Data Analytics & Society

dubit

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

# Using YouTube data to discover topics related to the 'slime' trend

Emily Coupland
ejcoupland1@sheffield.ac.uk
@emilycoupland

This study explored the capacity for exploring latent topics within a trend by applying textual analysis and topic discovery methods to YouTube data.

## Background

Content creators on YouTube use keywords to attract viewers, and YouTube titles and descriptions provide a rich source of textual data for analysis.

YouTube topic discovery is not widely used for social research purposes. Yet useful methodological frameworks exist for YouTube topic discovery for the sake of user search journeys.

This study explored the capacity to apply these methods to understand a social phenomenon, the kid's slime trend, which was the top Google trend of 2017.

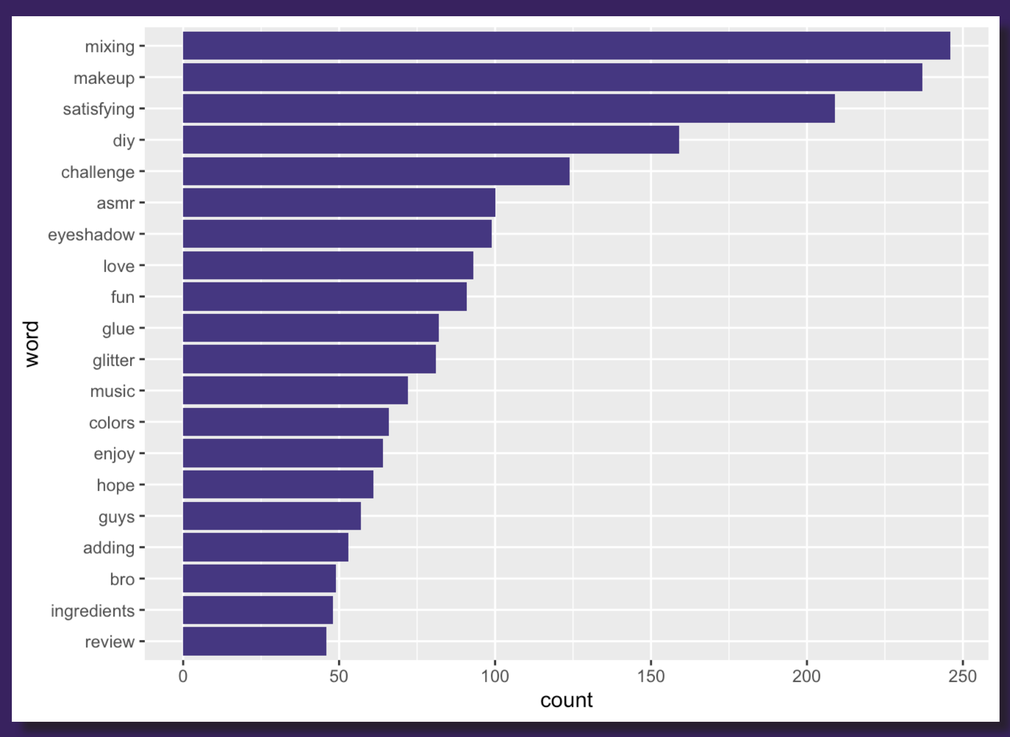
## Data

Data for 286 YouTube videos related to the search query 'slime' was scraped from YouTube's API, which was filtered to 221 videos which were likely to be in the English language.

Further cleaning was applied iteratively to filter stopwords, URLS and tags, and uninsightful frequent terms such as channel names.

## KEY WORDS AND PHRASES

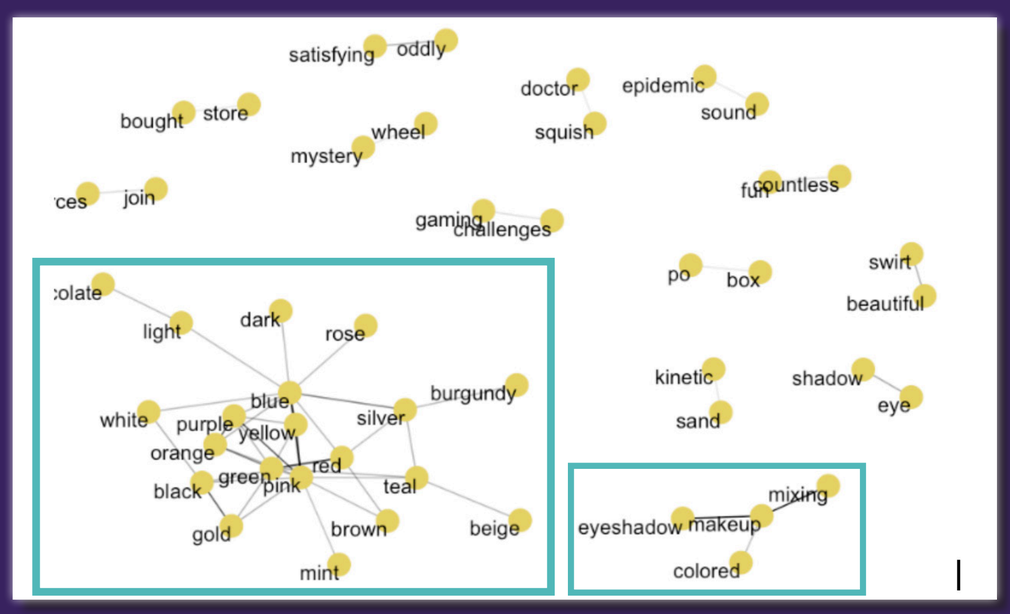
**The top 20 most frequent non-colour words provide some topical insight.**




It is evident that slime videos commonly relate to mixing, make-up, DIY, challenge, ASMR, etc.

Whilst this visualisation focuses on non-colour words, colour names were also prevalent within the top 20 most frequent words, which are topically ambiguous outside of context.

**Two distinct groups are evident in the top 55 most frequent two-word pairings; colour names and makeup.**




Observing commonly associated words provides useful context to some of the more ambiguous keywords;

- colour names form a distinct inter-related group
- the word 'mixing' is contextualised through it's strong association with makeup

## TOPIC MODELING

**Topic modeling produces semantically meaningful vocabulary groups which are useful for providing topical insight**




- Each topic seems to contain a vocabulary which is semantically distinct and meaningful:
  - Topic 1 relates to DIY and ingredients
  - Topic 2 relates to gaming
  - Topic 3 relates to colours and makeup
  - Topic 4 relates to challenge (bro/sis/challenge)
  - Topic 5 relates to products and commerce
  - Topic 6 relates to ASMR and satisfying content.
- It is useful that topic modeling inherently produces overlapping topics. For instance DIY can be understood in relation to multiple topics.
- Topic 6 contains related English and non-English words (i.e. "relaxante" and "relaxing"), which suggests a robustness of this method to multilingual data.
- The robustness to multi-lingual words, and the allowance of overlapping vocabulary, are advantages of topic modeling over the other clustering methods tested (K-Means and Hierarchical).