THE UNIVERSITY
OF BRITISH COLUMBIA

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

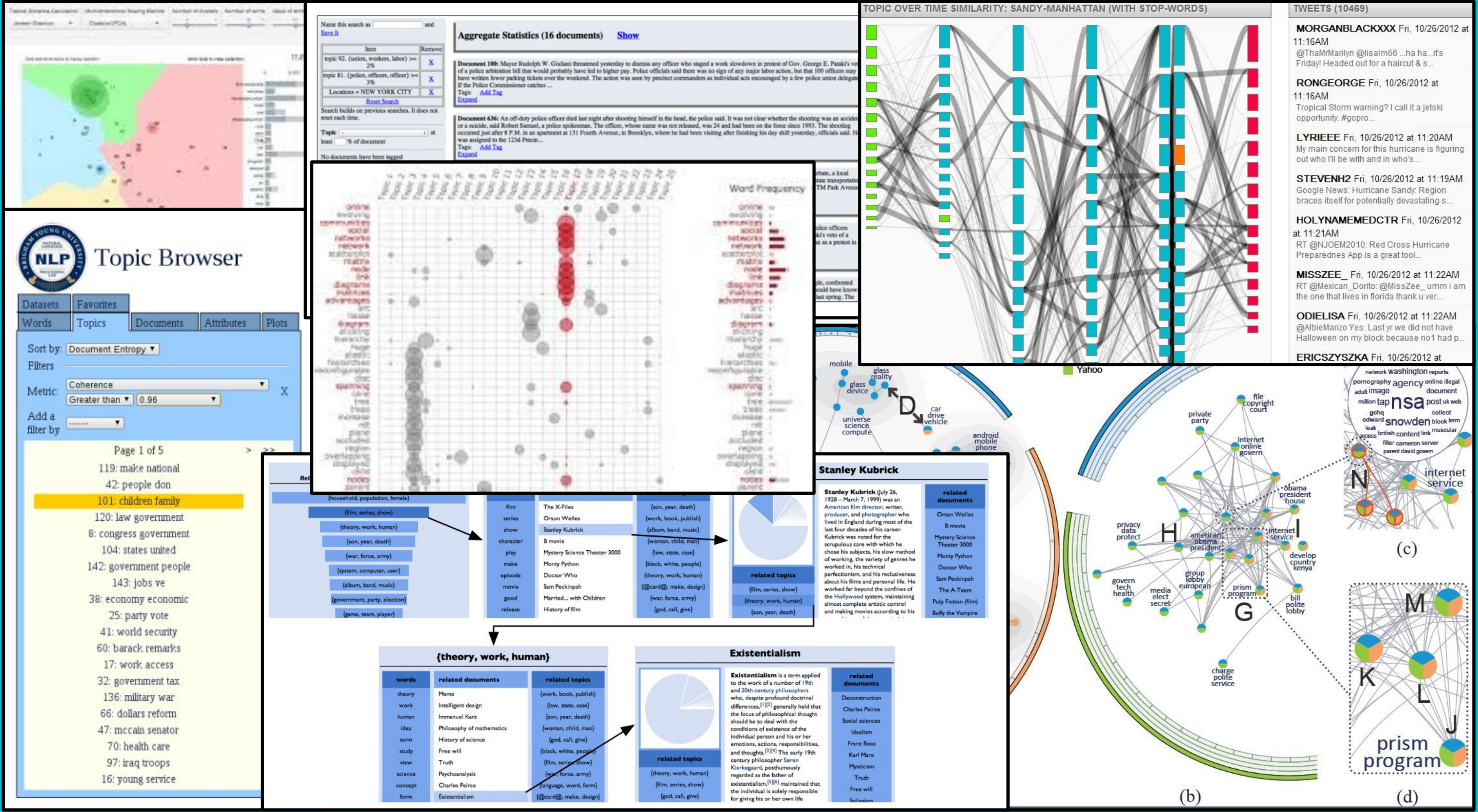
- It is needed to **summarize** and **understand** textual data

- 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]**

[1] Dou, W., Wang, X., Chang, R., & Ribarsky, W. (2011, October). Paralleltopics: A probabilistic approach to exploring document collections. In *2011 IEEE conference on visual analytics science and technology (VAST)* (pp. 231-240). IEEE.

[2] Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J. L., & Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. In *Advances in neural information processing systems* (pp. 288-296).



Name this search as ______ and
Save It

Item	Remove
topic 92. {union, workers, labor} >= 2%	X
topic 81. {police, officers, officer} >= 3%	X
Locations = NEW YORK CITY	X

Reset Search

Search builds on previous searches. It does not reset each time.

Topic ______ at least ___ % of document

No documents have been tagged

Aggregate Statistics (16 documents) Show

Document 100: Mayor Rudolph W. Giuliani threatened yesterday to dismiss any officer who staged a work slowdown in protest of Gov. George E. Pataki's ve of a police arbitration bill that would probably have led to higher pay. Police officials said there was no sign of any major labor action, but that 100 officers may have written fewer parking tickets over the weekend. The action was seen by precinct commanders as individual acts encouraged by a few police union delegat If the Police Commissioner catches ...
Tags: Add Tag
Expand

Document 636: An off-duty police officer died last night after shooting himself in the head, the police said. It was not clear whether the shooting was an accide or a suicide, said Robert Samuel, a police spokesman. The officer, whose name was not released, was 24 and had been on the force since 1993. The shooting occurred just after 8 P.M. in an apartment at 131 Fourth Avenue, in Brooklyn, where he had been visiting after finishing his day shift yesterday, officials said. H was assigned to the 123d Precin...
Tags: Add Tag
Expand

Topic Browser

Datasets | Favorites
Words | Topics | Documents | Attributes | Plots

Sort by: Document Entropy
Filters
Metric: Coherence
Greater than 0.96 X
Add a filter by

Page 1 of 5 > >>

119: make national
42: people don
101: children family
120: law government
8: congress government
104: states united
142: government people
143: jobs ve
38: economy economic
25: party vote
41: world security
60: barack remarks
17: work access
32: government tax
136: military war
66: dollars reform
47: mccain senator
70: health care
97: iraq troops
16: young service

TOPIC OVER TIME SIMILARITY: SANDY-MANHATTAN (WITH STOP-WORDS)

TWEETS (10469)

MORGANBLACKXXX Fri, 10/26/2012 at 11:16AM
@ThaMrMarilyn @lisalm66 ...ha ha...it's Friday! Headed out for a haircut & s...

RONGEORGE Fri, 10/26/2012 at 11:16AM
Tropical Storm warning? I call it a jetski opportunity. #gopro...

LYRIEEE Fri, 10/26/2012 at 11:20AM
My main concern for this hurricane is figuring out who I'll be with and in who's...

STEVENH2 Fri, 10/26/2012 at 11:19AM
Google News: Hurricane Sandy: Region braces itself for potentially devastating s...

HOLYNAMEMEDCTR Fri, 10/26/2012 at 11:21AM
RT @NJOEM2010: Red Cross Hurricane Preparednes App is a great tool...

MISSZEE_ Fri, 10/26/2012 at 11:22AM
RT @Mexican_Dorito: @MissZee_ umm i am the one that lives in florida thank u ver...

ODIELISA Fri, 10/26/2012 at 11:22AM
@AlbieManzo Yes. Last yr we did not have Halloween on my block because no1 had p...

ERICSZYSZKA Fri, 10/26/2012 at

{household, population, female}
{film, series, show}
{theory, work, human}
{son, year, death}
{war, force, army}
{system, computer, user}
{album, band, music}
{government, party, election}
{game, team, player}

film	The X-Files	{son, year, death}
series	Orson Welles	{work, book, publish}
show	Stanley Kubrick	{album, band, music}
character	B movie	{woman, child, man}
play	Mystery Science Theater 3000	{law, state, case}
make	Monty Python	{black, white, people}
episode	Doctor Who	{theory, work, human}
movie	Sam Peckinpah	{@card@, make, design}
good	Married... with Children	{war, force, army}
release	History of film	{god, call, give}

related topics
{film, series, show}
{theory, work, human}
{son, year, death}

Stanley Kubrick

Stanley Kubrick (July 26, 1928 – March 7, 1999) was an American film director, writer, producer, and photographer who lived in England during most of the last four decades of his career. Kubrick was noted for the scrupulous care with which he chose his subjects, his slow method of working, the variety of genres he worked in, his technical perfectionism, and his reclusiveness about his films and personal life. He worked far beyond the confines of the Hollywood system, maintaining almost complete artistic control and making movies according to his

related documents
Orson Welles
B movie
Mystery Science Theater 3000
Monty Python
Doctor Who
Sam Peckinpah
The A-Team
Pulp Fiction (film)
Buffy the Vampire

{theory, work, human}

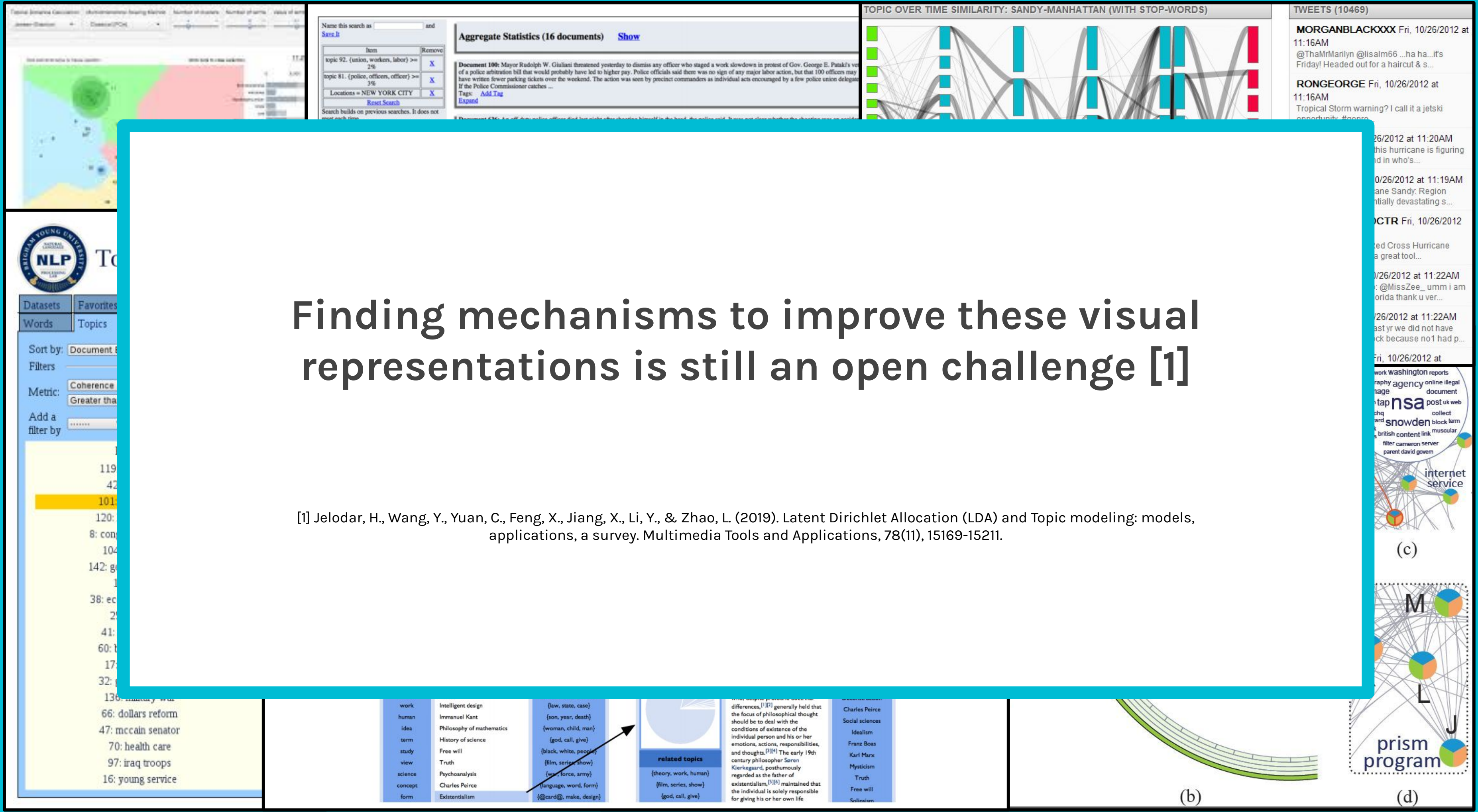
words	related documents	related topics
theory	Meme	{work, book, publish}
work	Intelligent design	{law, state, case}
human	Immanuel Kant	{son, year, death}
idea	Philosophy of mathematics	{woman, child, man}
term	History of science	{god, call, give}
study	Free will	{black, white, people}
view	Truth	{film, series, show}
science	Psychoanalysis	{war, force, army}
concept	Charles Peirce	{language, word, form}
form	Existentialism	{@card@, make, design}

Existentialism

related topics
{theory, work, human}
{film, series, show}
{god, call, give}

Existentialism is a term applied to the work of a number of 19th and 20th-century philosophers who, despite profound doctrinal differences,[1][2] generally held that the focus of philosophical thought should be to deal with the conditions of existence of the individual person and his or her emotions, actions, responsibilities, and thoughts.[3][4] The early 19th century philosopher Søren Kierkegaard, posthumously regarded as the father of existentialism,[5][6] maintained that the individual is solely responsible for giving his or her own life

related documents
Deconstruction
Charles Peirce
Social sciences
Idealism
Franz Boas
Karl Marx
Mysticism
Truth
Free will
Solipsism


Aggregate Statistics (16 documents)
Show
TOPIC OVER TIME SIMILARITY: SANDY-MANHATTAN (WITH STOP-WORDS)
TWEETS (10469)
MORGANBLACKXXX Fri, 10/26/2012 at 11:16AM
@ThaMrMarilyn @lisalm66 ...ha ha...it's Friday! Headed out for a haircut & s...
RONGEORGE Fri, 10/26/2012 at 11:16AM
Tropical Storm warning? I call it a jetski
NLP
Datasets
Favorites
Words
Topics
Sort by:
Filters
Metric:
Coherence
Greater tha
Add a filter by
66: dollars reform
47: mccain senator
70: health care
97: iraq troops
16: young service
related topics
internet service
nsa
snowden
washington
agency
prism program
M
L
J
(b)
(c)
(d)

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

[1] Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., & Zhao, L. (2019). Latent Dirichlet Allocation (LDA) and Topic modeling: models, applications, a survey. Multimedia Tools and Applications, 78(11), 15169-15211.

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.**

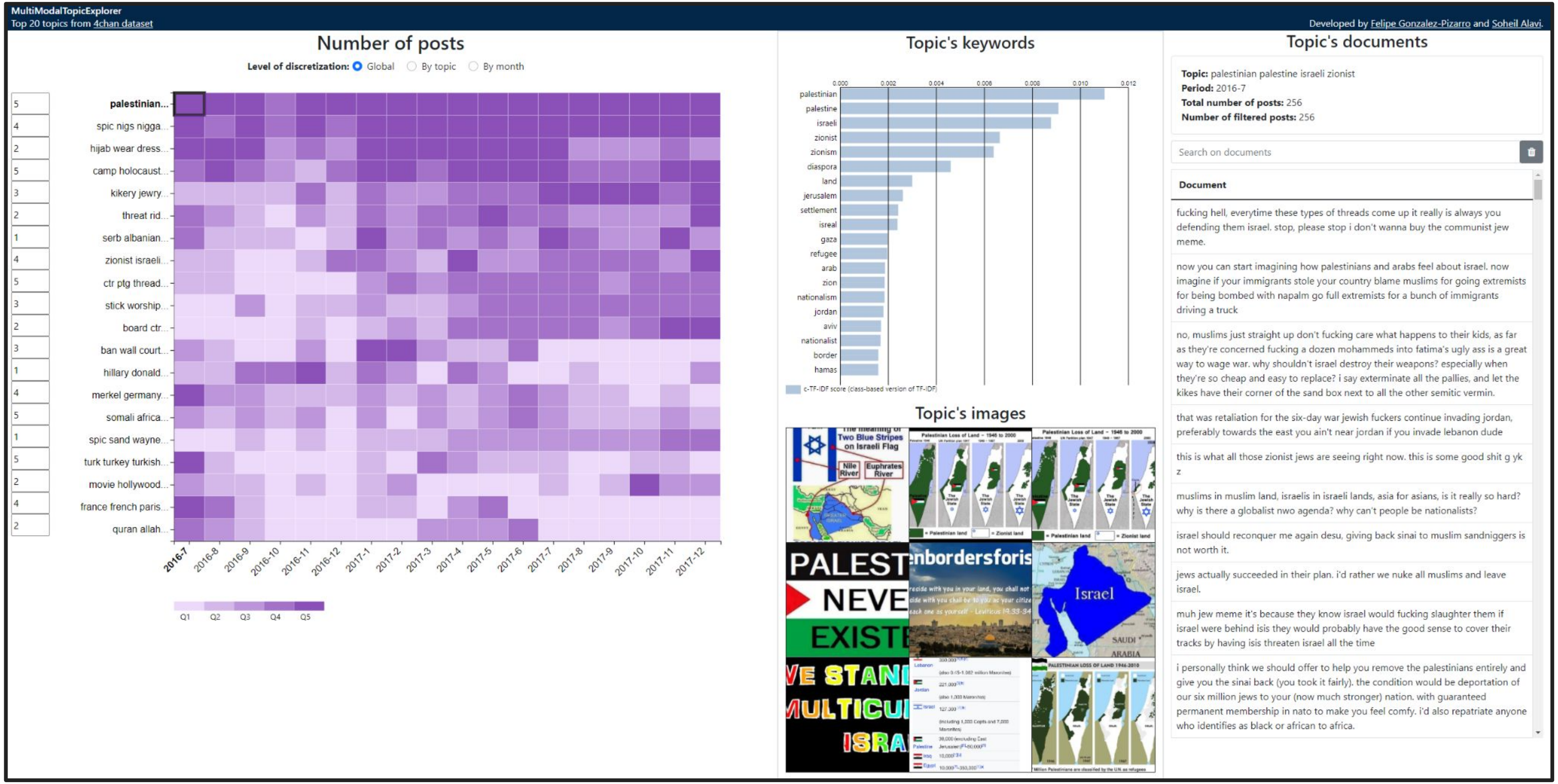
[1] Li, R., Xiao, W., Wang, L., Jang, H., & Carenini, G. (2021). T3-Vis: a visual analytic framework for Training and fine-Tuning Transformers in NLP.
[2] Nan-Chen Chen, Margaret Drouhard, Rafal Kocielnik, Jina Suh, and Cecilia R. Aragon. 2018. Using Machine Learning to Support Qualitative Coding in Social Science: Shifting the Focus to Ambiguity. <i>ACM Trans. Interact. Intell. Syst.</i> 8, 2, Article 9 (July 2018), 20 pages.

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**

[1] Hoque, E., & Carenini, G. (2015, March). Convisit: Interactive topic modeling for exploring asynchronous online conversations. In *Proceedings of the 20th International Conference on Intelligent User Interfaces* (pp. 169-180).

Our proposal: MultiModalTopicExplorer



MultiModalTopicExplorer
Top 20 topics from 4chan dataset

Developed by Felipe Gonzalez-Pizarro and Soheil Alavi.

Number of posts

Level of discretization: ● Global ○ By topic ○ By month

Topic's keywords

Topic's images

The meaning of Two Blue Stripes on Israeli Flag
Nile River
Euphrates River
Palestinian Loss of Land – 1946 to 2000
Palestinian Loss of Land – 1946 to 2000
= Palestinian land
= Zionist land
PALEST
NEVE
EXISTE
VE STAN
MULTICU
ISRA
enbordersforis
Israel
SAUDI ARABIA
PALESTINIAN LOSS OF LAND 1946-2010

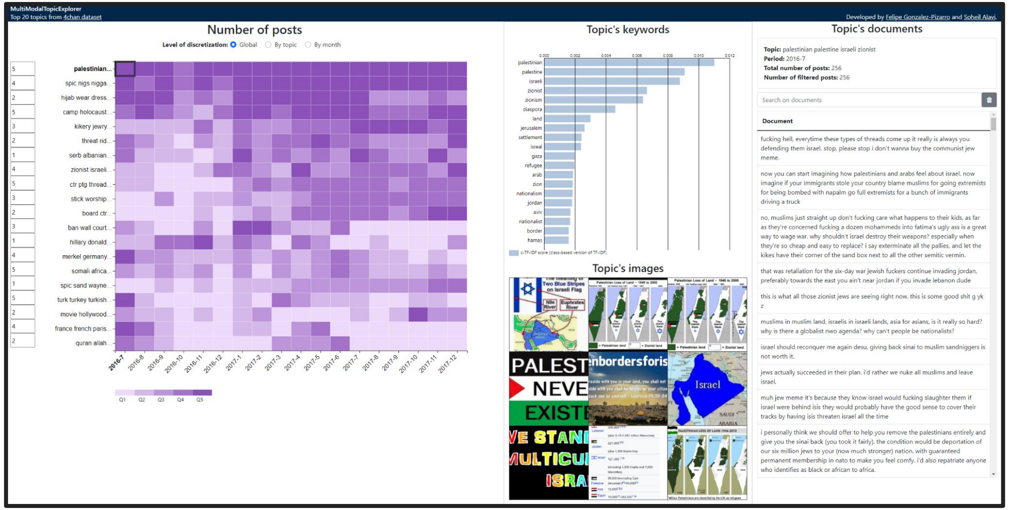
Topic's documents

Topic: palestinian palestine israeli zionist
Period: 2016-7
Total number of posts: 256
Number of filtered posts: 256

Search on documents

Document
fucking hell, everytime these types of threads come up it really is always you defending them israel. stop, please stop i don't wanna buy the communist jew meme.
now you can start imagining how palestinians and arabs feel about israel. now imagine if your immigrants stole your country blame muslims for going extremists for being bombed with napalm go full extremists for a bunch of immigrants driving a truck
no, muslims just straight up don't fucking care what happens to their kids, as far as they're concerned fucking a dozen mohammeds into fatima's ugly ass is a great way to wage war. why shouldn't israel destroy their weapons? especially when they're so cheap and easy to replace? i say exterminate all the pallies, and let the kikes have their corner of the sand box next to all the other semitic vermin.
that was retaliation for the six-day war jewish fuckers continue invading jordan, preferably towards the east you ain't near jordan if you invade lebanon dude
this is what all those zionist jews are seeing right now. this is some good shit g yk z
muslims in muslim land, israelis in israeli lands, asia for asians, is it really so hard? why is there a globalist nwo agenda? why can't people be nationalists?
israel should reconquer me again desu, giving back sinai to muslim sandniggers is not worth it.
jews actually succeeded in their plan. i'd rather we nuke all muslims and leave israel.
muh jew meme it's because they know israel would fucking slaughter them if israel were behind isis they would probably have the good sense to cover their tracks by having isis threaten israel all the time
i personally think we should offer to help you remove the palestinians entirely and give you the sinai back (you took it fairly). the condition would be deportation of our six million jews to your (now much stronger) nation. with guaranteed permanent membership in nato to make you feel comfy. i'd also repatriate anyone who identifies as black or african to africa.

Our proposal: MultiModalTopicExplorer



MultiModalTopicExplorer
Top 20 topics from 4chan dataset

Developed by Felipe Gonzalez-Pizarro and Soheil Alavi.

Number of posts

Level of discretization: ● Global ○ By topic ○ By month

Topic's keywords

Topic's images

The meaning of Two Blue Stripes on Israeli Flag
Nile River
Euphrates River
Palestinian Loss of Land – 1946 to 2000
PALESTINE NEVER EXISTED
Israel
SAUDI ARABIA

Topic's documents

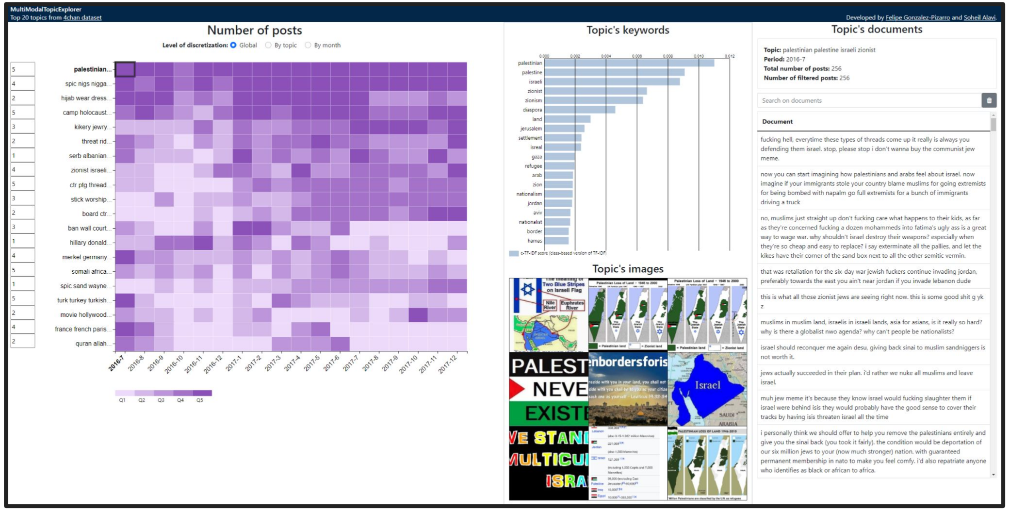
Topic: palestinian palestine israeli zionist
Period: 2016-7
Total number of posts: 256
Number of filtered posts: 256

Search on documents

Document
fucking hell, everytime these types of threads come up it really is always you defending them israel. stop, please stop i don't wanna buy the communist jew meme.
now you can start imagining how palestinians and arabs feel about israel. now imagine if your immigrants stole your country blame muslims for going extremists for being bombed with napalm go full extremists for a bunch of immigrants driving a truck
no, muslims just straight up don't fucking care what happens to their kids, as far as they're concerned fucking a dozen mohammeds into fatima's ugly ass is a great way to wage war. why shouldn't israel destroy their weapons? especially when they're so cheap and easy to replace? i say exterminate all the pallies, and let the kikes have their corner of the sand box next to all the other semitic vermin.
that was retaliation for the six-day war jewish fuckers continue invading jordan, preferably towards the east you ain't near jordan if you invade lebanon dude
this is what all those zionist jews are seeing right now. this is some good shit g yk z
muslims in muslim land, israelis in israeli lands, asia for asians, is it really so hard? why is there a globalist nwo agenda? why can't people be nationalists?
israel should reconquer me again desu, giving back sinai to muslim sandniggers is not worth it.
jews actually succeeded in their plan. i'd rather we nuke all muslims and leave israel.
muh jew meme it's because they know israel would fucking slaughter them if israel were behind isis they would probably have the good sense to cover their tracks by having isis threaten israel all the time
i personally think we should offer to help you remove the palestinians entirely and give you the sinai back (you took it fairly). the condition would be deportation of our six million jews to your (now much stronger) nation. with guaranteed permanent membership in nato to make you feel comfy. i'd also repatriate anyone who identifies as black or african to africa.

- **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


MultiModalTopicExplorer
Top 20 topics from 4chan dataset

Developed by Felipe Gonzalez-Pizarro and Soheil Alavi.

Number of posts

Level of discretization: ● Global ○ By topic ○ By month

Topic's keywords

Topic's images

Topic's documents

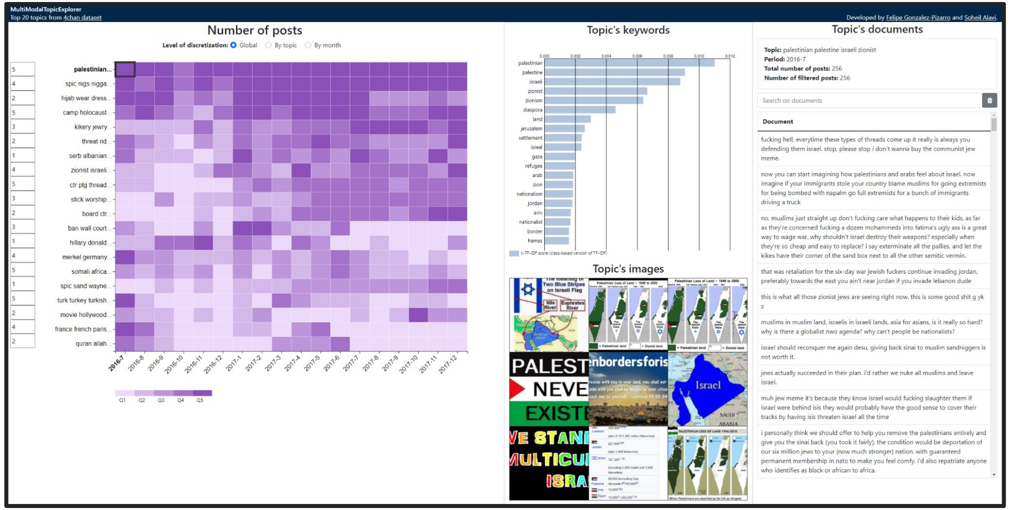
Topic: palestinian palestine israeli zionist
Period: 2016-7
Total number of posts: 256
Number of filtered posts: 256

Search on documents

Document
fucking hell, everytime these types of threads come up it really is always you defending them israel. stop, please stop i don't wanna buy the communist jew meme.
now you can start imagining how palestinians and arabs feel about israel. now imagine if your immigrants stole your country blame muslims for going extremists for being bombed with napalm go full extremists for a bunch of immigrants driving a truck
no, muslims just straight up don't fucking care what happens to their kids, as far as they're concerned fucking a dozen mohammeds into fatima's ugly ass is a great way to wage war. why shouldn't israel destroy their weapons? especially when they're so cheap and easy to replace? i say exterminate all the pallies, and let the kikes have their corner of the sand box next to all the other semitic vermin.
that was retaliation for the six-day war jewish fuckers continue invading jordan, preferably towards the east you ain't near jordan if you invade lebanon dude
this is what all those zionist jews are seeing right now. this is some good shit g yk z
muslims in muslim land, israelis in israeli lands, asia for asians, is it really so hard? why is there a globalist nwo agenda? why can't people be nationalists?
israel should reconquer me again desu, giving back sinai to muslim sandniggers is not worth it.
jews actually succeeded in their plan. i'd rather we nuke all muslims and leave israel.
muh jew meme it's because they know israel would fucking slaughter them if israel were behind isis they would probably have the good sense to cover their tracks by having isis threaten israel all the time
i personally think we should offer to help you remove the palestinians entirely and give you the sinai back (you took it fairly). the condition would be deportation of our six million jews to your (now much stronger) nation. with guaranteed permanent membership in nato to make you feel comfy. i'd also repatriate anyone who identifies as black or african to africa.

- 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



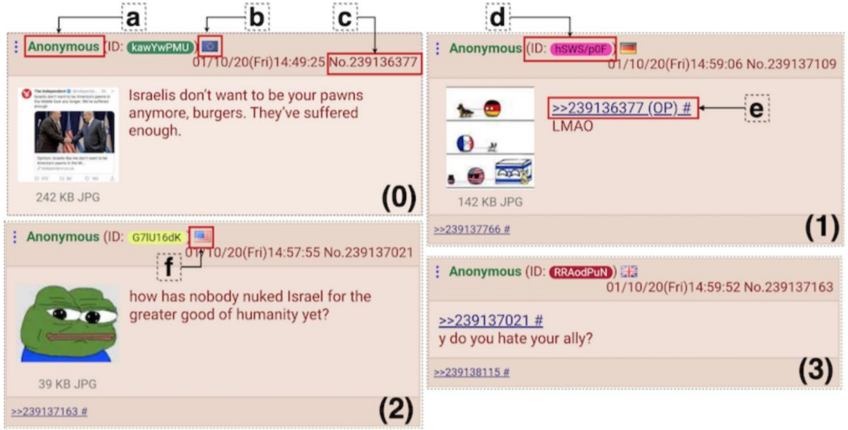
Topic: palestinian palestine israeli zionist
Period: 2016-7
Total number of posts: 256
Number of filtered posts: 256

Document
fucking hell, everytime these types of threads come up it really is always you defending them israel. stop, please stop i don't wanna buy the communist jew meme.
now you can start imagining how palestinians and arabs feel about israel. now imagine if your immigrants stole your country blame muslims for going extremists for being bombed with napalm go full extremists for a bunch of immigrants driving a truck
no, muslims just straight up don't fucking care what happens to their kids, as far as they're concerned fucking a dozen mohammeds into fatima's ugly ass is a great way to wage war. why shouldn't israel destroy their weapons? especially when they're so cheap and easy to replace? i say exterminate all the pallies, and let the kikes have their corner of the sand box next to all the other semitic vermin.
that was retaliation for the six-day war jewish fuckers continue invading jordan, preferably towards the east you ain't near jordan if you invade lebanon dude
this is what all those zionist jews are seeing right now. this is some good shit g yk z
muslims in muslim land, israelis in israeli lands, asia for asians, is it really so hard? why is there a globalist nwo agenda? why can't people be nationalists?
israel should reconquer me again desu, giving back sinai to muslim sandniggers is not worth it.
jews actually succeeded in their plan. i'd rather we nuke all muslims and leave israel.
muh jew meme it's because they know israel would fucking slaughter them if israel were behind isis they would probably have the good sense to cover their tracks by having isis threaten israel all the time
i personally think we should offer to help you remove the palestinians entirely and give you the sinai back (you took it fairly). the condition would be deportation of our six million jews to your (now much stronger) nation. with guaranteed permanent membership in nato to make you feel comfy. i'd also repatriate anyone who identifies as black or african to africa.

- 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



Anonymous (ID: kawYwPMU) 01/10/20(Fri)14:49:25 No.239136377

242 KB JPG

Israelis don't want to be your pawns anymore, burgers. They've suffered enough.

Anonymous (ID: hSWS/p0F) 01/10/20(Fri)14:59:06 No.239137109

142 KB JPG

>>239136377 (OP) #
LMAO

>>239137766 #

Anonymous (ID: G7lU16dK) 01/10/20(Fri)14:57:55 No.239137021

39 KB JPG

how has nobody nuked Israel for the greater good of humanity yet?

>>239137163 #

Anonymous (ID: RRAodPuN) 01/10/20(Fri)14:59:52 No.239137163

>>239137021 #
y do you hate your ally?

>>239138115 #

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

arXiv:2001.07487v2 [cs.CY] 1 Apr 2020

Raiders of the Lost Kek: 3.5 Years of Augmented 4chan Posts from the Politically Incorrect Board*

Antonis Papasavva[1,☞], Savvas Zannettou[2,☞], Emiliano De Cristofaro[1,☞],
Gianluca Stringhini[3,☞], and Jeremy Blackburn[4,☞]
[1]University College London, [2]Max-Planck-Institut für Informatik,
[3]Boston University, [4]Binghamton University, [☞]iDRAMA Lab
{antonis.papasavva.19, e.decristofaro}@ucl.ac.uk, szannett@mpi-inf.mpg.de,
gian@bu.edu, jblackbu@binghamton.edu

Abstract

This paper presents a dataset with over 3.3M threads and 134.5M posts from the Politically Incorrect board (/pol/) of the imageboard forum 4chan, posted over a period of almost 3.5 years (June 2016–November 2019). To the best of our knowledge, this represents the largest publicly available 4chan dataset, providing the community with an archive of posts that have been permanently deleted from 4chan and are otherwise inaccessible. We augment the data with a set of additional labels, including toxicity scores and the named entities mentioned in each post. We also present a statistical analysis of the dataset, providing an overview of what researchers interested in using it can expect, as well as a simple content analysis, shedding light on the most prominent discussion topics, the most popular entities mentioned, and the toxicity level of each post. Overall, we are confident that our work will motivate and assist researchers in studying and understanding 4chan, as well as its role on the greater Web. For instance, we hope this dataset may be used for cross-platform studies of social media, as well as being useful for other types of research like natural language processing. Finally, our dataset can assist qualitative work focusing on in-depth case studies of specific narratives, events, or social theories.

1 Introduction

Modern society increasingly relies on the Internet for a wide range of tasks, including gathering, sharing, and commenting on content, events, and discussions. Alas, the Web has also enabled anti-social and toxic behavior to occur at an unprecedented scale. Malevolent actors routinely exploit social networks to target other users via hate speech and abusive behavior, or spread extremist ideologies [3, 12, 13, 40].

A non-negligible portion of these nefarious activities often originate on "fringe" online platforms, e.g., 4chan, 8chan, Gab. In fact, research has shown how influential 4chan is in spreading disinformation [11, 43], hateful memes [42], and coordinating harassment campaigns on other platforms [21, 25, 34]. These platforms are also linked to various real-world violent events, including the radicalization of users who committed mass shootings [2, 6, 16].

4chan is an imageboard where users (aka Original Posters, or OPs) can create a thread by posting an image and a message to a board; others can post in the OP's thread, with a message and/or an image. Among 4chan's key features are anonymity and ephemerality; users do not need to register to post content, and in fact the overwhelming majority of posts are anonymous. At most, threads are archived after they become inactive and deleted within 7 days.

Overall, 4chan is widely known for the large amount of content, memes, slang, and Internet culture it has generated over the years [15]. For example, 4chan popularized the "lolcat" meme on the early Web. More recently, politically charged memes, e.g., "God Emperor Trump" [24] have also originated on the platform.

Data Release. In this work, we focus on the *"Politically Incorrect" board* (/pol/),[1] given the interest it has generated in prior research and the influential role it seems to play on the rest of the Web [7, 21, 43, 34, 42, 39]. Along with the paper, we release a dataset [44] including 134.5M posts from over 3.3M /pol/ conversation threads, made over a period of approximately 3.5 years (June 2016–November 2019). Each post in our dataset has the text provided by the poster, along with various post metadata (e.g., post id, time, etc.).

We also *augment* the dataset by attaching additional set of labels to each post, including: 1) the named entities mentioned in the post, and 2) the toxicity scores of the post. For the former, we use the spaCy library [35], and for the latter, Google's Perspective API [30].

We also wish to warn the readers that some of the content in our dataset, as well as in this paper, is highly toxic, racist, and hateful, and can be rather disturbing.

Relevance. We are confident that our dataset will be useful to the research community in several ways. First, /pol/ con-

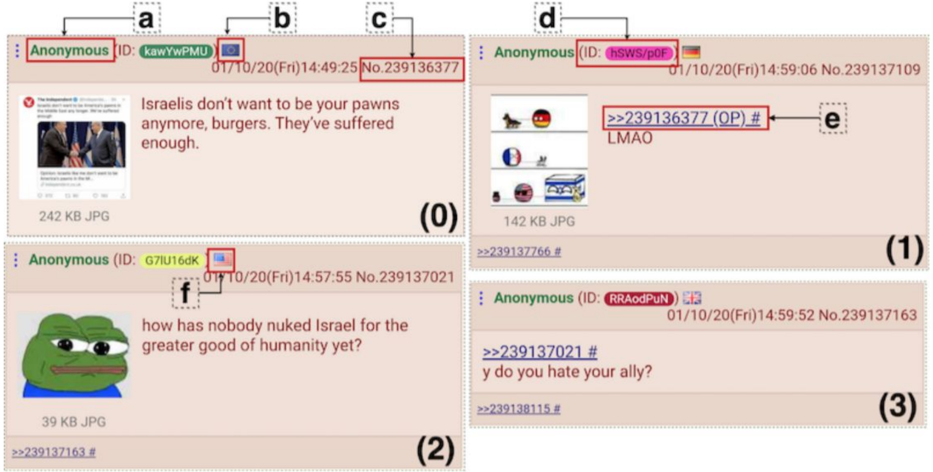
*Published at the 14th International AAAI Conference on Web and Social Media (ICWSM 2020). Please cite the ICWSM version.

[1] http://boards.4chan.org/pol/

1

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

Why 4chan dataset?



Anonymous (ID: kawYwPMU) 01/10/20(Fri)14:49:25 No.239136377

Israelis don't want to be your pawns anymore, burgers. They've suffered enough.

242 KB JPG

(0)

Anonymous (ID: hSWS/p0F) 01/10/20(Fri)14:59:06 No.239137109

>>239136377 (OP) #
LMAO

142 KB JPG

>>239137766 #

(1)

Anonymous (ID: G7lU16dK) 01/10/20(Fri)14:57:55 No.239137021

how has nobody nuked Israel for the greater good of humanity yet?

39 KB JPG

>>239137163 #

(2)

Anonymous (ID: RRAodPuN) 01/10/20(Fri)14:59:52 No.239137163

>>239137021 #
y do you hate your ally?

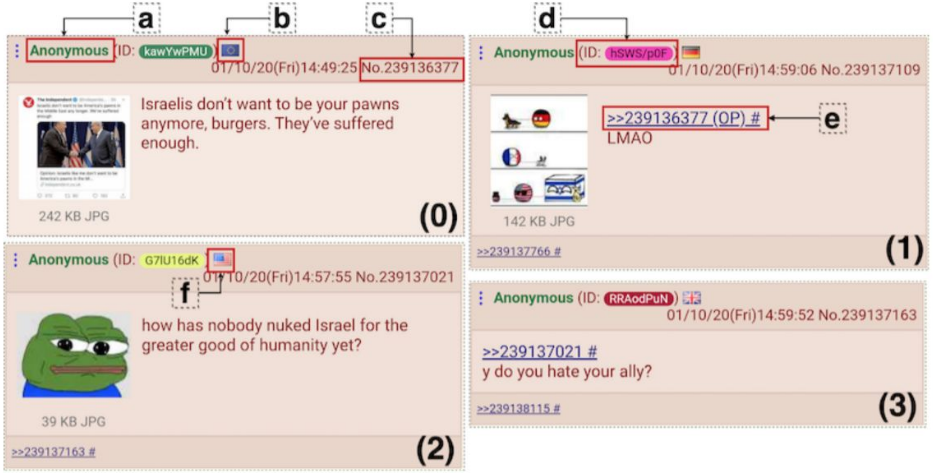
>>239138115 #

(3)

- 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

Base Dataset



Anonymous (ID: kawYwPMU) 01/10/20(Fri)14:49:25 No.239136377

Israelis don't want to be your pawns anymore, burgers. They've suffered enough.

242 KB JPG

Anonymous (ID: hSWS/p0F) 01/10/20(Fri)14:59:06 No.239137109

>>239136377 (OP) #
LMAO

142 KB JPG

>>239137766 #

Anonymous (ID: G7lU16dK) 01/10/20(Fri)14:57:55 No.239137021

how has nobody nuked Israel for the greater good of humanity yet?

39 KB JPG

>>239137163 #

Anonymous (ID: RRAodPuN) 01/10/20(Fri)14:59:52 No.239137163

>>239137021 #
y do you hate your ally?

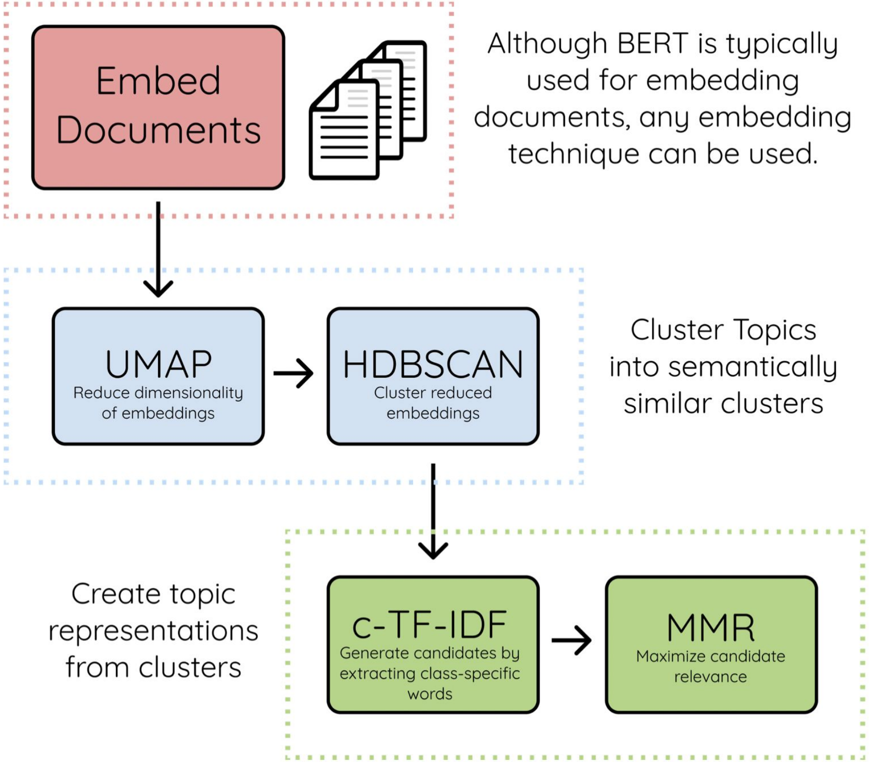
>>239138115 #

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

BERTOPIC



Embed Documents

Although BERT is typically used for embedding documents, any embedding technique can be used.

UMAP

Reduce dimensionality of embeddings

HDBSCAN

Cluster reduced embeddings

Cluster Topics into semantically similar clusters

Create topic representations from clusters

c-TF-IDF

Generate candidates by extracting class-specific words

MMR

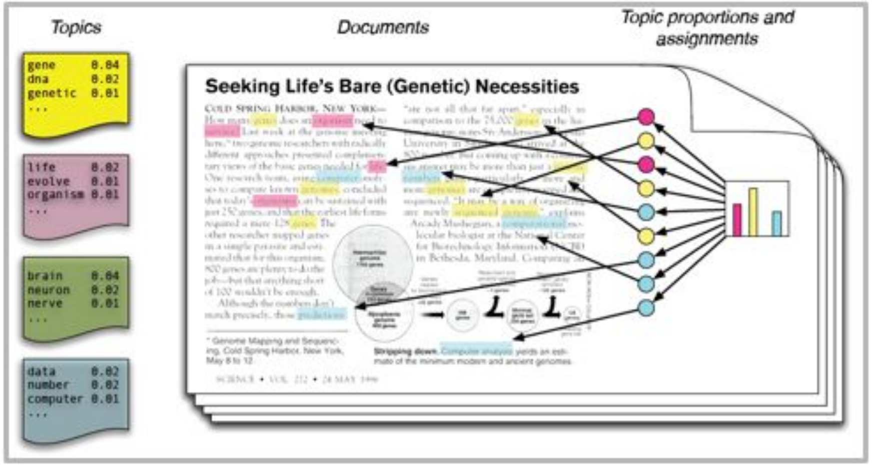
Maximize candidate relevance

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

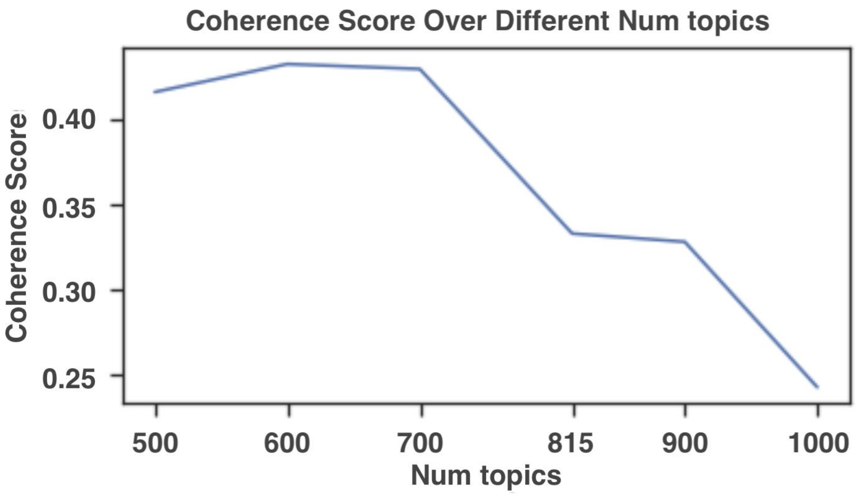
LDA

- 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**

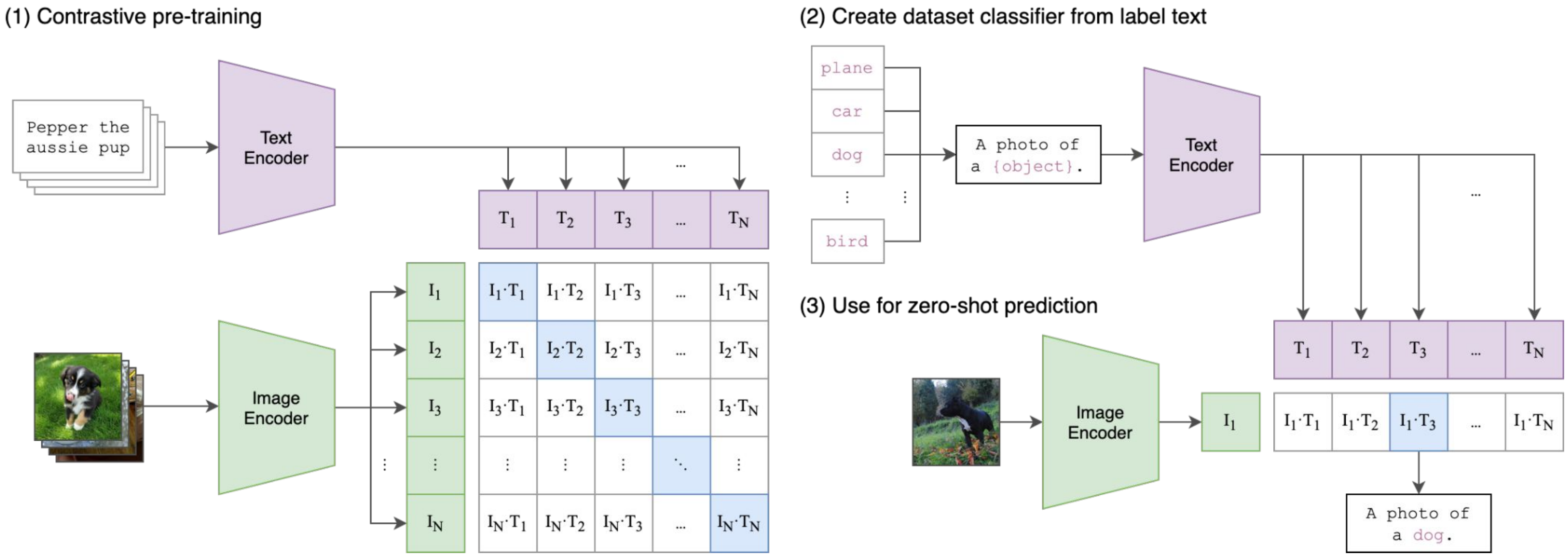


LDA-Implementation

- 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



CLIP: Contrastive Language-Image Pre-Training



	T_1	T_2	T_3	...	T_N
I_1	$I_1{\cdot}T_1$	$I_1{\cdot}T_2$	$I_1{\cdot}T_3$	...	$I_1{\cdot}T_N$
I_2	$I_2{\cdot}T_1$	$I_2{\cdot}T_2$	$I_2{\cdot}T_3$	...	$I_2{\cdot}T_N$
I_3	$I_3{\cdot}T_1$	$I_3{\cdot}T_2$	$I_3{\cdot}T_3$	...	$I_3{\cdot}T_N$
⋮	⋮	⋮	⋮	⋱	⋮
I_N	$I_N{\cdot}T_1$	$I_N{\cdot}T_2$	$I_N{\cdot}T_3$	...	$I_N{\cdot}T_N$

Neural network trained on a variety of (image, text) pairs.

Source: https://github.com/openai/CLIP

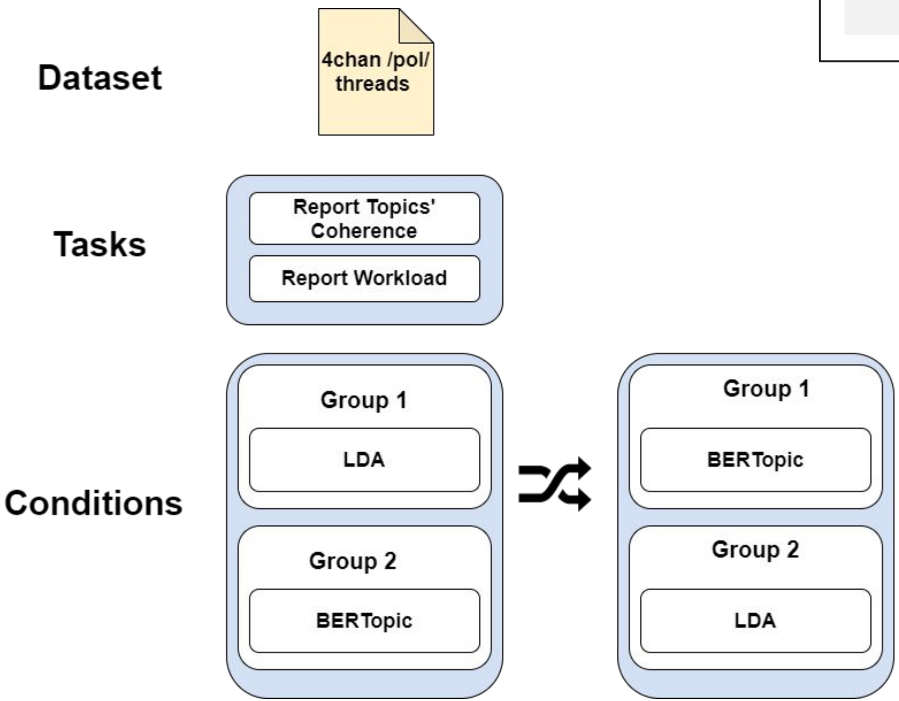
Content Warning

*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

Group 2

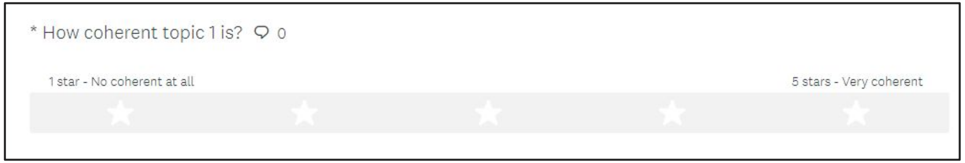
BERTopic

Group 1

BERTopic

Group 2

LDA

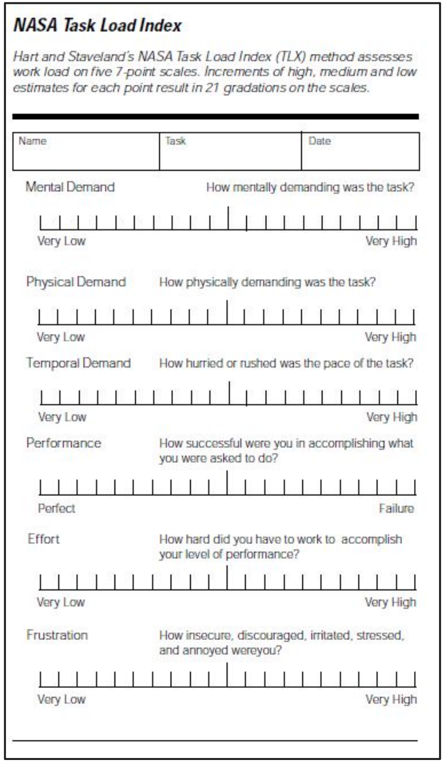


* How coherent topic 1 is?

1 star - No coherent at all

5 stars - Very coherent

☆ ☆ ☆ ☆ ☆



NASA Task Load Index

Hart and Staveland's NASA Task Load Index (TLX) method assesses work load on five 7-point scales. Increments of high, medium and low estimates for each point result in 21 gradations on the scales.

Name	Task	Date

Mental Demand — How mentally demanding was the task?

Very Low — Very High

Physical Demand — How physically demanding was the task?

Very Low — Very High

Temporal Demand — How hurried or rushed was the pace of the task?

Very Low — Very High

Performance — How successful were you in accomplishing what you were asked to do?

Perfect — Failure

Effort — How hard did you have to work to accomplish your level of performance?

Very Low — Very High

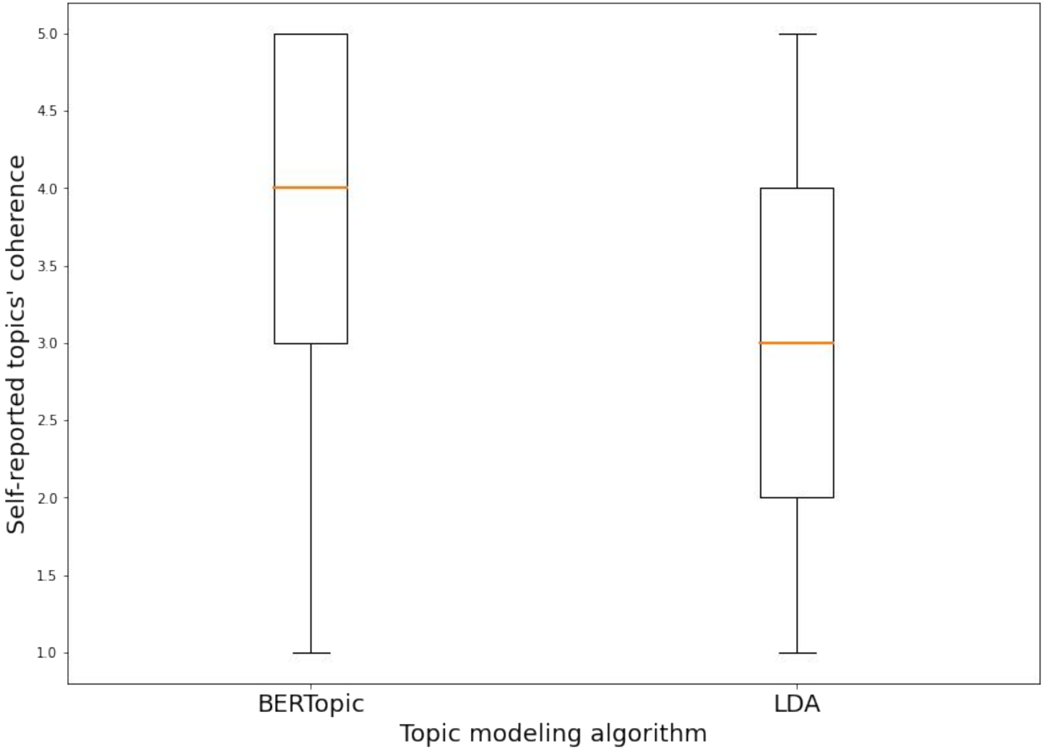
Frustration — How insecure, discouraged, irritated, stressed, and annoyed wereyou?

Very Low — Very High

Null hypothesis (H_0)

H_0: There are no differences in the coherence of topics emerged from BERTopic and LDA

Results - Quality of topics



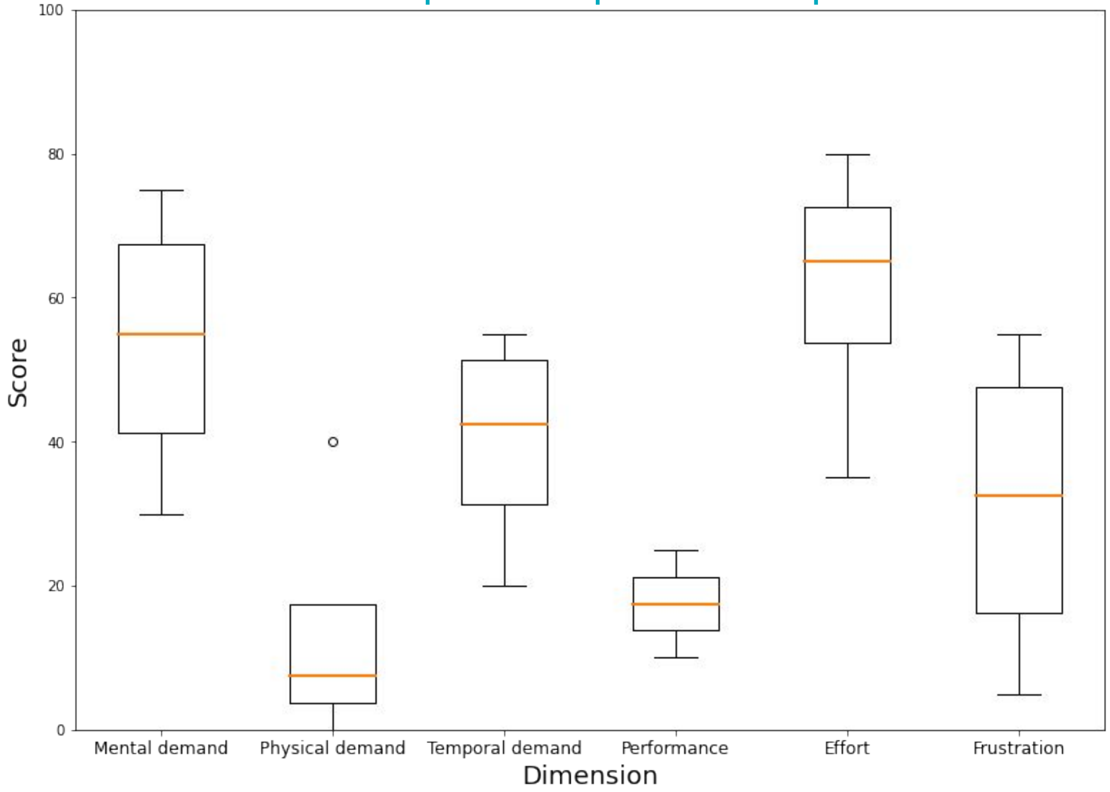
We found statistically significant differences between conditions

(U = 4811.0 $N_{bertopic}$=80,N_{lda} = 84, **p = <.001**)

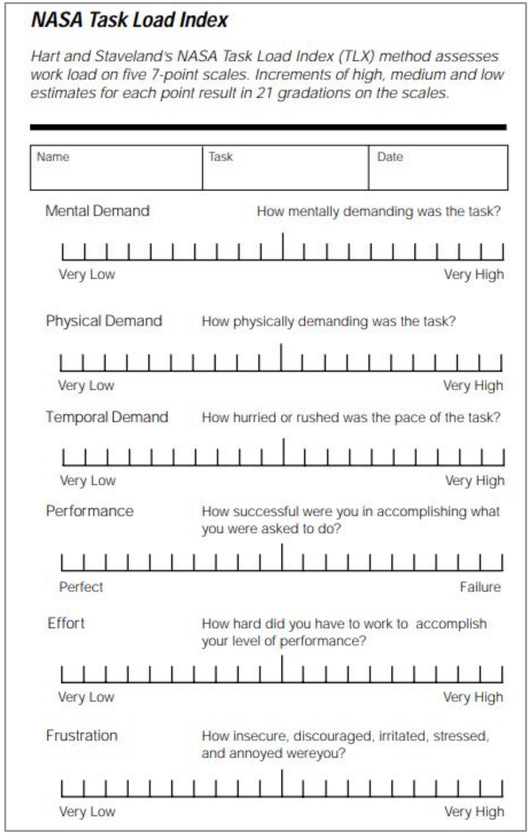
We can reject Null hypothesis

H0: There are no differences in the coherence of topics emerged from BERTopic and LDA

Distribution of participant responses to the NASA TLX questionnaire



Cao, A., Chintamani, K. K., Pandya, A. K., & Ellis, R. D. (2009). NASA TLX: Software for assessing subjective mental workload. *Behavior research methods*, *41*(1), 113-117.



NASA Task Load Index

Hart and Staveland's NASA Task Load Index (TLX) method assesses work load on five 7-point scales. Increments of high, medium and low estimates for each point result in 21 gradations on the scales.

Name	Task	Date

Mental Demand — How mentally demanding was the task?

Very Low — Very High

Physical Demand — How physically demanding was the task?

Very Low — Very High

Temporal Demand — How hurried or rushed was the pace of the task?

Very Low — Very High

Performance — How successful were you in accomplishing what you were asked to do?

Perfect — Failure

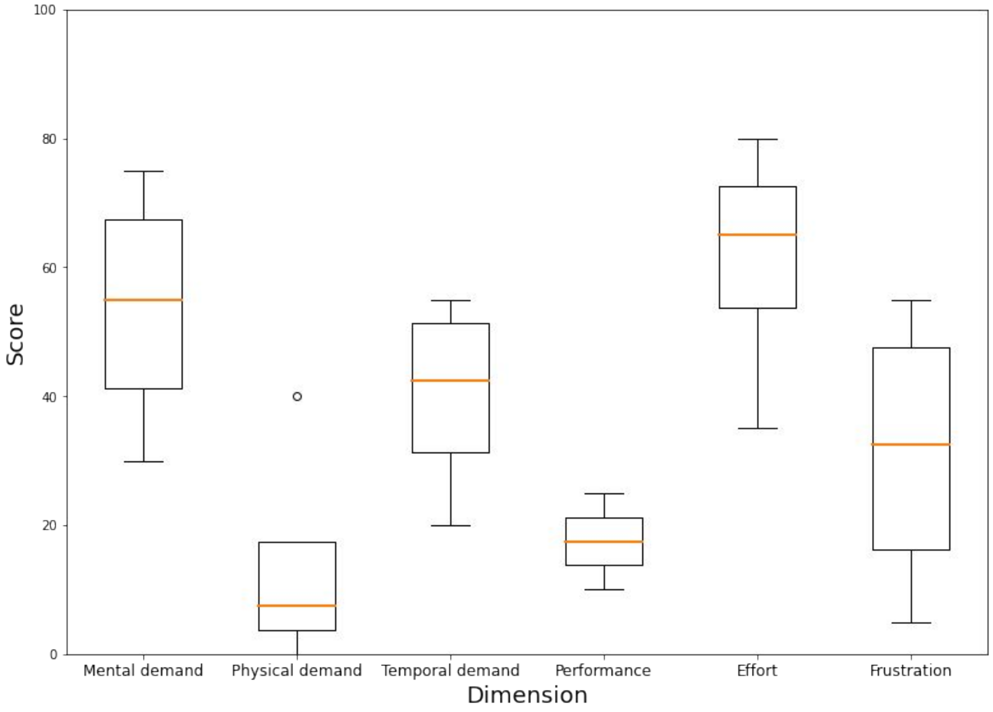
Effort — How hard did you have to work to accomplish your level of performance?

Very Low — Very High

Frustration — How insecure, discouraged, irritated, stressed, and annoyed wereyou?

Very Low — Very High

Distribution of participant responses to the NASA TLX questionnaire



A lower score indicates a better result

The results **hint** that:

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

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 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! :)