NLP Analysis of Media Coverage: Freedom in Tibet - Capstone Project Proposal

Professor : scott.crossley@Vanderbilt.Edu

Project Overview

Objective: Analyze how different news sources frame discussions of Tibetan freedom using NLP techniques to reveal patterns in media discourse and potential biases across outlets, time periods, and geographic regions.

Research Questions & Hypotheses

Primary Question: How do different news sources linguistically frame concepts of freedom and human rights in Tibet?

Key Hypotheses:

- **H1:** Western vs. Chinese state media will show statistically significant sentiment differences in Tibetan freedom coverage
- **H2:** Different outlets will use distinct terminology patterns ("autonomy" vs. "independence" vs. "separatism")
- **H3:** Coverage sentiment will correlate with major political events (2008 protests, policy changes)
- **H4:** Topic modeling will reveal distinct thematic clusters (cultural, political, economic, religious)

Methodology

NLP Techniques:

1. **Sentiment Analysis:** BERT-based classification, aspect-based sentiment, temporal tracking
2. **Topic Modeling:** LDA, dynamic topic modeling for temporal analysis

Technical Stack: Python, spaCy, transformers (Hugging Face), scikit-learn, pandas

Dataset Strategy

Target: 6400 (2008-2024)

Primary Sources:

- **Western Media:** BBC, CNN, Guardian, NYT, Reuters
- **Chinese State Media:** China Daily, Global Times, Xinhua (English)
- **International:** Al Jazeera, Deutsche Welle, SCMP
- **Tibetan-focused:** Tibet Post, Phayul

Collection Methods:

- Web scraping with custom scrapers
- GDELT Project database (structured event data)
- News APIs (Guardian, Reuters, News API)
- Archive.org for historical content

Corpus

- Political leaning of new media
- 2008 - 2024- sentiment changing-
- Articles
- Keyword : Tibet
- 100 news paper per sources per year in 4 = 100*4*16 = 6400
- Sentiment has change
- Time series
- Topic modeling : fine grain index , Sentiment analysis
-