Navigating Social Media Text Analytics: Overcoming Linguistic Complexity via Advanced Modeling Techniques

Cecelia F. Gordon Joseph W. Stewart Ares Boira Lopez Dr. Hairong Song Dr. Shane Connelly Dr. Matthew Jensen
University of Oklahoma

NCITE NATIONAL COUNTERTERRORISM INNOVATION, TECHNOLOGY, AND EDUCATION CENTER
A U.S. DEPARTMENT OF HOMELAND SECURITY CENTER OF EXCELLENCE

Introduction

Alignment with Research Objectives

This research supports the National Counterterrorism Innovation Technology and Education Center's (NCITE) objectives by offering new methodologies to analyze and visualize social media data, aiming to enhance detection and understanding of terrorist communications online.

University of Oklahoma Collaboration

Highlighting the strategic partnership and expertise of the University of Oklahoma team in leveraging these advanced techniques.

Problem Statement

Challenges in Social Media Text Analysis

•**Data Abnormalities**:

•Ideological datasets often have excessive zeros, non-normal distributions, and semi-continuous data, complicating accurate analysis, particularly for rare phenomena like radical language and hate speech (King & Zeng, 2001; Wiegand et al., 2019) .

•**Traditional Methods' Limitations**:

•Existing linguistic processing tools (e.g., LIWC, WordNet) struggle with dynamic and context-dependent language on platforms like Twitter (Tausczik & Pennebaker, 2010; Boyd, 2017).

The above challenges are further magnified in studies involving sensitive topics like terrorism-related content, where the extremely low base rates of relevant terms often lead to potential biases in analytical results (Conway et al., 2012; Scrivens et al., 2020).

Our Approach

To address the complex structure of social media and linguistic data, our research team has implemented advanced text analytic procedures:

1. ***Multilevel Modeling***
2. ***Mixed Effect Modeling with a Gamma Link***
3. ***Two-parts Mixed Effect Modeling***

These sophisticated methods enhance the reliability and depth of our social media text analysis, addressing the limitations of traditional tools and overcoming data sparsity challenges.

Innovative Analytical Techniques

1. Hierarchical Linear Modeling (HLM)

To account for the data having a nested structure (i.e., users nested within groups), we utilized hierarchical linear modeling (HLM) techniques to evaluate the effects of the role of the user in the group (e.g., generic group account, prominent member, or leader), the role of violence classification associated with the user's group (e.g., violent or non-violent), and the role of political ideology of the user's group (e.g., left or right-leaning).

•**Two-level HLM:**

•Role of the user account served as the level one variable, while violence classification and political ideology served as level two variables. See Figure 1 and Table 1.

2. Mixed Effects Modeling with a Gamma Link

This model is used for proportional and positively skewed data. These phenomena occur at low base rates, thus much of the data is positively skewed, which requires the use of more appropriate distributions to model the data. This technique was successfully used to model the use of moral foundations in ideological group messaging.

3. Two-Parts Mixed Effects Modeling

Zero-Inflated Models

Two-parts mixed effect model with a Poisson link for zero-inflated count data.

•This model separates zero-count data from non-zero (positive) occurrences. The two parts allow for the examination of non-zero data at varying levels of the outcome variable for a nuanced analysis.

Semi-Continuous Data Models

Two-parts mixed effect model for datasets with zero values and continuous distributions among non-zero observations.

•*Part 1:* A logistic regression is used to predict the probability of an observation being zero or non-zero based on fixed effects.

•*Part 2:* A standard linear mixed model is applied to the logarithmic transformation of the non-zero data to examine the effects of the predictors on the magnitude of the non-zero values.

Figure 1

Proposed HLM Model

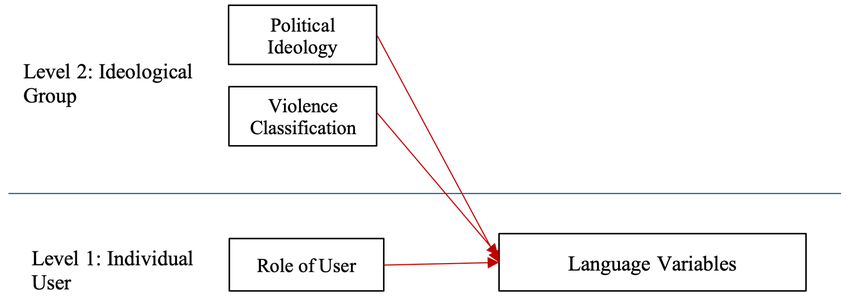

Table 1

Results from HLM Analysis of "Trust" Language Used

Predictors	Model 1 (β)	Model 2 (β)	Model 3 (β)
Intercept	1.10**	1.19**	1.20**
Level 1			
Role	-0.04	-0.05	-0.07
Level 2			
Violence	-	-0.22*	-
Ideology	-	-	-0.12
AIC / BIC	239.36 / 258.18	237.00 / 252.65	240.99 / 256.64

Note. $N = 172$. Role is coded 1 = group account, 2 = prominent member, 3=leader; Violence is coded 0 = non-violent, 1 = violent; Ideology is coded 0 = right-leaning, 1 = left-leaning.
**$p < 0.01$, *$p < 0.05$.

Contributions

- These sophisticated techniques aim to enhance the reliability and depth of test analysis in social media and linguistic research.
  - This helps to overcome limitations posed by traditional linguistic analysis tools.
  - Addresses the inherent complexities of data sparsity, excessive zeros, and positively skewed distributions.

Limitations and Future Directions

- The current models need further evaluation for their robustness and scalability to larger datasets and different social media platforms.
  - Future research can test and refine the models across various contexts and data sources.
  - Future research should continue to develop and uncover adaptive models that better capture the nuances of social media language.

*Acknowledgement: This project was funded by NCITE, a Center of Excellence for the U.S. Department of Homeland Security based at the University of Nebraska-Omaha