Research Design Final Project Computational Social Science - 26/01/2025

Title: *"Analysis of Public Sentiment and Discussion Topics on the Israeli-Palestinian Conflict: An In-depth Study and Comparison of YouTube Comments Across Different News Channels"*

Author: Christian Braga

Research Questions:
1. What is the prevailing sentiment among YouTube users about the Israeli-Palestinian conflict, and what are the primary topics of discussion for each side?
2. Has the sentiment and the topic of discussion evolved from the beginning of the conflict?
3. Is the sentiment expressed by YouTube users consistent across different news channels? Do users from different sides have specific preferences for the news channels they use to gather information?
4. Is it possible to determine any relevant information about different news channels based on user sentiment and preferences, such as whether users perceive that an information network implicitly supports one faction over another?

Experiment / Methodology:
To address these research questions, the study will analyze YouTube comments on videos covering the current state of the conflict and the actions of the involved parties. It will employ transformer-based models and topic modeling techniques to extract and assess user sentiments and discussion themes. Additionally, the experiment will compare the results across different time periods and among various news channels.

Population (Target):
YouTube Users.

Sample:
The analysis targets YouTube users who comment on videos concerning the Israeli-Palestinian conflict. Since verifying each commenter's nationality is not feasible, only English-language comments will be considered. The aim is to include a globally diverse user base to ensure a comprehensive perspective. Accordingly, videos from news channels around the world have been selected to represent a broad array of viewpoints.

Unit:
The fundamental unit of analysis is the individual YouTube comment. The comments will then be aggregated based on several factors: the video they belong to, the timeframe in which the video was published, the news channel that released the video, the sentiment they express, and the topics they discuss.

Data / Observations:
The YouTube video comments are collected via API and aggregated according to 2 purposes:
1. Fine tuning of the model, this dataset consists of comments collected from different videos dealing with the Israeli Palestinian conflict that have been manually labeled and will be used to train the model to understand the context of our analysis in a supervised way.
2. Inference, comments from specific videos that differ based on the date of publication (beginning of the conflict / 1 year after the beginning of the conflict) and the news channel (for each news channel 2 videos will be analyzed one published at the start of the conflict, the second approximately 1 year later).

NOTE: "the beginning of the conflict" is defined as October 2023. This specific timeframe was chosen as it marks a significant escalation in tensions relevant to the Israeli-Palestinian war, providing a clear starting point for examining changes in public sentiment and discussion themes over time.

Measurement:
For sentiment analysis, the comments will be classified into one of 3 possible target variables:
0 = comment that supports the Israeli cause
1 = comment that supports the Palestinian cause
2 = neutral comment / not expressing an opinion on the topic
For the topic modeling analysis, unsupervised learning techniques will be used, so by definition there are no target variables.

Measurement Error:
As for sentiment analysis, errors in the measurements will certainly be present and will be entirely determined by the accuracy of the model and the difficulty (even for human persons) of interpreting the comments (due to sarcasm, allusions, and irony).
As for topic modeling, any errors in identifying the topic of discussion will predominantly be determined by the model developed and the pre-processing techniques used to prepare the data.

Machine Learning Methods:
Transformers Based Models, Topic Modeling Algorithms (Unsupervised).

ML Models:
Sentiment Analysis: BERT (transformers based model) - Topic Modeling: BERT-topic / LDA.

Theory / Mechanism:
One possible theory/mechanism connecting a comment to its sentiment could be the subject of the video, a video showing Israeli army attacks on the Palestinian population should evoke a feeling of rejection and opposition from pro-Palestinian users and a feeling of approval from the other side, and vice versa.

Training / Testing:
BERT is a pre-trained model on which we will perform fine tuning to adapt it for the specific sentiment classification task. Fine tuning will be performed on the specifically created and labelled dataset which is split into train, validation, and test set (we will not adopt cross-validation for computation/time reasons).

Assessment of Accuracy:
Cross Entropy Loss - Prediction Accuracy - Confusion Matrix.

Previous Studies:
Multi-Level Analysis of political sentiments using twitter data: a case study of the palestinian-israeli confict, Iyad Al-Agha and Osama Abu-Dahrooj, December 2019. This paper performs an analysis of this conflict using Twitter data and a traditional ML model (Logistic Regression). A potential scientific contribution of my project is to leverage the use of innovative transformer-based architectures for text classification. Additionally, my project proposes studying this conflict through a different social media platform and focuses on various news channels in hopes of uncovering more insights and comparing this work with previous studies.