MultiModalTopicExplorer: A Visual Text Analytics System for Exploring a Collection of Multi-modal Online Conversations

Soheil Alavi
salavis@cs.ubc.ca

Felipe González-Pizarro
felipegp@cs.ubc.ca
Content Warning

Our results contain textual and graphic elements that are anti-semitic, anti-muslim, racist, sexist, homophobic, and offensive in many other ways.
Topic modeling and InfoVis

- It is needed to **summarize** and **understand** textual data

- Promising solution: Topic modeling
  - Statistical approach for extracting **topics** from large text corpora.
  - Topic models do not provide meanings and interpretation directly
    - **Humans must be involved** [1]

- Humans who directly interact with and interpret the output of topic modeling **may rely on visualization tools to better interpret the results** [2]

---


Finding mechanisms to improve these visual representations is still an open challenge [1].

Lack of support on qualitative analysis of topic models

- Visual Analytics systems can provide valuable insights about machine learning model’s intrinsic properties and behaviors [1][2]
  - NLP experts can use these systems to evaluate the quality of topic models
  - Current topic modeling visualizations tools do not provide explicit functionalities to support this task.

Lack of support on multi-modal conversations

- With the proliferation of web-based social media, there has been an exponential growth of asynchronous online conversations discussing a large variety of popular issues [1]
  - To discuss these and other topics, social media users post textual and image data.
  - To the best of our knowledge, none of the current topic modeling visualization tools support image representation of topics

Our proposal: MultiModalTopicExplorer
Our proposal: MultiModalTopicExplorer

- Two key innovations:
  - It allows to report the quality of the most frequent topics
  - Show the most relevant images for each topic
Our proposal: MultiModalTopicExplorer

- **Functionalities:**
  - Identify most relevant keywords, documents, and images for each topic
  - Evolution of topics over time
  - Rate topics

- **End Goal:** Helping users to evaluate topics’ quality
MultiModalTopicExplorer - Task abstraction

- Functionalities:
  - Identify frequent topics
  - Identify most relevant keywords, documents, and images for each topic
  - Evolution of topics over time
  - Rate topics

- End Goal: Helping users to evaluate topics’ quality
Dataset: 4chan dataset

3.3M threads and 134M posts from the Politically incorrect board (/pol/), posted over a period of almost 3.5 years

Multimodal (Image + Text): 4chan https://zenodo.org/record/3606810#.YU-wSLhKiUk
Why 4chan dataset?

- An exploration of these conversations could help understand how these communities interact on these platforms.
- It is the first step before creating automated hate speech detection and mitigation systems.

Multimodal (Image + Text): 4chan [https://zenodo.org/record/3606810#.YU-wSLhKiUk](https://zenodo.org/record/3606810#.YU-wSLhKiUk)
Our adopted Base Dataset:

- More than 0.5 million randomly selected samples
- Over a period of 1.5 years (June 2016-Dec 2017)
- Remove HTML tags and punctuations
- Lowercase the words
- Lemmatization, Stop words removal,

Multimodal (Image + Text) : 4chan [https://zenodo.org/record/3606810#.YU-wSLhKiUk](https://zenodo.org/record/3606810#.YU-wSLhKiUk)
BERTOPIC

A Visual Text Analytics System for Exploring a Collection of Multi-modal Online Conversations
BERTOPIC-Implementation

- BERTopic finds the number of topics automatically
  - We found 815 topics
- No Bigrams and Trigrams calculations needed for phrase generation!
- Training:
  - 2 hours: 4 GeForce GTX 1080 Ti GPUs
  - Chunks of size 130k samples
LDA

- It is based on the assumption that document collections have latent topics in the form of a multinomial distribution of words, which is typically presented to users via its top-N highest probability words (Lau et al., 2014)
- Traditional and popular method even today
LDA-Implementation

- Calculated bigrams and trigrams for finding phrase keywords
- Used gensim LDA multicore
  - 9 cores cpu took 6 hours to train the best model
- Num topics = 600 yields best Coherence Score

![Coherence Score Over Different Num topics](chart.png)
**CLIP: Contrastive Language-Image Pre-Training**

1. **Contrastive pre-training**
   - Text Encoder
   - Image Encoder
   - Neural network trained on a variety of (image, text) pairs.

2. **Create dataset classifier from label text**
   - A photo of a (object).
   - Text Encoder

3. **Use for zero-shot prediction**
   - A photo of a dog.
   - Image Encoder

Source: [https://github.com/openai/CLIP](https://github.com/openai/CLIP)
Content Warning

Our results contain textual and graphic elements that are anti-semitic, anti-muslim, racist, sexist, homophobic, and offensive in many other ways.
Demo Url

http://MultiModalTopicExplorer.ml
User study design

Dataset

- 4chan/pol/threads

Tasks

- Report Topics' Coherence
- Report Workload

Conditions

- Group 1
  - LDA
  - BERTopic
- Group 2
  - BERTopic
  - LDA

A Visual Text Analytics System for Exploring a Collection of Multi-modal Online Conversations
Null hypothesis ($H_0$)

$H_0$: There are no differences in the coherence of topics emerged from BERTopic and LDA
We found statistically significant differences between conditions
\((U = 4811.0, N_{bertopic}=80, N_{lda} = 84, p < .001)\)
We can reject Null hypothesis

H₀: There are no differences in the coherence of topics emerged from BERTopic and LDA
Distribution of participant responses to the NASA TLX questionnaire

Distribution of participant responses to the NASA TLX questionnaire

The results hint that:

Users participants felt successful in accomplishing the task, but it required effort and mental demand.

A lower score indicates a better result
MultiModalTopicExplorer functionalities allow users to feel successful while evaluating topic models.
Future work

❖ **Datasets:** Investigate MultiModelTopicExplorer functionalities in other domains

❖ **Scalability:** Seek options to visualize a larger number of topics (e.g., 300 topics) in a longer period of time (e.g., 100 months) in a compact manner

❖ **User study:**
  ➢ Increase the number of users participants
  ➢ Compare our tool with other topic modeling visualization tools

❖ **Users in control:** Allow users to change the number of keywords and images displayed
Future work

❖ Use of automatically calculated metrics such as Coherence to compare the BERTopic vs LDA
❖ Consider topic hierarchy
❖ Consider hierarchy in conversation threads (Replies, comments, etc.)
❖ Improve the BERTopic model to be more scalable (Right now it can only fit to 130000 samples, and predict the rest)
❖ Find a way to boost LDA’s training speed with GPUs
Questions?
Thank You! :)