.
   - Incorporate a labeled dataset (manually labeled claims into themes) to evaluate and validate topic coherence more effectively.

Part 3: Social, Cultural, and Behavioral Implications of Findings

Social Implications

Increased Polarization: The findings reveal clusters of misinformation that exploit sensitive issues, such as communal tensions and political conflicts. By targeting these topics, misinformation contributes to social polarization, creating divisions within communities, especially between religious and political groups.

Amplified Emotional Reactions: Misinformation narratives are crafted to elicit strong emotional responses. Visual misinformation, as seen in photos of violence or protests, can provoke fear, anger, or outrage, potentially inciting real-world conflicts or protests.

Cultural Implications

Reinforcement of Stereotypes: The frequent targeting of religious and ethnic groups (e.g., Hindu-Muslim dynamics) perpetuates cultural stereotypes and biases, solidifying divisive narratives. This affects inter-group relations and can lead to prejudice and discrimination against certain communities.

Cross-Border Cultural Tensions: Misinformation involving neighboring countries, such as India and Bangladesh, contributes to cultural misunderstandings and tensions. By highlighting incidents with a communal angle, these narratives can fuel cross-border hostilities and cultural isolation.

Impact on Cultural Institutions: Narratives that involve cultural symbols or national icons (e.g., political leaders, religious figures) manipulate public perception, potentially undermining respect for cultural institutions and heritage. This erodes the cultural fabric by pitting individuals against one another based on manipulated cultural identities.

Behavioral Implications

Spread of Misleading Information: The prevalence of misinformation on platforms like WhatsApp and social media encourages the habit of forwarding and sharing content without verification. This behavior exacerbates the spread of false information, leading to a hard-to-control cycle of misinformation.

Influence on Voting and Civic Behavior: Misinformation about elections, political candidates, and policies can influence voting behavior, potentially skewing democratic processes. Voters may form opinions based on fabricated stories rather than factual information, affecting political outcomes.

Conclusion

In conclusion, the analysis of misinformation themes using PCA, K-means clustering, and LDA reveals how misinformation strategically targets sensitive social, cultural, and political issues to exploit public sentiment and amplify division. These narratives, especially those centered around communal tensions, political controversies, and sensationalized incidents, contribute to social polarization, reinforce harmful stereotypes, and foster distrust in institutions. The behavioral implications—such as increased sharing of unverified content and skewed civic behavior—highlight the urgent need for effective misinformation countermeasures. By understanding the patterns and impacts of misinformation, we can develop more targeted approaches to educate the public, promote critical media literacy, and safeguard social cohesion.