Topic Model for Image and Text Analysis

Justin Chun-ting Ho, A. Marthe Möller, Joanna Strycharz, Rhianne W. Hoek

Amsterdam School of Communication Research

Introduction

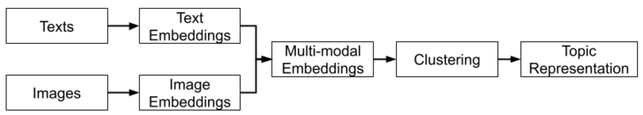
We examine three approaches that leverage **Vision-Language Models (VLM)** for topic extraction from visual and textual content. Using data from Instagram posts by green influencers, **the study evaluates the performance of LVM against manually coded themes**. The study underscores the potential of VLM in the computational analysis of visual content and suggests avenues for further research to enhance the robustness and applicability of its use in communication research.

Data

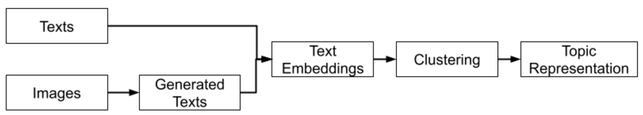
- Sustainability-related Instagram posts from 44 Dutch green influencers
- A dataset of 794 image-caption pairs (translated into English, **manually annotated in a previous work**)
- Convert the image and textual content into numerical representations using **CLIP**, apply clustering algorithm using **BERTopic**, manually match each cluster to a pre-defined theme in the codebook
- Calculate Correct Classification Rate

Approaches

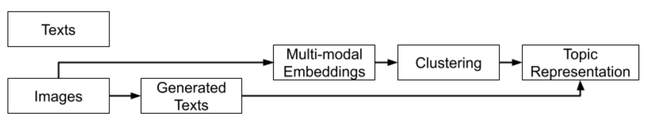
1. Multi-modal Approach



2. Text-based Approach

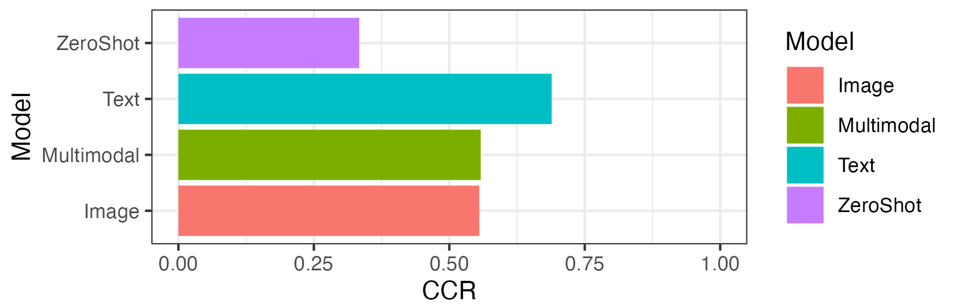


3. Image-only Approach



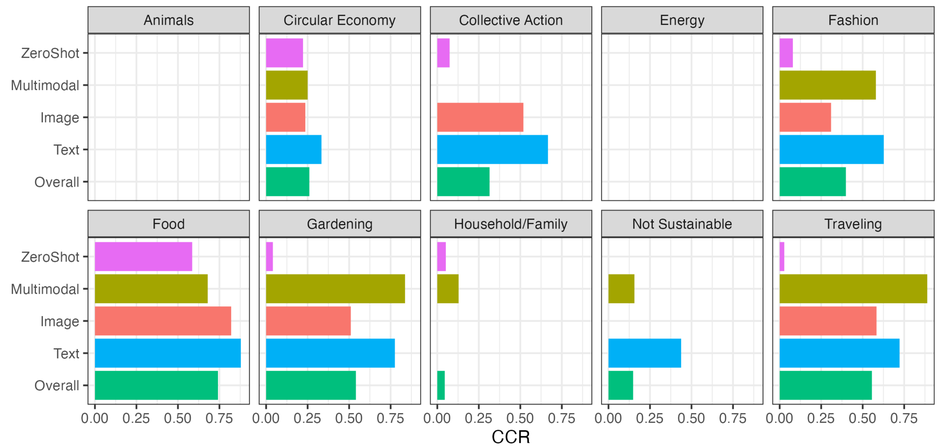
We also use the **open vocabulary image classification** (Zero-Shot) approach suggested by OpenAI as baseline

Text-based Approach Outperforms the Rest in Correct Classification Rate



The graph shows the overall Correct Classification Rate (CCR). **Text-based approach shows the highest CCR of 0.689.** The multi-modal approach yields a CCR of 0.558, the image-only approach yields a similar CCR of 0.555. **Zero-Shot approach yields the lowest CCR of 0.334.**

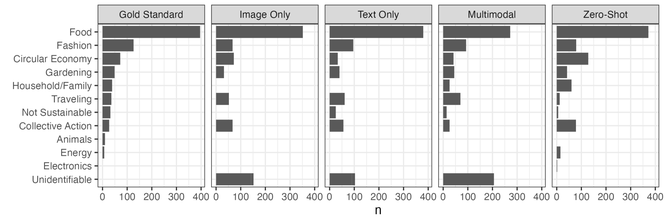
Correct Classification Rate Varies Significantly across Themes



We observe **great discrepancies across themes and between approaches**. In general, all three approaches are able to reliably identify Food, Travel, and Gardening with an average CCR of 0.793, 0.731, and 0.707 respectively. Rare themes such as Animals and Energy have zero CCR as none of the models can identify them.

The level of complexity of the themes seems to have an effect on CCR. Themes about abstract concepts, such as Circular Economy and Sustainability, have extremely low CCR. **Among these abstract topics, text-based approach seems to perform the best.** One possible explanation is that some themes can be completely communicated with images alone (a photo can perfectly describe food) while others require some forms of textual explanation (it is much harder to describe the concept of circular economy without using texts).

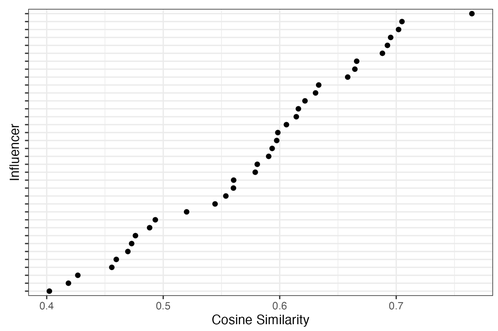
Topic Distributions



The Gold Standard sub-graph shows the counts as identified by the manual coding. In general, all approaches follow similar distributions. **Multimodal approach fails to classify the highest amount of posts**. Zero-Shot approach is able to return a theme for every post.

Image-Text Alignment Matters

We also explore the effect of image-text alignment on CCR. The embeddings of the textual and visual content were estimated for each post using CLIP, cosine similarity was then calculated. In general, we observe **a large variation in image-text alignment across influencers**.



To estimate its effect on CCR, we run four logistic regression models. We created a dummy variable (correctly classified = 1, otherwise = 0) as the dependent variable and the cosine similarity is used as the independent variable, We also controlled for the topic and influencer.

For the **multi-modal approach**, the logistic regression model yielded an odds ratio (OR) of 7.543 for cosine similarity ($p = 0.027$), indicating that for **each unit increase in cosine similarity, there is a 654% increase in the odds of correct classification**. We cannot observe any significant association (alpha = 0.05) between cosine similarity and correct classification for the other three approaches.

j.c.ho@uva.nl