UNIVERSITÀ DEGLI STUDI DI MILANO

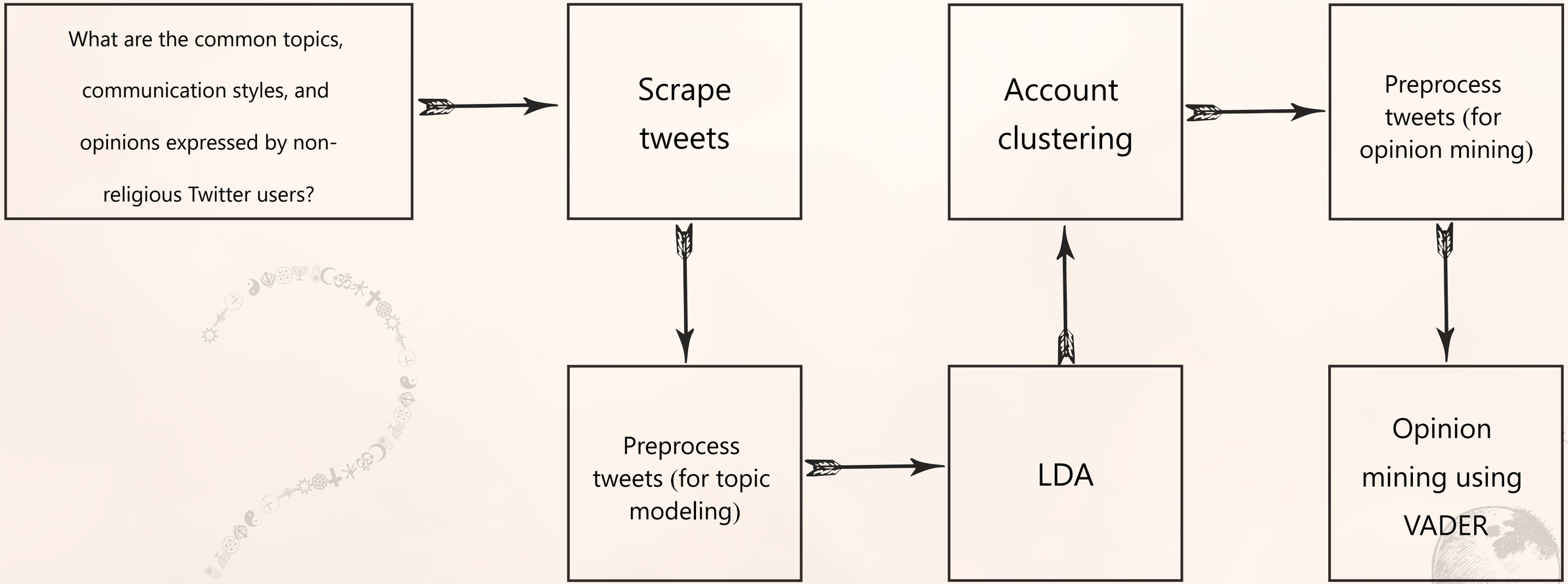
LA STATALE

In God we trust

Information Retrieval

Vojimir Ranitovic 963780

Map of the project



Dataset and preprocessing

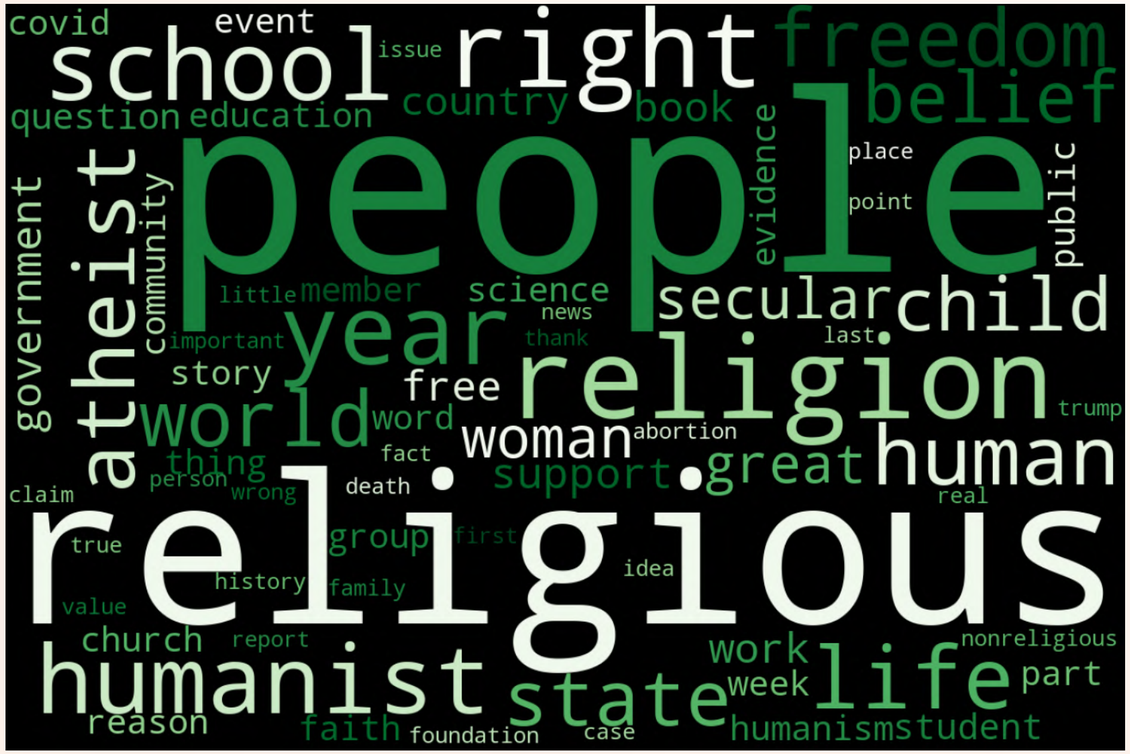
- The dataset contains **1,224,446** scraped tweets in the period from 2007-01-01 to 2023-01-30.
- For quicker analysis, only tweets from 2020-01-01 were used, which is **195,053** instances.
- In preprocessing phase stopwords, links, mentions, hashtags, punctuations, and numbers were removed.
- Lemmatization of unigrams and bigrams.
- Lemmatization of nouns and adjectives only.

Preprocessing

- "@juss_professor @kathrynfenlodge @AccordCoalition A good take away from today's session, says @JohnAdenitire, senior Lecturer @QMSchoolofLaw is: collective worship undermines the notion that the state must provide objective and pluralistic education. 'Compulsory worship interferes with freedom of expression and conscience'."

- ['today', 'session', 'senior', 'lecturer', 'undermine', 'notion', 'state', 'objective', 'pluralistic', 'education', 'compulsory', 'worship', 'freedom', 'expression', 'conscience']

Top unigrams and bigrams



Latent Dirichlet Allocation

Important parts of LDA:

- **id2word** is a dictionary that maps words in the corpus to unique integer ids.
- **corpus** is a list of documents, where each document is represented as a list of integer word ids. LDA treats each document as a bag of words, meaning it doesn't consider the order in which the words appear within the document.

Latent Dirichlet Allocation

Important parts of LDA:

- **id2word** is a dictionary that maps words in the corpus to unique integer ids.
- **corpus** is a list of documents, where each document is represented as a list of integer word ids. LDA treats each document as a bag of words, meaning it doesn't consider the order in which the words appear within the document.

Important parameters of LDA:

- **Alpha** is a hyperparameter that controls the sparsity of the topic distribution for each document. A lower alpha value will lead to a more sparse distribution, meaning that fewer topics will be assigned to each document.
- **Eta**: This is a hyperparameter that controls the sparsity of the word distribution for each topic. A lower eta value will lead to a more sparse distribution, meaning that fewer words will be assigned to each topic.

Latent Dirichlet Allocation

- LDA starts by **randomly** assigning each word in the corpus to a random topic. A topic is simply a probability distribution over words.
- LDA then iteratively refines the topic assignments for each word in the corpus, until it finds a set of topics that best explains the corpus. This process is called **inference**.
- During inference, LDA considers each word in the corpus and looks at the topics of the other words in the same document. It then updates the topic assignment of the current word based on these context words.
- LDA **repeats** this process for all words in the corpus, and over multiple **iterations**, refines the topic assignments until **convergence**.
- Once inference is complete, LDA **returns the learned topics** as a probability distribution over words. These topics represent the underlying themes or concepts that are present in the corpus.

LDA hyperparameter tuning

- Beside alpha and eta there are also chunksize, passes, iterations etc.
- For parameters tuning only alpha, eta and number of topics were chosen.
- **Number of topics** = range(20, 65, 5)
- **Alpha** = (0.001, 1, 0.001)
- **Eta** = (0.001, 0.5, 0.001)
- **Random search** with 300 iterations.

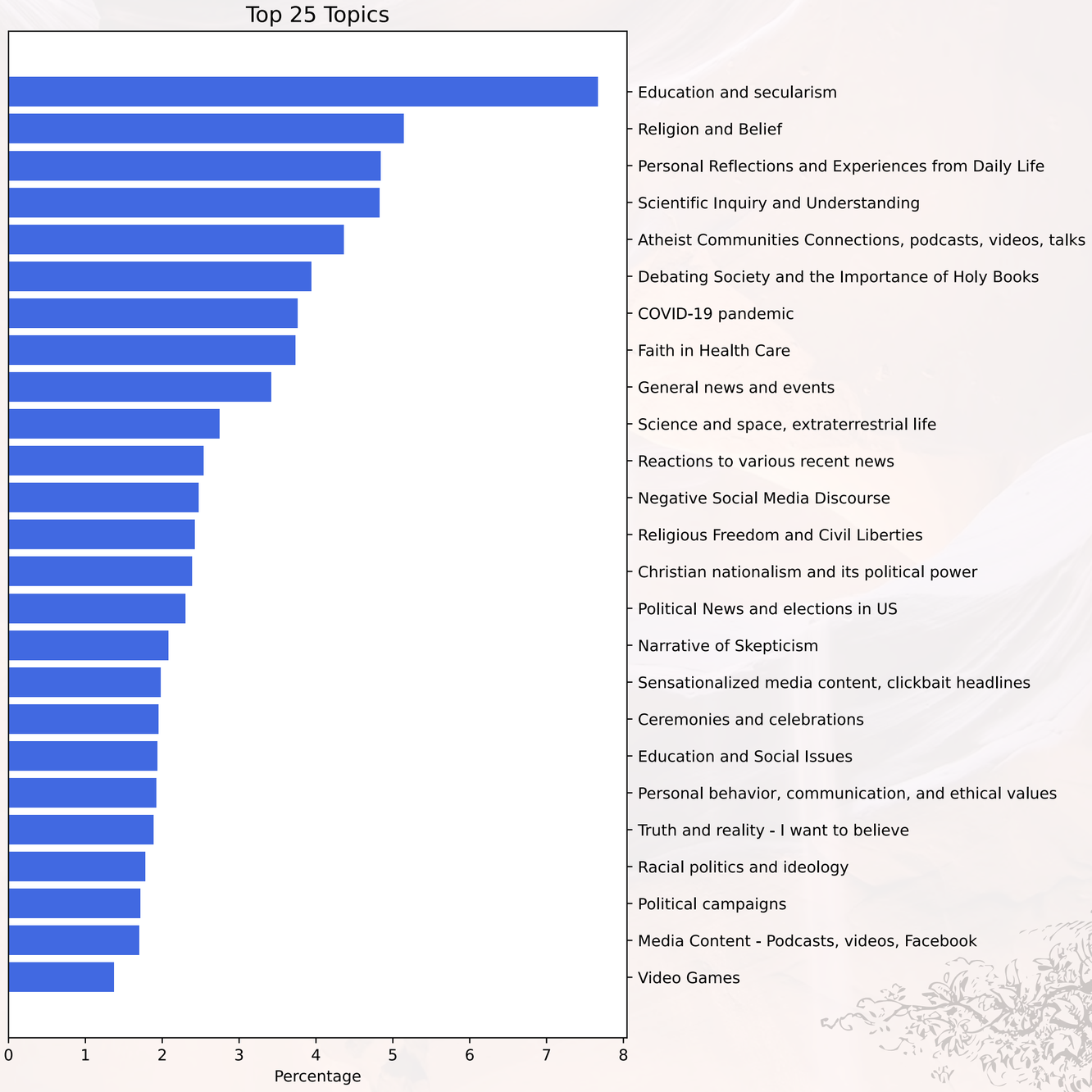
LDA hyperparameter tuning

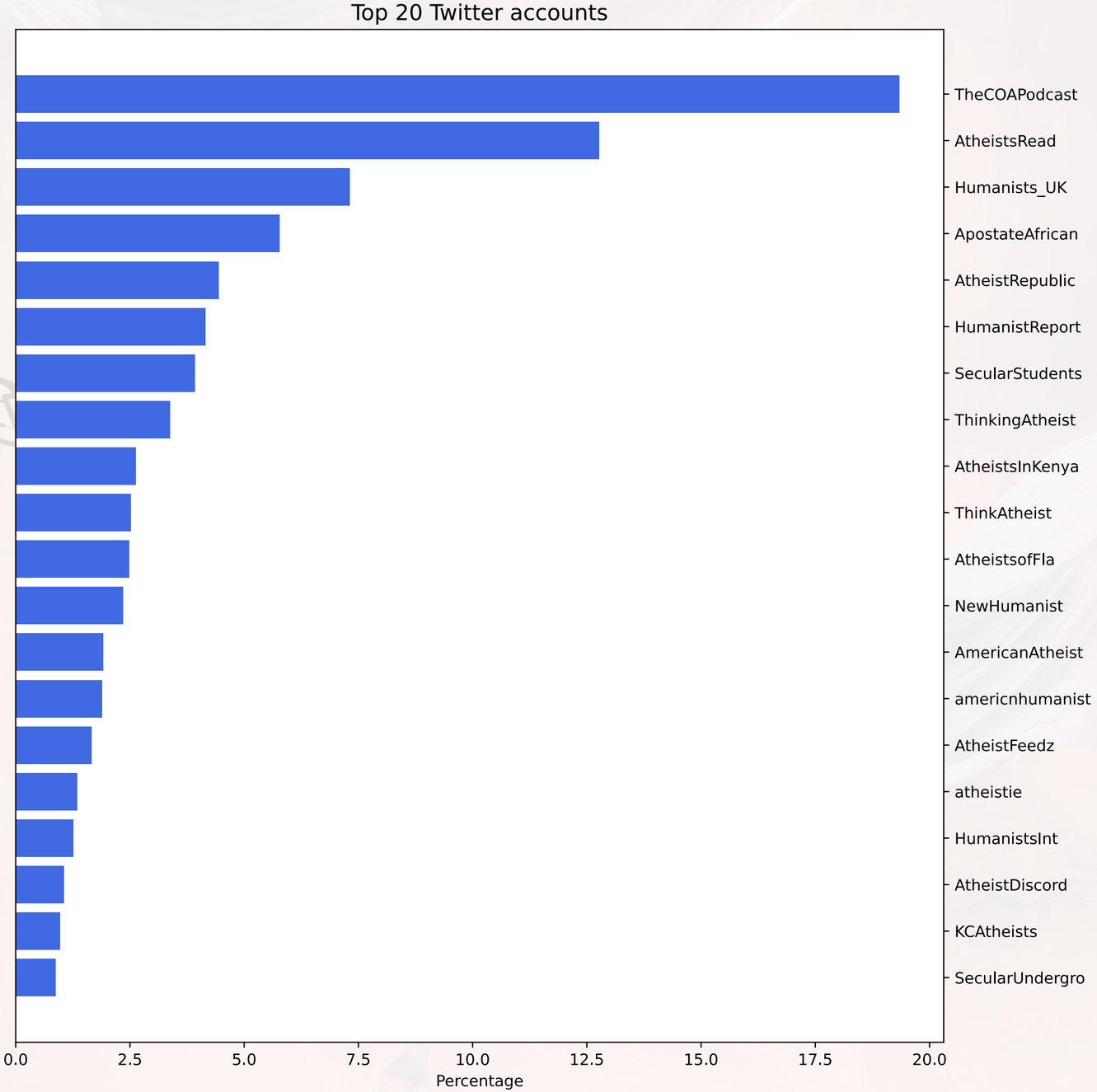
- Beside alpha and eta there are also chunksize, passes, iterations etc.
- For parameters tuning only alpha, eta and number of topics were chosen.
- **Number of topics** = range(20, 65, 5)
- **Alpha** = (0.001, 1, 0.001)
- **Eta** = (0.001, 0.5, 0.001)
- **Random search** with 300 iterations.
- **Return the best** model according to coherence value.
- C_V measures how semantically related are the words within a topic, and how distinct topics are from one another. Coherence measures compute the similarity between the top words in a topic or between the top words in a pair of topics, and usually, **the higher coherence** scores indicate that the words in a topic are **more semantically related** and that the topics are more distinct from other ones.

LDA hyperparameter tuning

- Best model → 60 topics, alpha=0.429, eta=0.464, cv=0.479

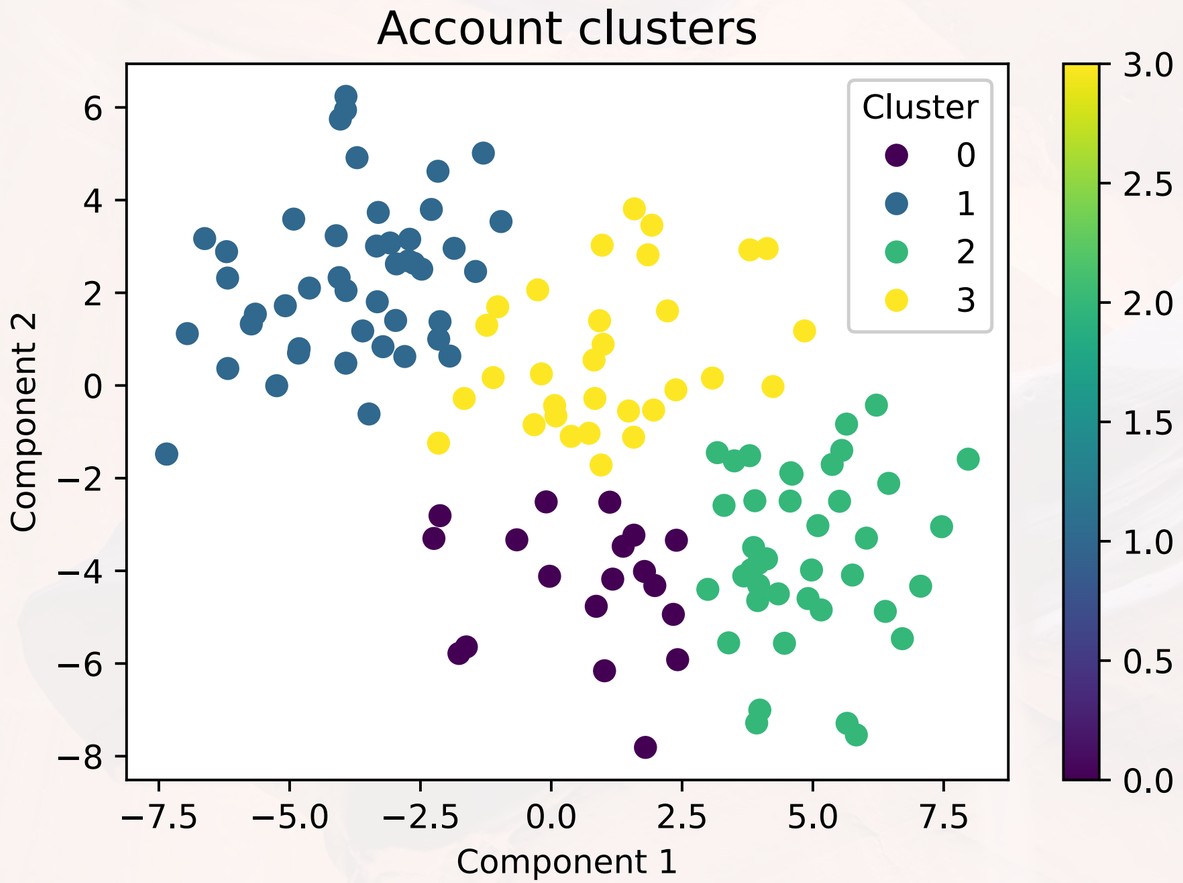
- (6, '0.055*"right" + 0.045*"secular" + 0.043*"school" + 0.030*"state" + 0.025*"public" + 0.023*"government" + 0.023*"member" + 0.022*"child" + 0.019*"free" + 0.019*"education" + 0.018*"group" + 0.018*"support" + 0.015*"student" + 0.013*"bill" + 0.011*"organization" + 0.011*"catholic" + 0.011*"action" + 0.011*"parent" + 0.011*"help" + 0.011*"court" + 0.009*"policy" + 0.009*"service" + 0.009*"discrimination" + 0.009*"legal" + 0.008*"american" + 0.008*"prayer" + 0.008*"national" + 0.008*"official" + 0.008*"high" + 0.007*"coalition"')
- (32, '0.046*"covid" + 0.035*"vaccine" + 0.034*"report" + 0.029*"country" + 0.028*"church" + 0.026*"case" + 0.025*"death" + 0.019*"family" + 0.018*"pandemic" + 0.013*"leader" + 0.011*"blasphemy" + 0.011*"nation" + 0.010*"evangelical" + 0.009*"conspiracy_theorie" + 0.009*"coronavirus" + 0.008*"threat" + 0.008*"able" + 0.008*"effort" + 0.008*"bad" + 0.008*"global" + 0.008*"conspiracy" + 0.007*"vaccination" + 0.007*"dangerous" + 0.007*"crisis" + 0.006*"majority" + 0.006*"antivaxxer" + 0.006*"million" + 0.006*"big" + 0.006*"test" + 0.005*"fight"')


Top 25 Topics
Education and secularism
Religion and Belief
Personal Reflections and Experiences from Daily Life
Scientific Inquiry and Understanding
Atheist Communities Connections, podcasts, videos, talks
Debating Society and the Importance of Holy Books
COVID-19 pandemic
Faith in Health Care
General news and events
Science and space, extraterrestrial life
Reactions to various recent news
Negative Social Media Discourse
Religious Freedom and Civil Liberties
Christian nationalism and its political power
Political News and elections in US
Narrative of Skepticism
Sensationalized media content, clickbait headlines
Ceremonies and celebrations
Education and Social Issues
Personal behavior, communication, and ethical values
Truth and reality - I want to believe
Racial politics and ideology
Political campaigns
Media Content - Podcasts, videos, Facebook
Video Games
0
1
2
3
4
5
6
7
8
Percentage

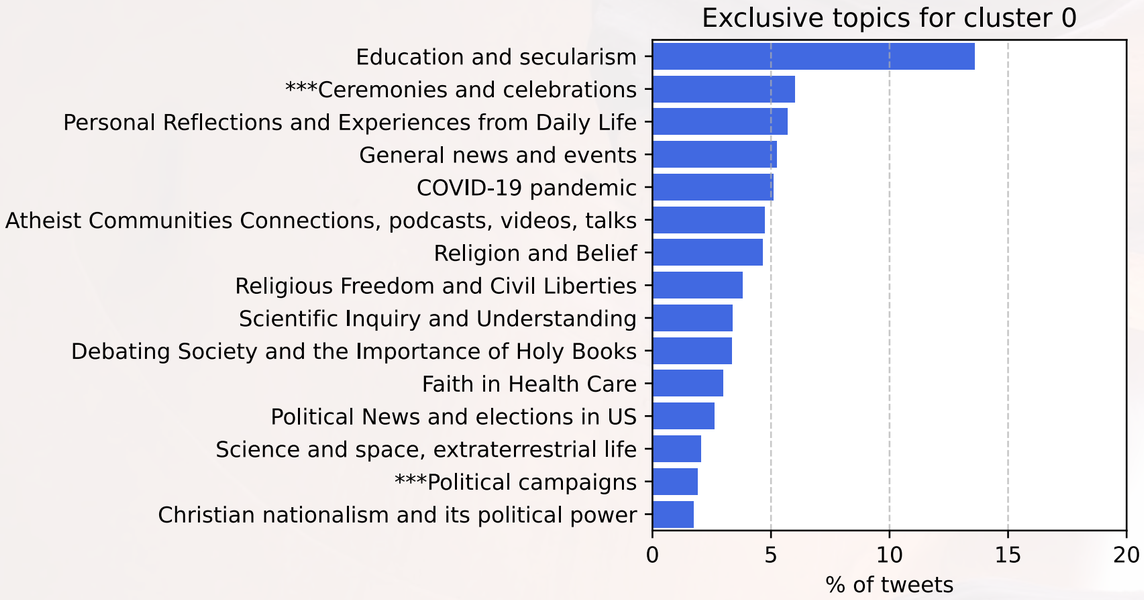

Top 20 Twitter accounts
TheCOAPodcast
AtheistsRead
Humanists_UK
ApostateAfrican
AtheistRepublic
HumanistReport
SecularStudents
ThinkingAtheist
AtheistsInKenya
ThinkAtheist
AtheistsofFla
NewHumanist
AmericanAtheist
americnhumanist
AtheistFeedz
atheistie
HumanistsInt
AtheistDiscord
KCAtheists
SecularUndergro
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
Percentage

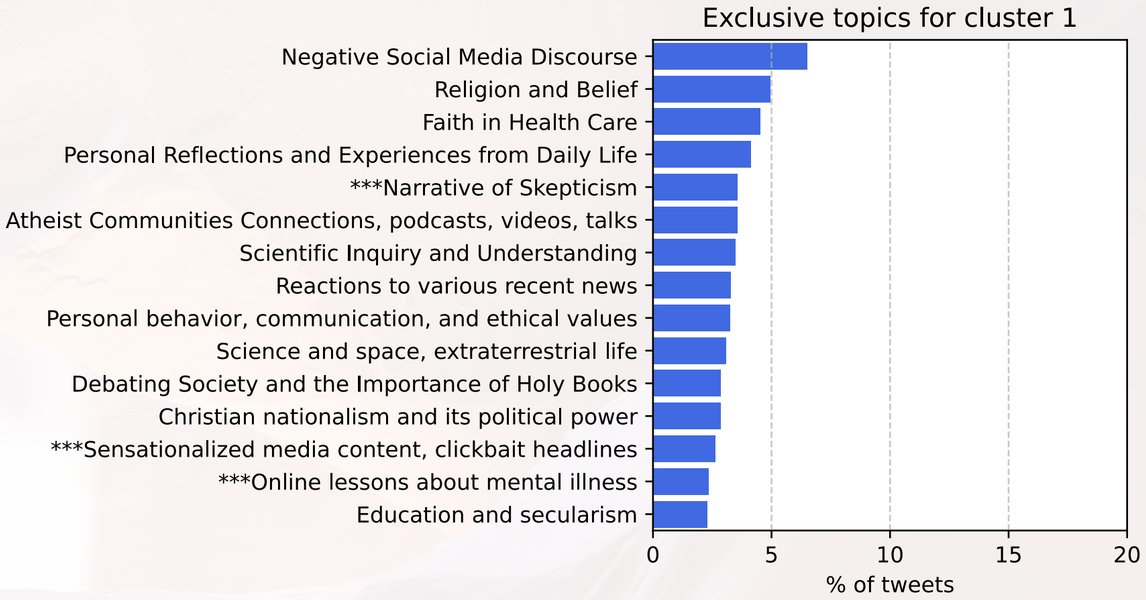
Clustering Twitter accounts

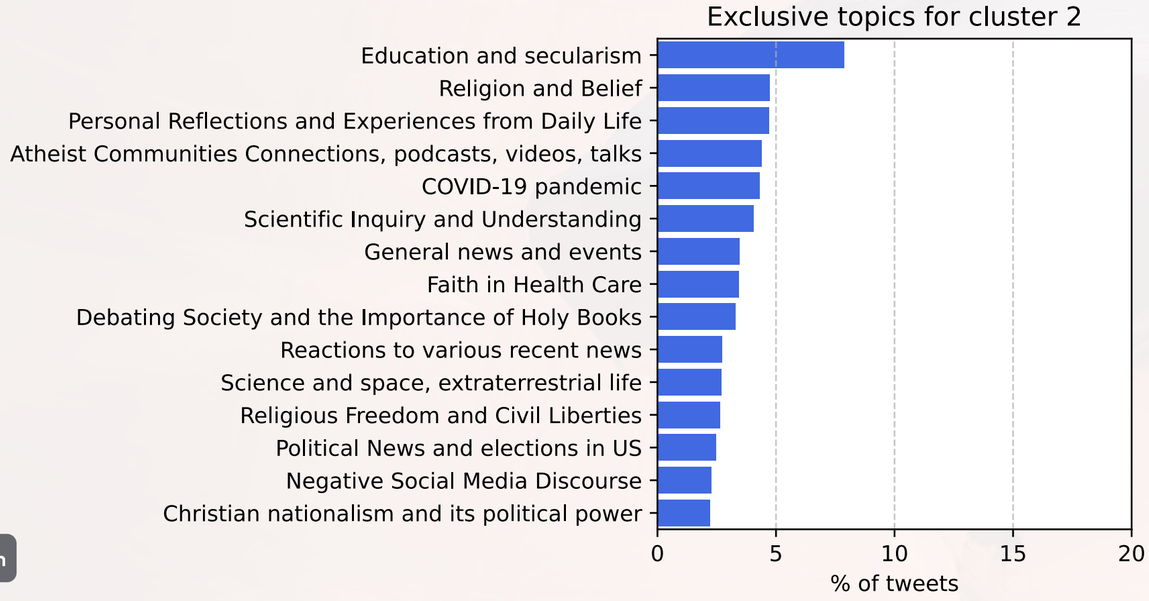
- After finding the topic distribution for each account, they are clustered into 4 groups.

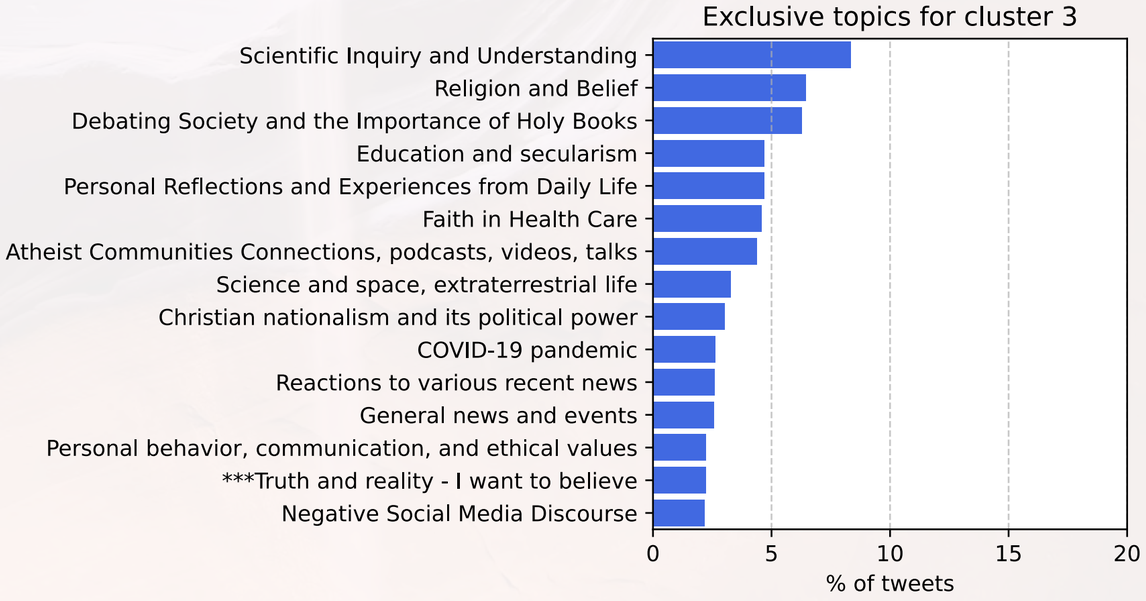


Clustering Twitter accounts

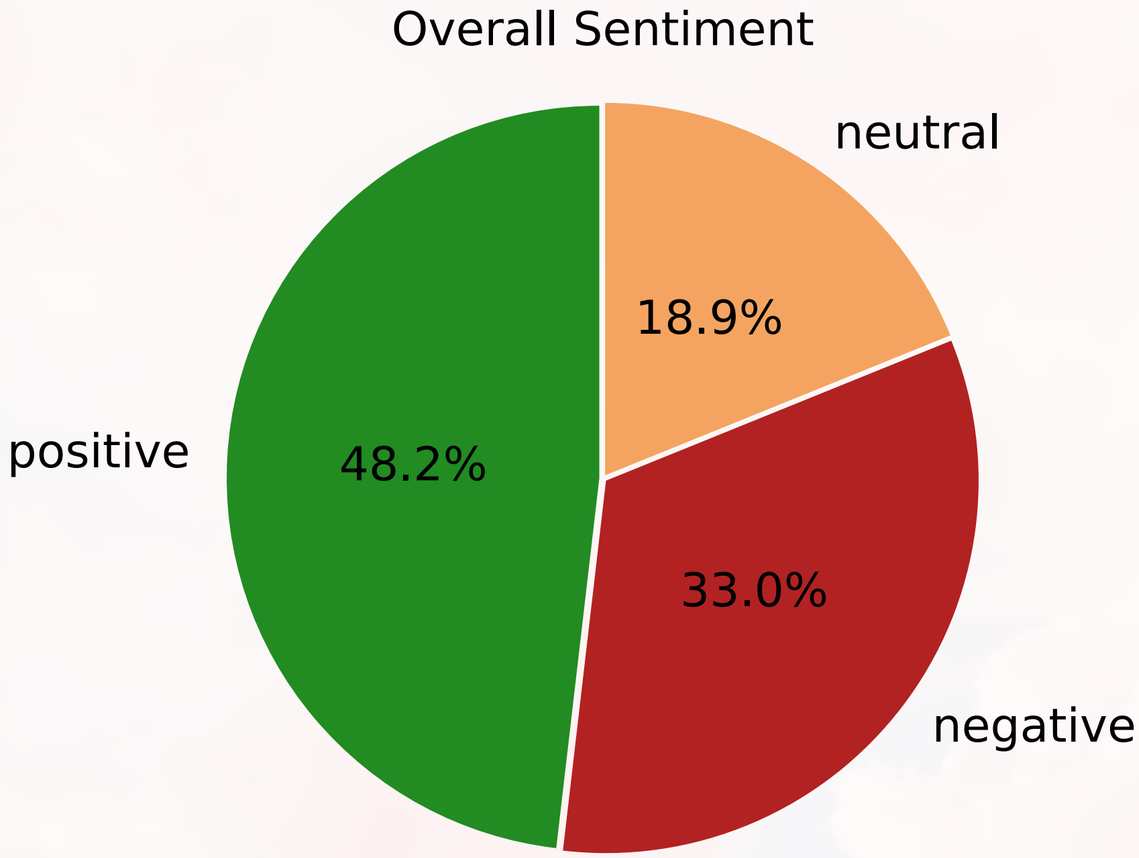




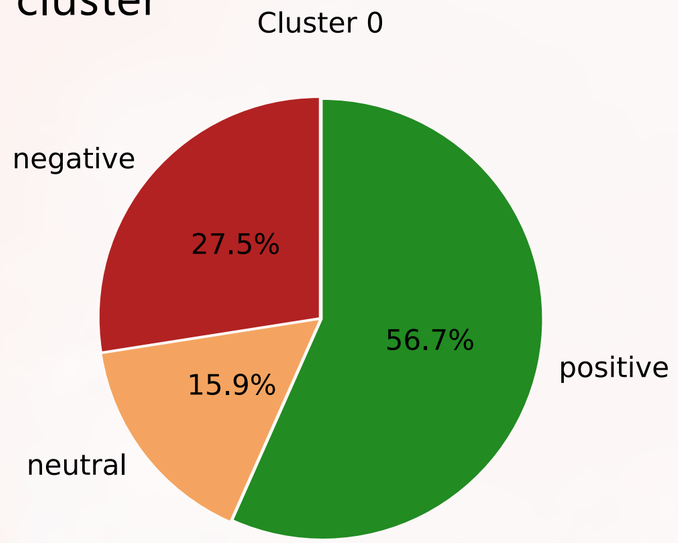


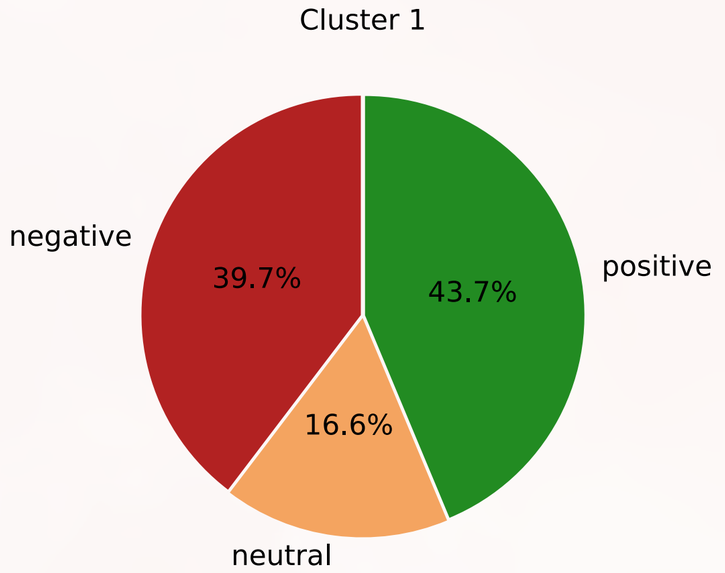


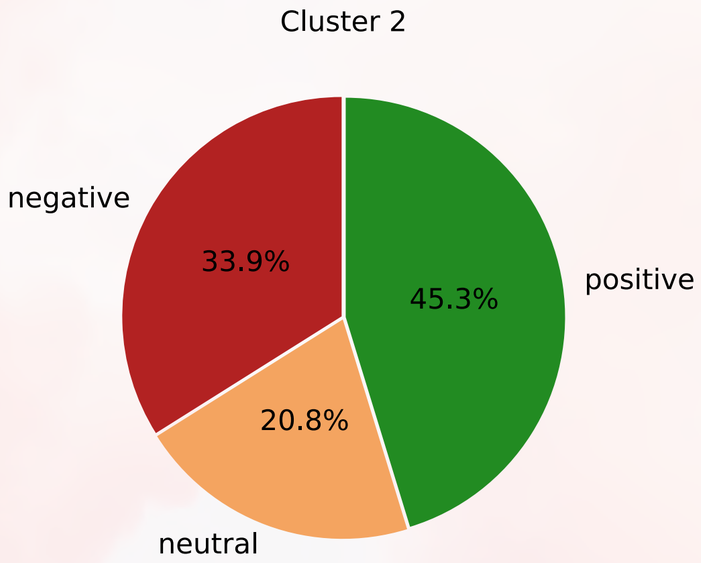
Opinion mining (VADER)

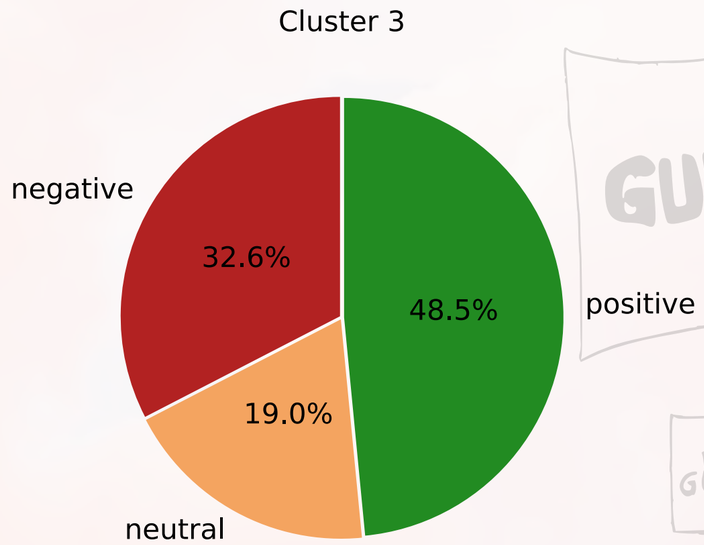


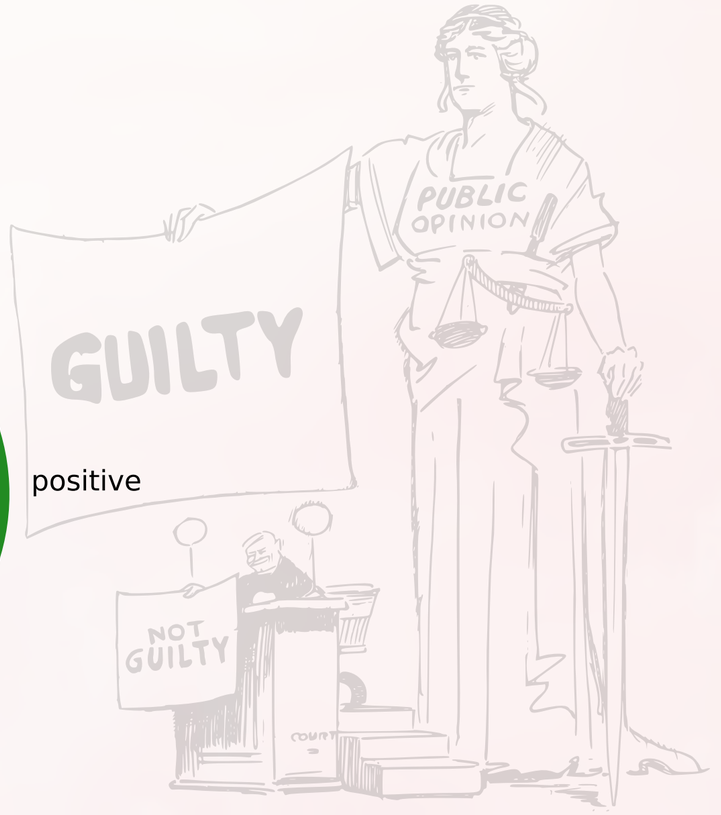
Overall sentiment per cluster





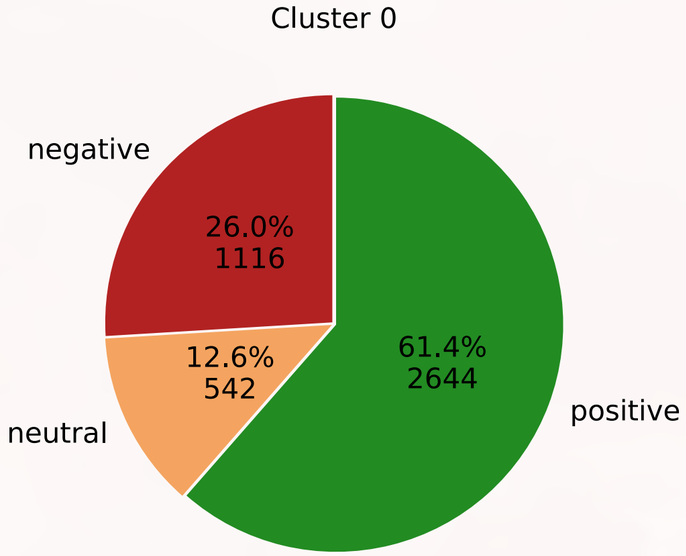


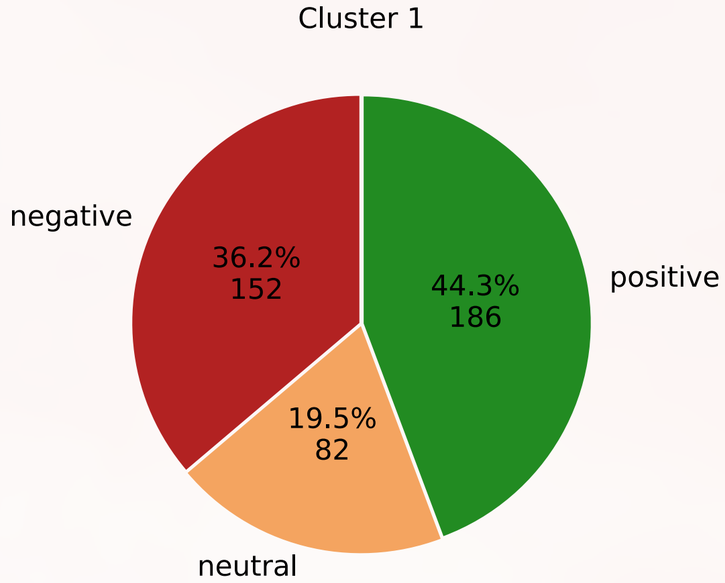


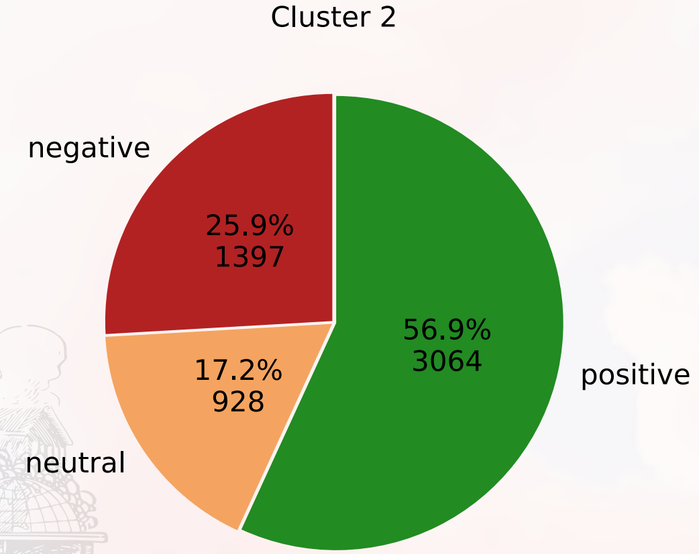


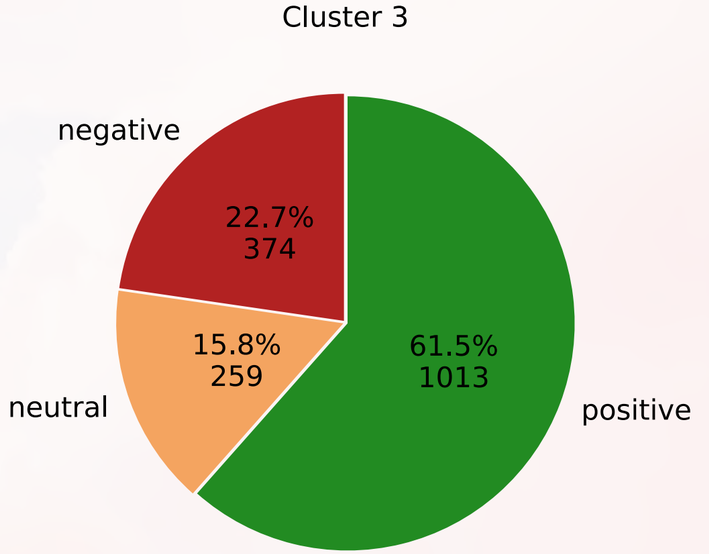
Education and Secularism

- "right"
- "secular"
- "school"
- "state"
- "public"
- "government"
- "member"
- "child"
- "free"
- "education"
- "group"
- "support"
- "student"
- "bill"
- "organization"



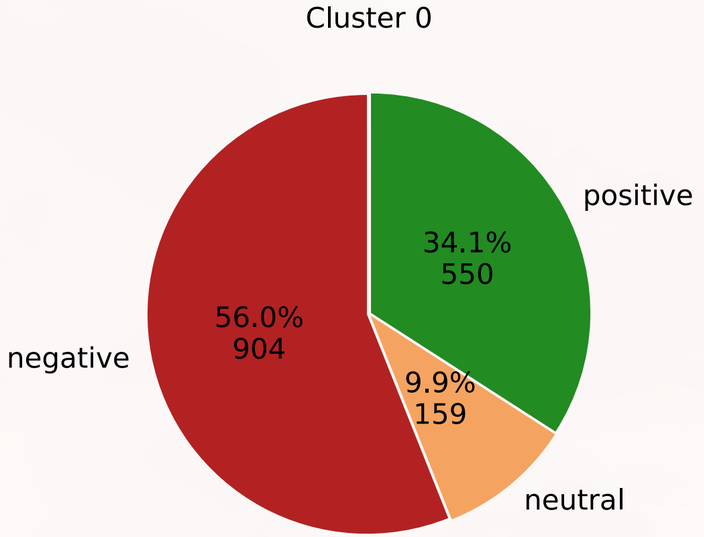


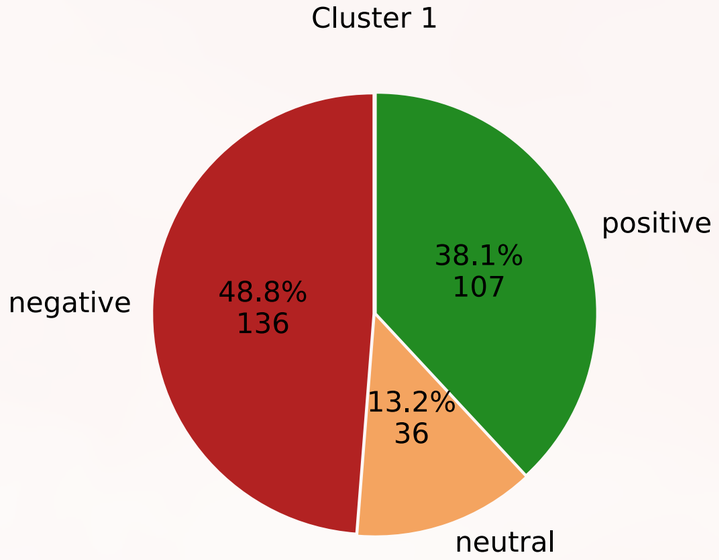


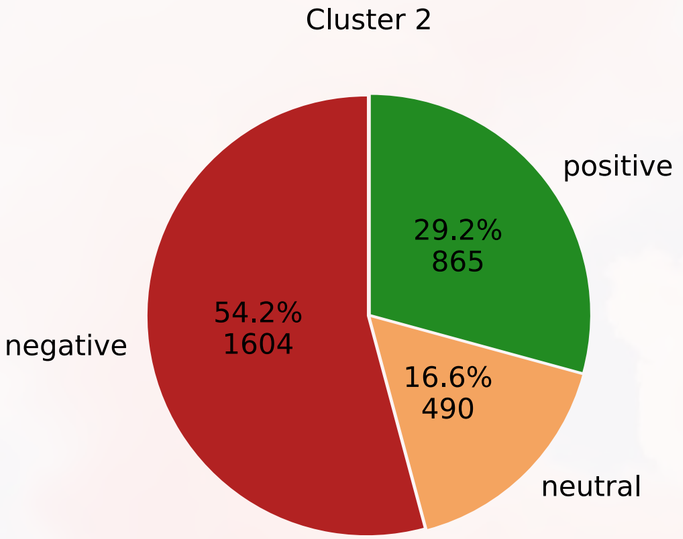


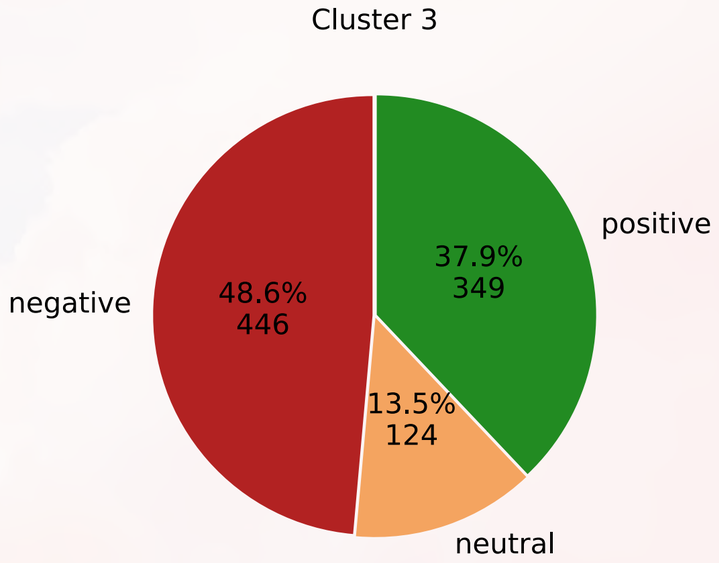
Covid-19 pandemic

- "covid"
- "vaccine"
- "report"
- "country"
- "church"
- "case"
- "death"
- "family"
- "pandemic"
- "leader"
- "blasphemy"
- "nation"
- "evangelical"
- "conspiracy_theorie"
- "coronavirus"



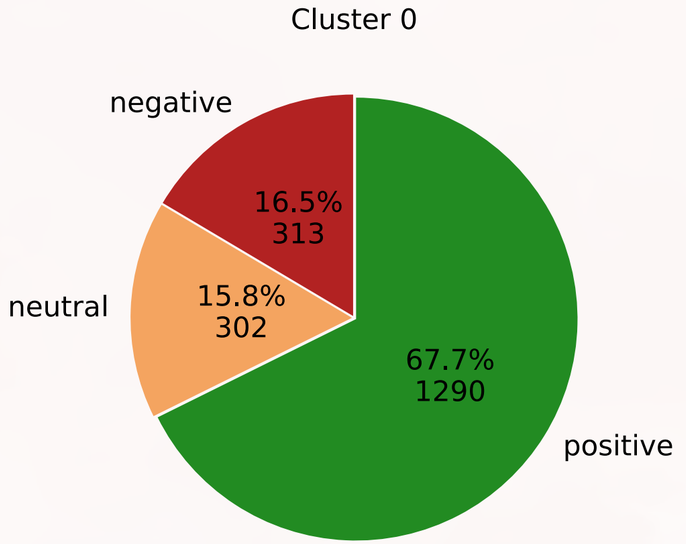


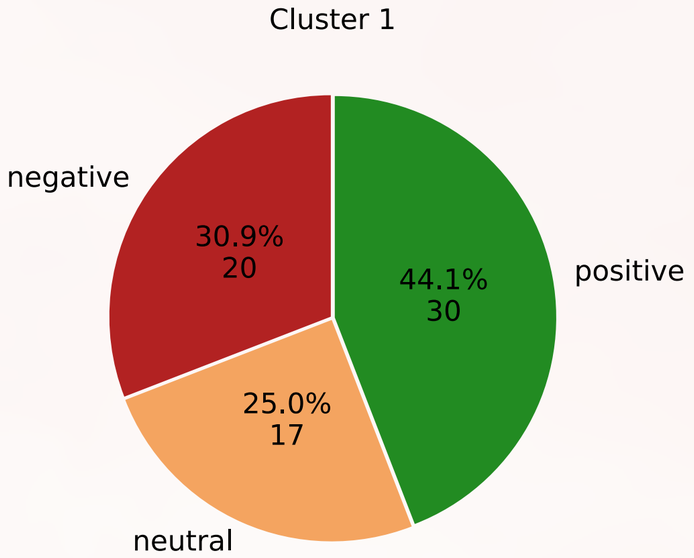


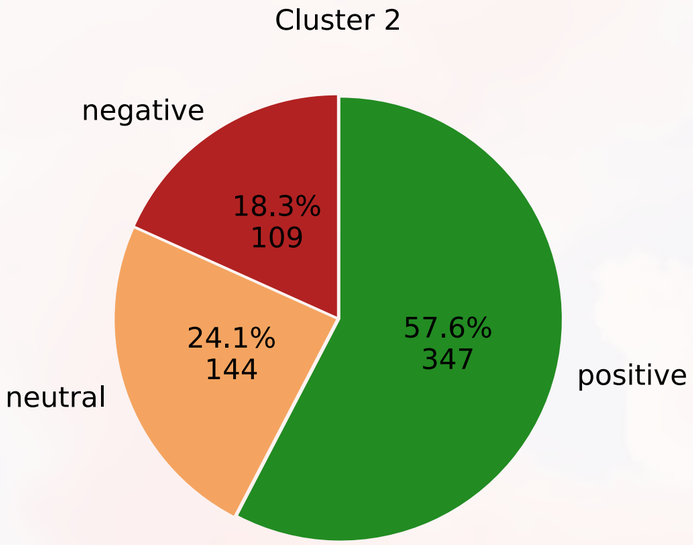


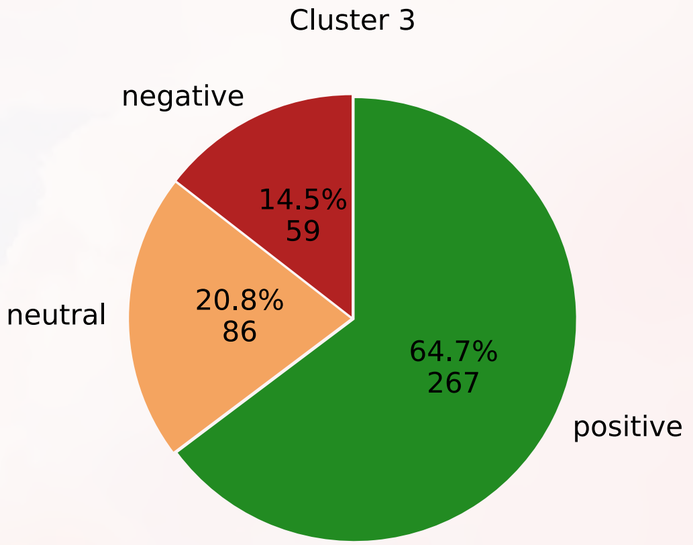
Ceremonies and celebrations

- "humanist"
- "wedding"
- "celebrant"
- "ceremony"
- "young"
- "humanism"
- "funeral"
- "movement"
- "couple"
- "marriage"
- "award"
- "music"
- "celebration"
- "memorial"
- "anniversary"



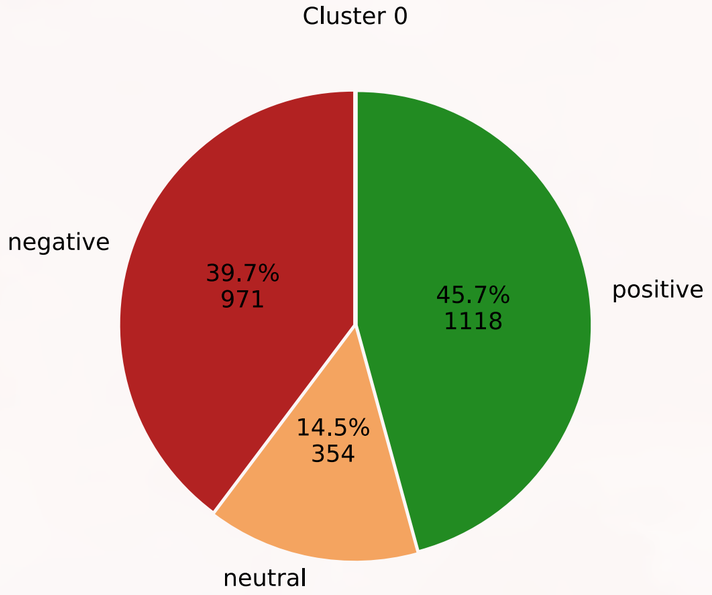


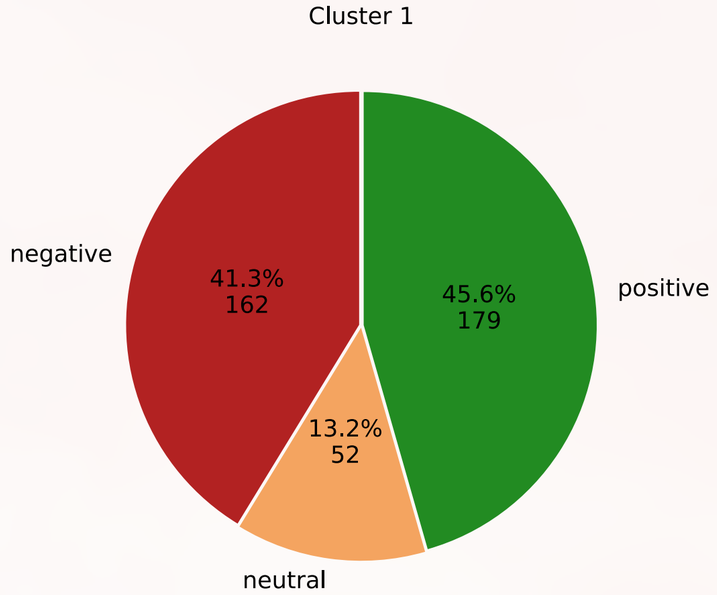


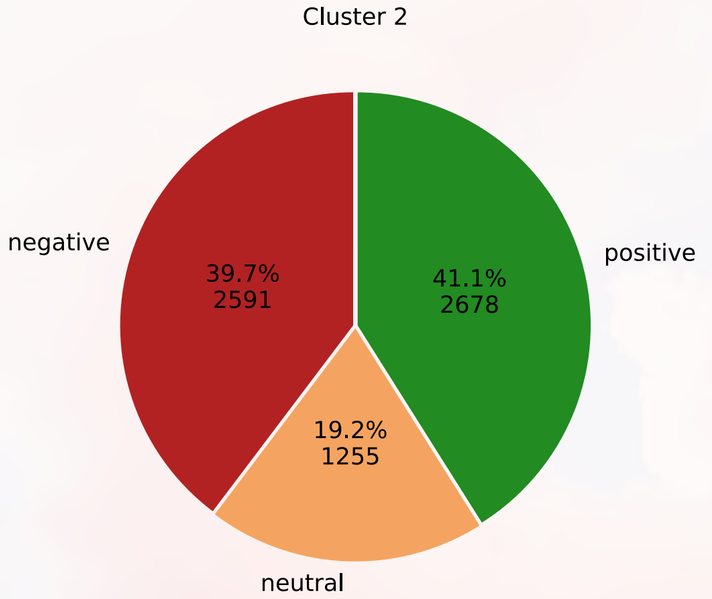


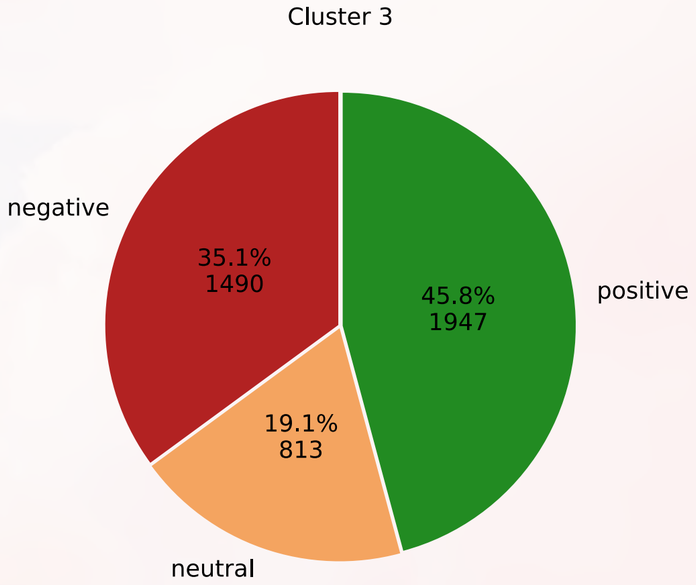
Question about religions

- christianity
- islam
- judaism
- jew
- muslim
- christian
- bible
- quran
- torah
- jesus
- yahweh
- allah



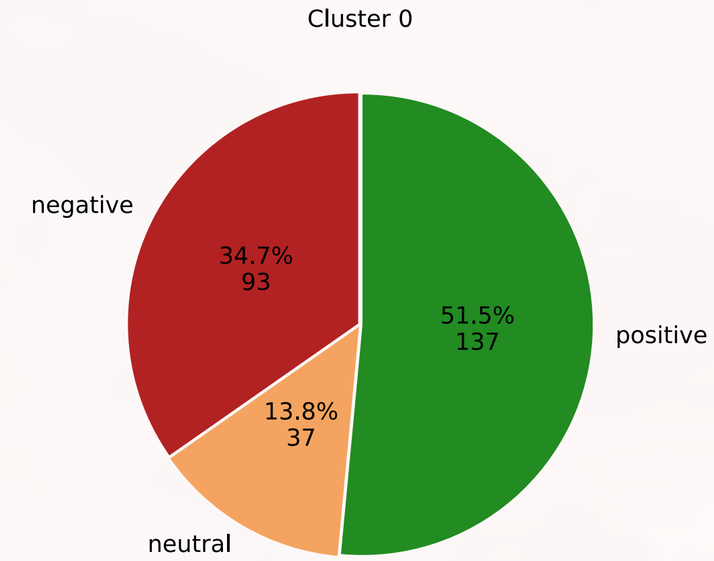


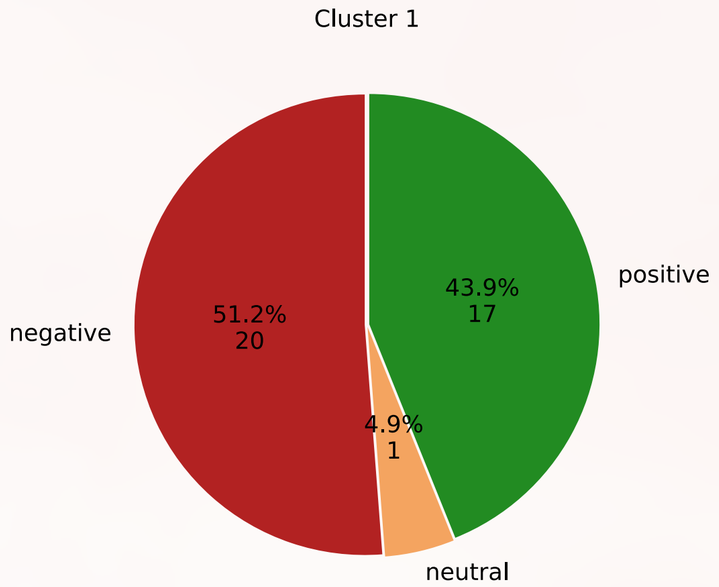


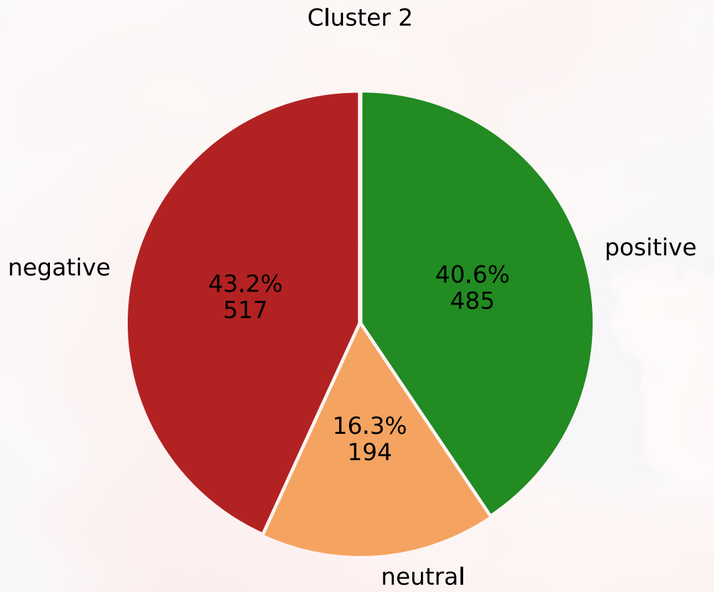


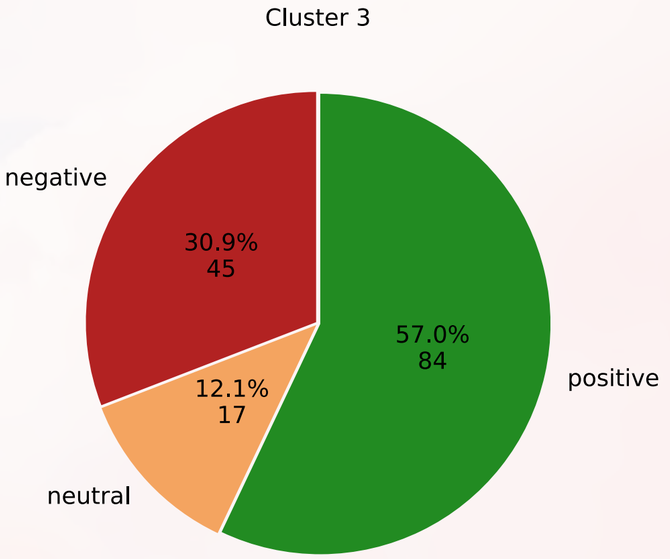
Question about vaccines

- vaccine
- vaccination
- immunization
- pfizer
- astrazeneca
- moderna
- booster



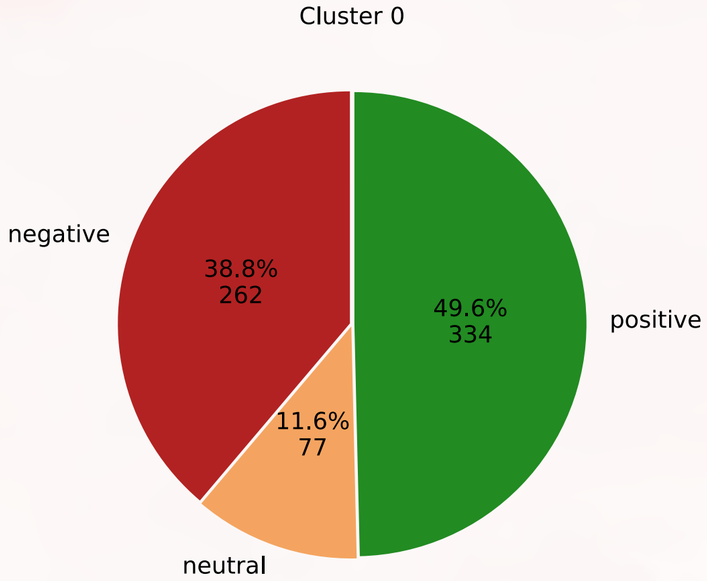


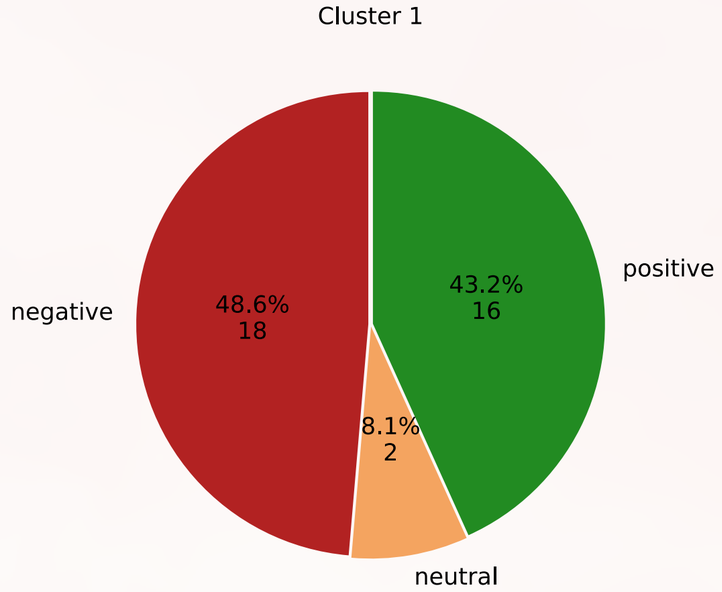


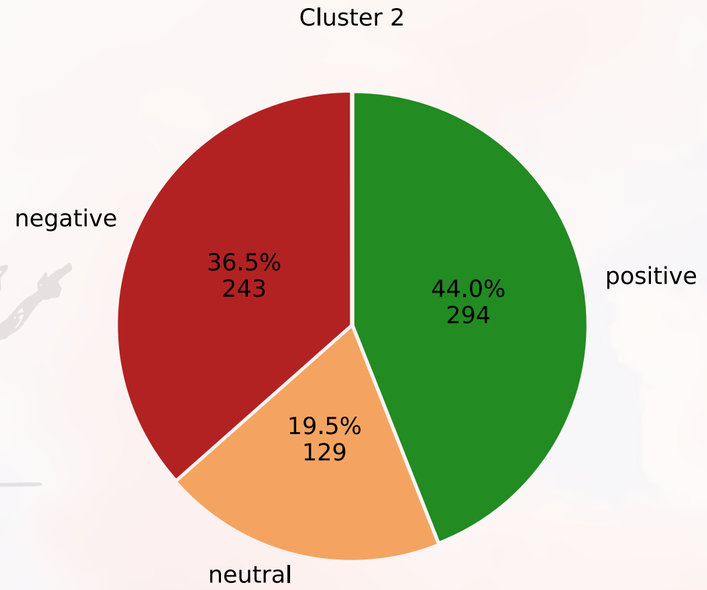


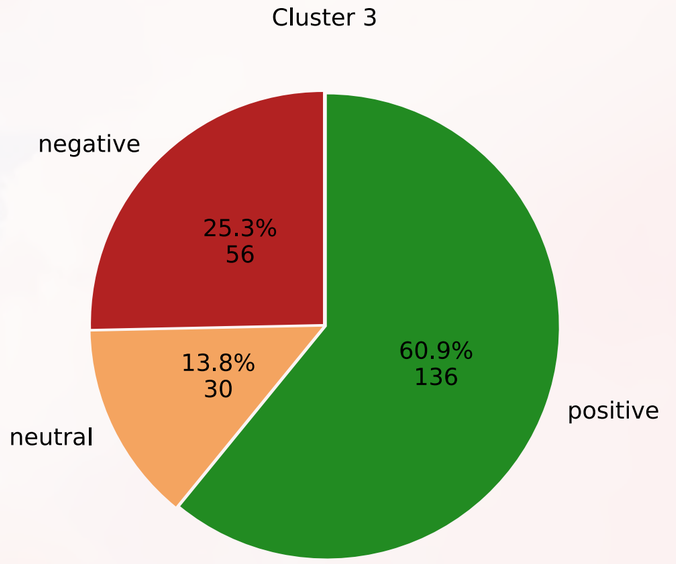
Question about abortus

- abortion
- abortus



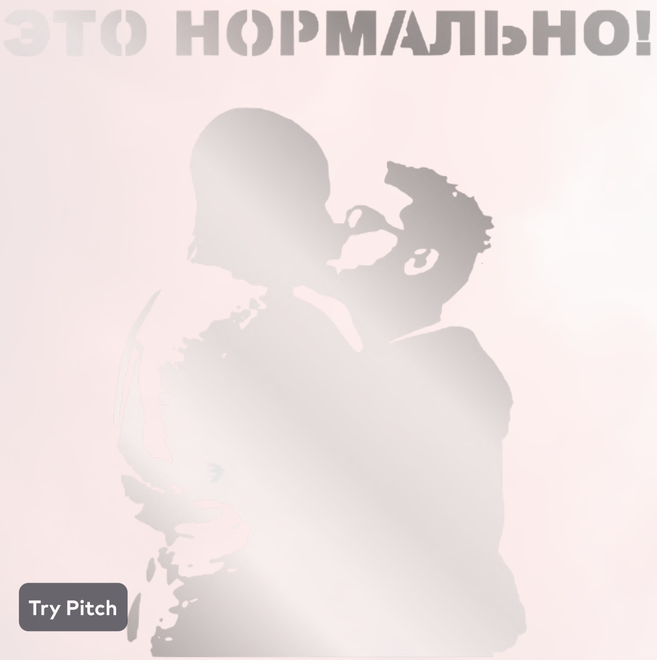


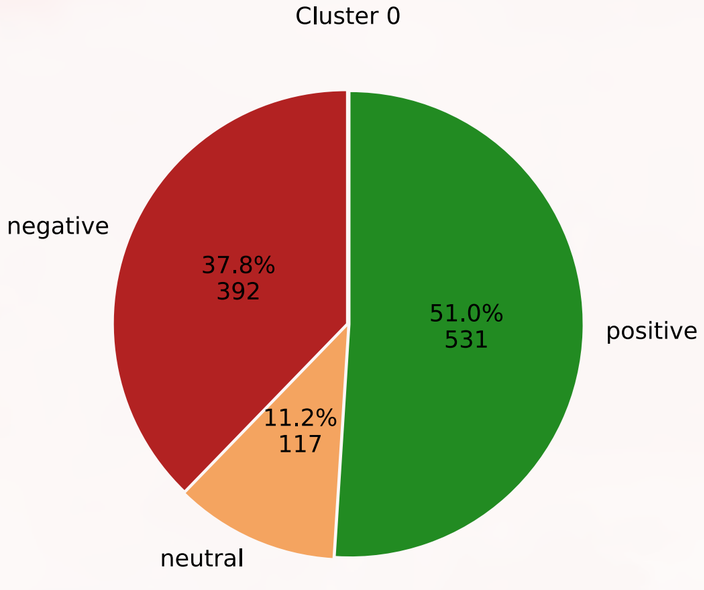


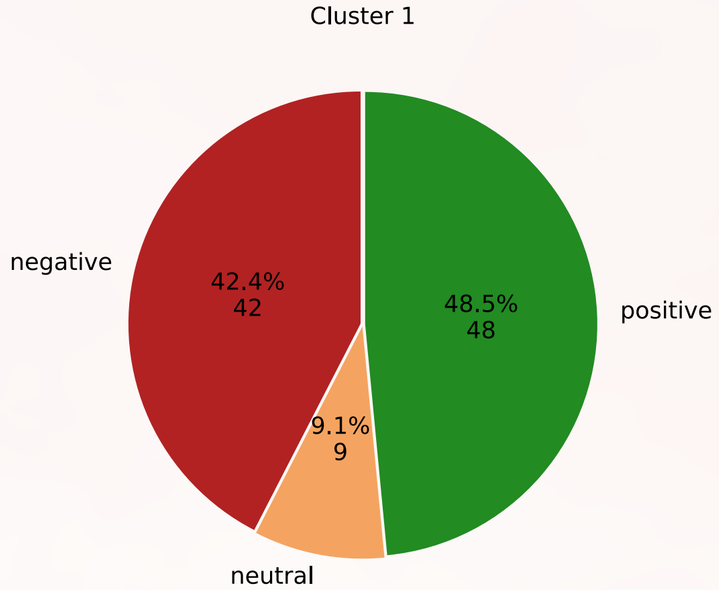


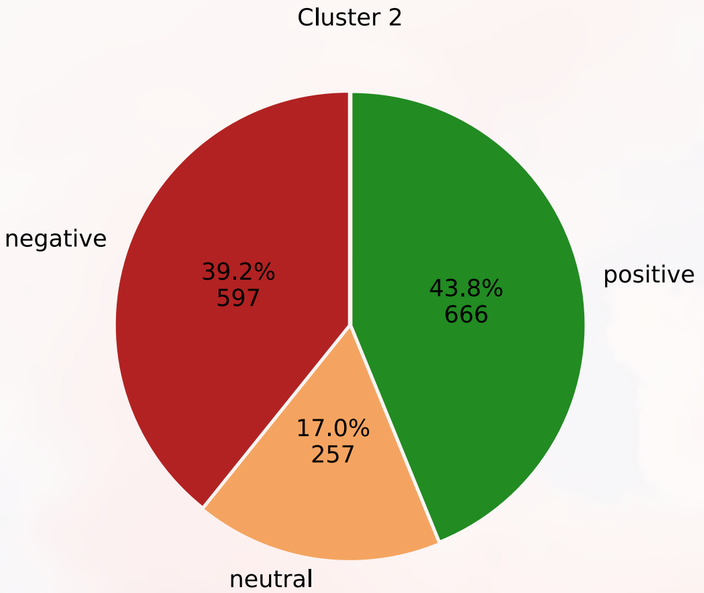
Question about LGBT

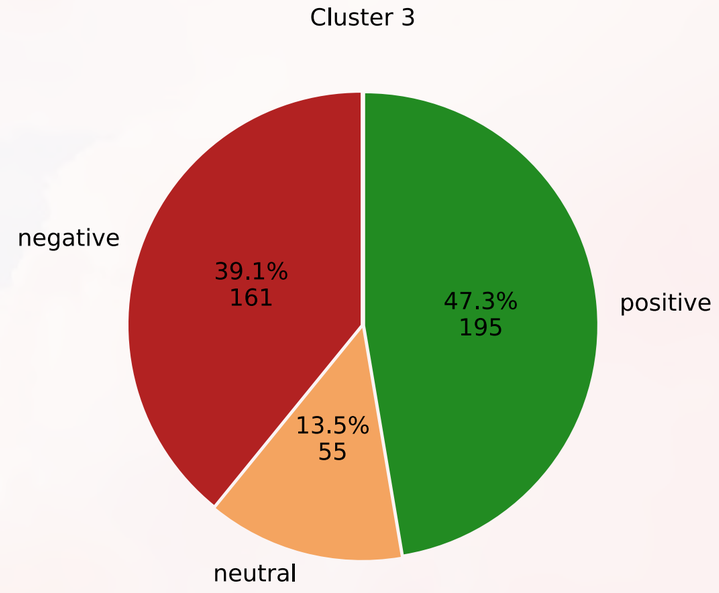
- gay
- lgbt
- lgbt+
- lesbian
- bisexual
- transgender
- same sex
- same-sex







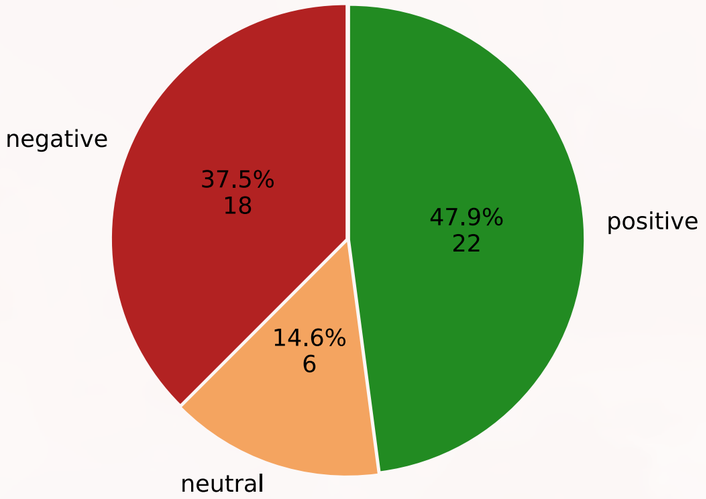




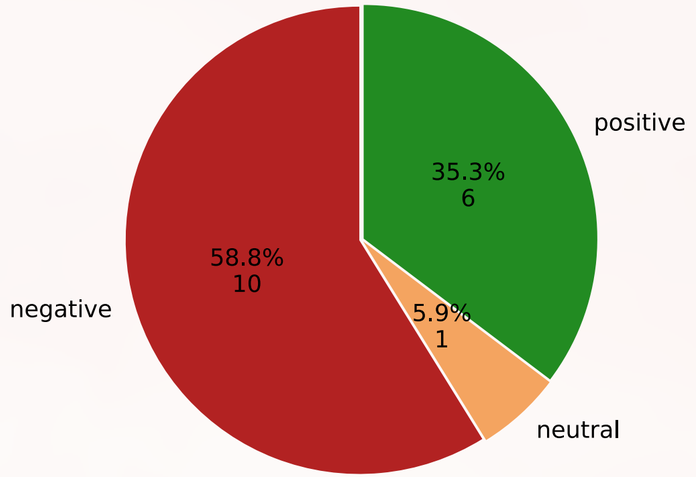
Question about aliens

- alien
- extraterrestrial
- ufo

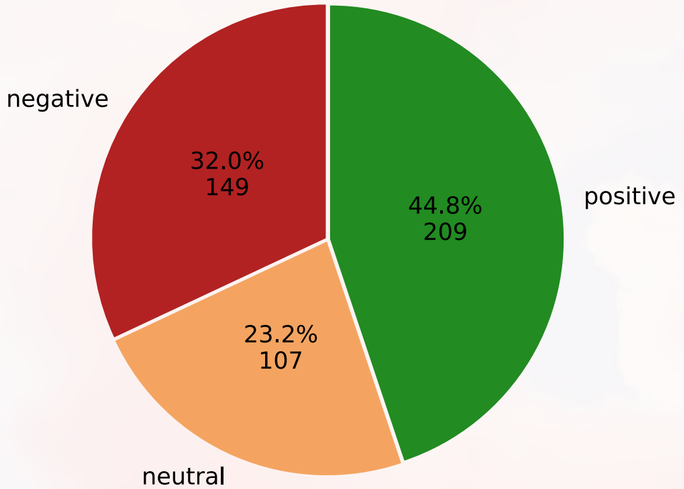
Cluster 0



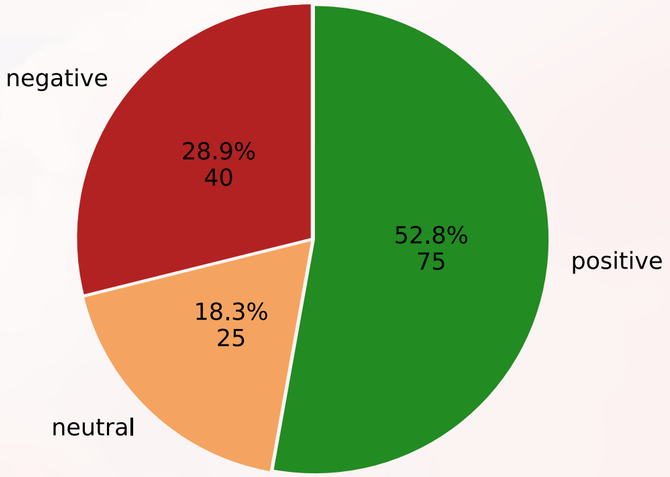
Cluster 1



Cluster 2

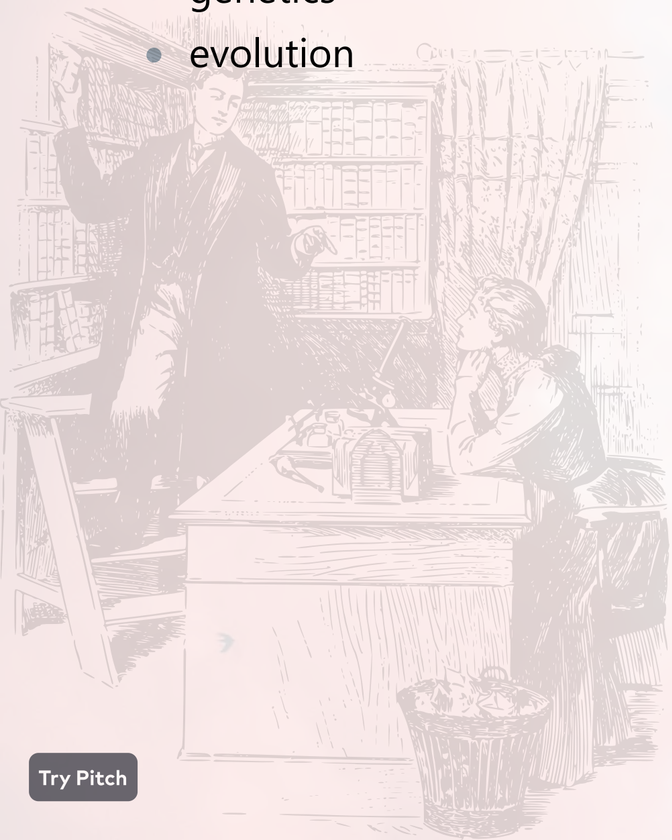


Cluster 3

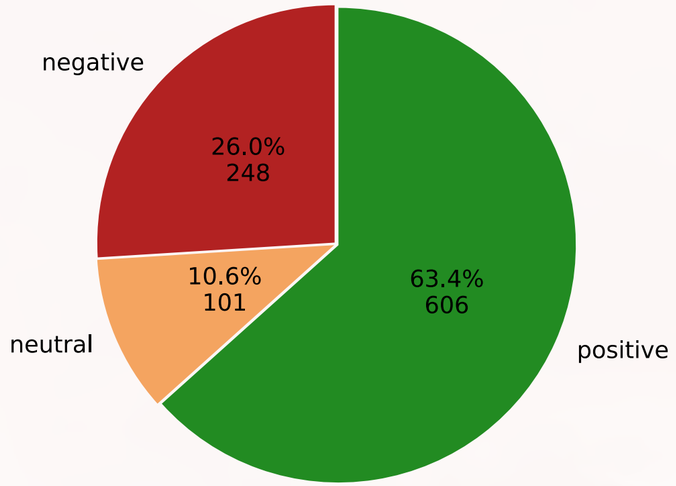


Question about science

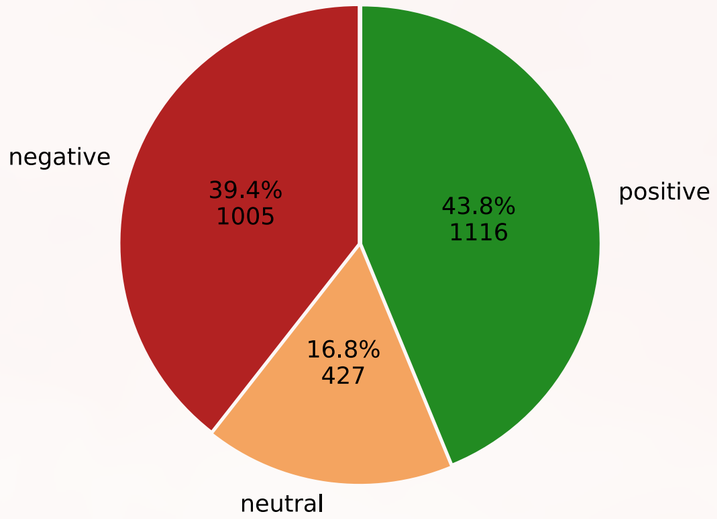
- science
- scientific
- scientific proof
- scientific evidence
- scientific method
- genetics
- evolution



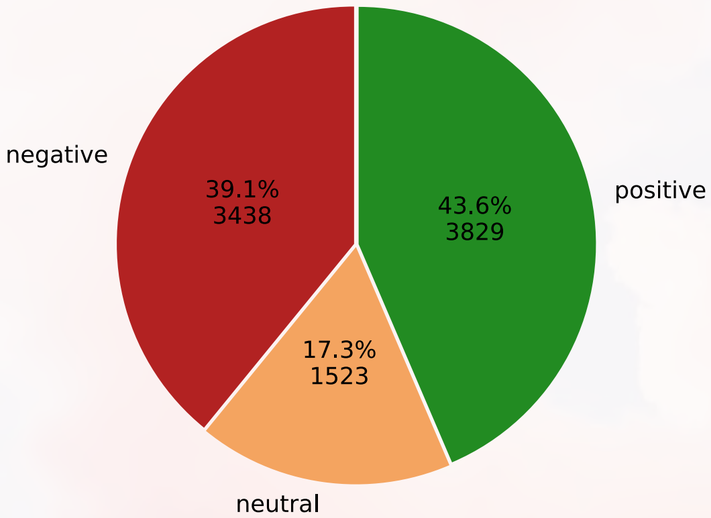
Cluster 0



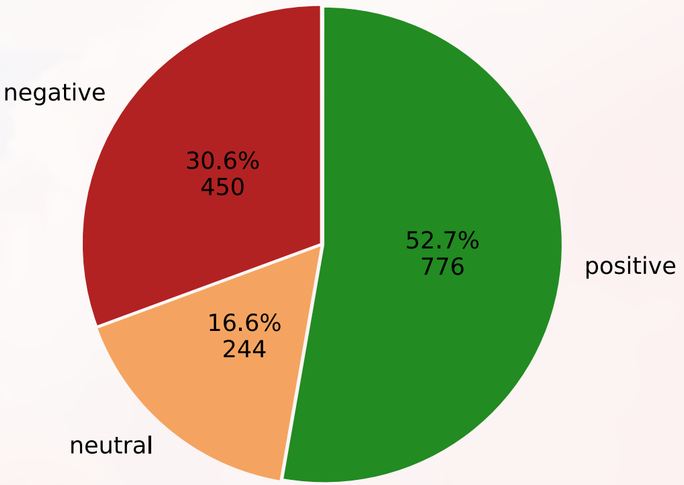
Cluster 1



Cluster 2



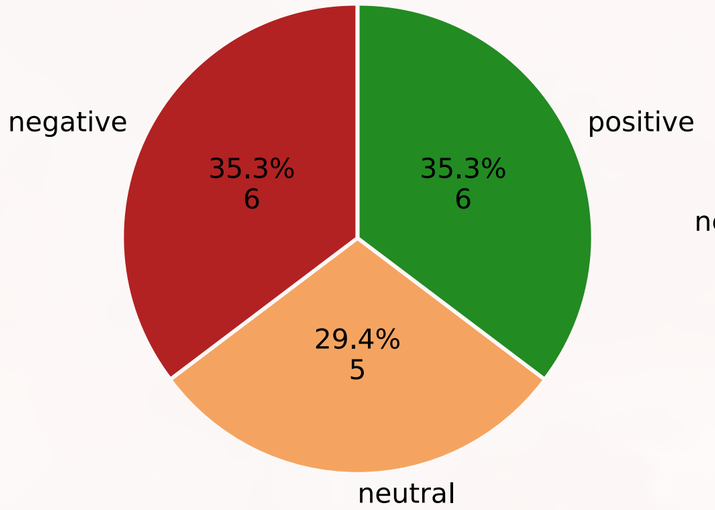
Cluster 3



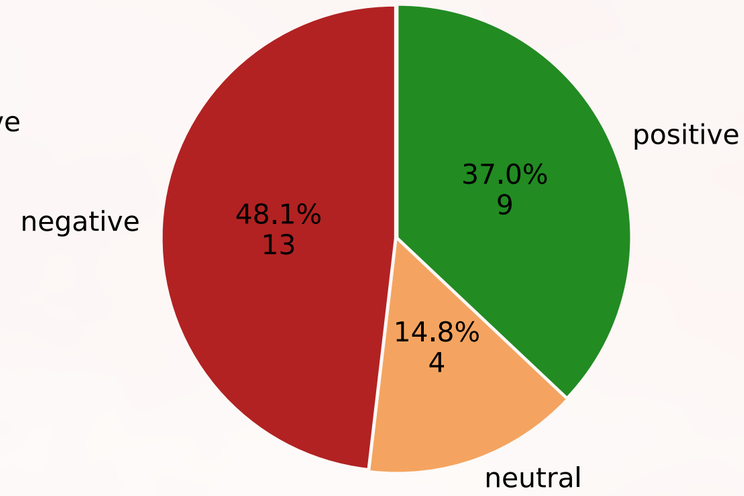
Question about woke culture

- woke
- woke culture
- wokeism

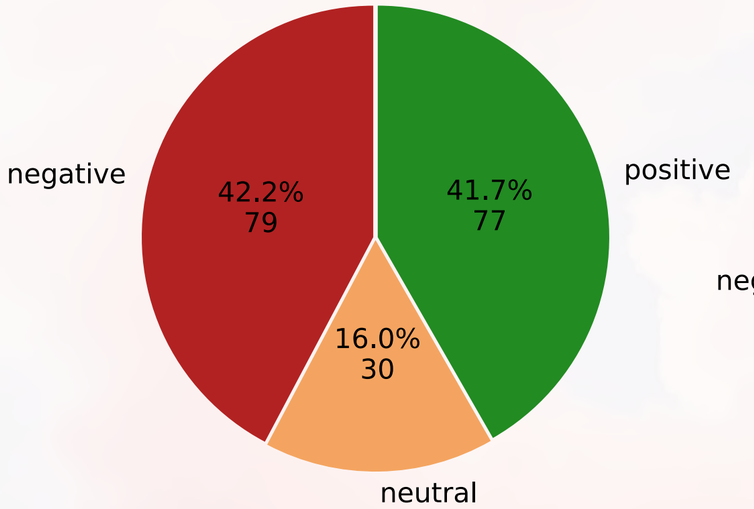
Cluster 0



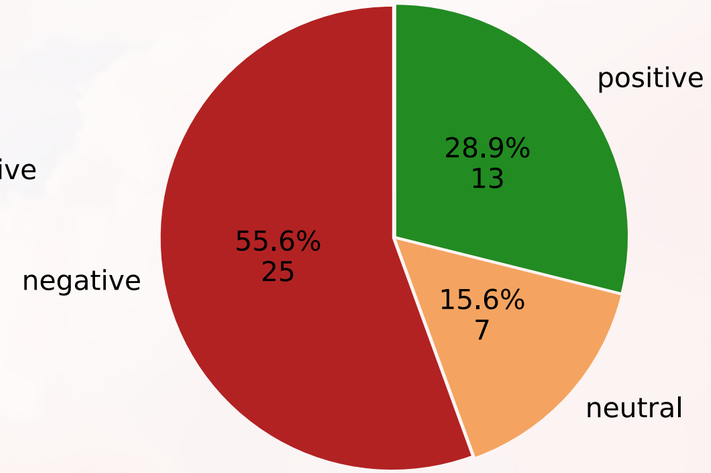
Cluster 1



Cluster 2



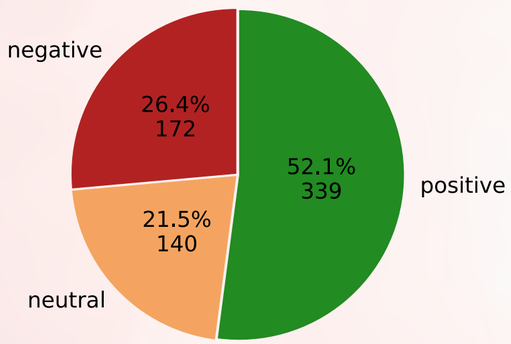
Cluster 3



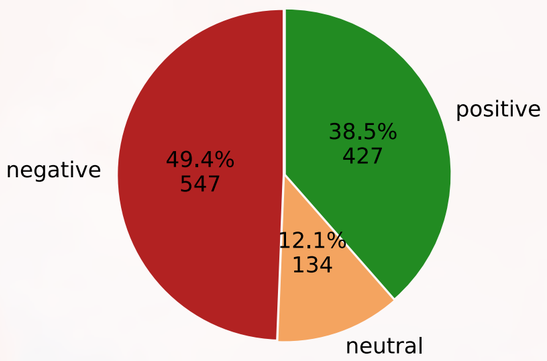
Question about politics

About Republicans (Trump, Republicans, Republican, Donald Trump)

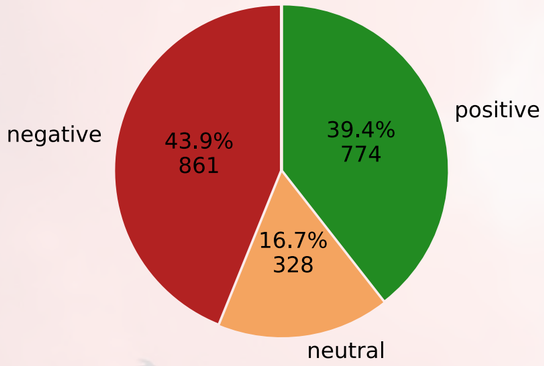
Cluster 0



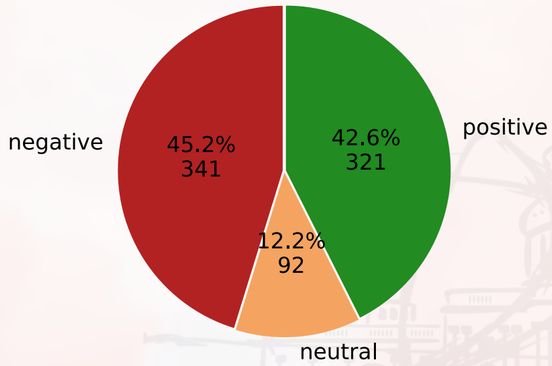
Cluster 1



Cluster 2

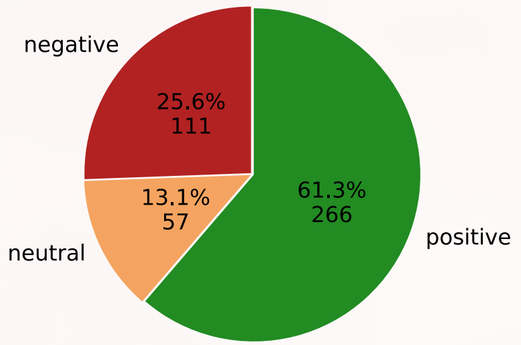


Cluster 3

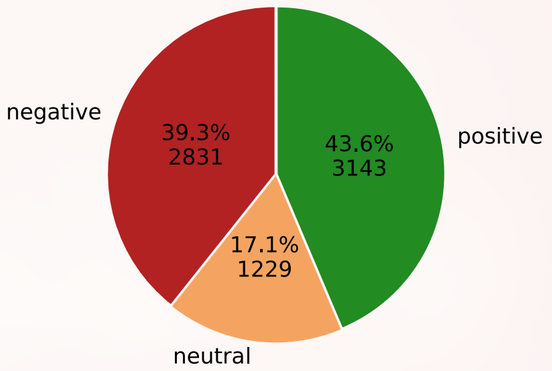


About Democrats (Joe Biden, Democrats, Biden, Democrat)

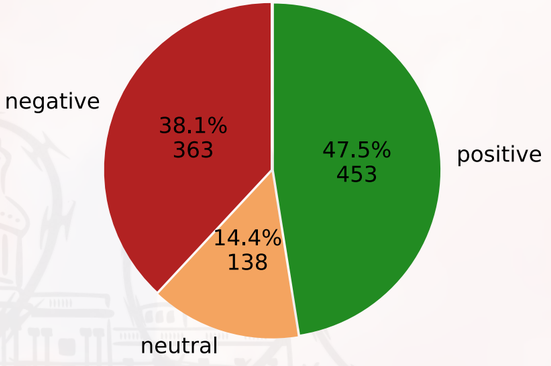
Cluster 0



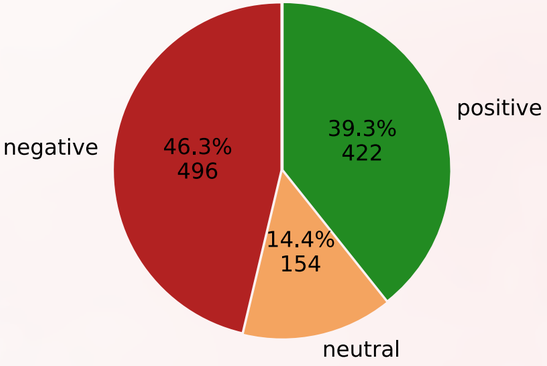
Cluster 1



Cluster 2



Cluster 3



Conclusion and future work

- Cluster 0 has a positive opinion of every custom topic.
- Cluster 1 usually does not have the significant number of instances, except the question is about politics.
- Cluster 2 is the most "conservative" cluster.
- Cluster 3 has the highest positive opinion about custom topics, except the political questions.

Conclusion and future work

- Cluster 0 has a positive opinion of every custom topic.
- Cluster 1 usually does not have the significant number of instances, except the question is about politics.
- Cluster 2 is the most "conservative" cluster.
- Cluster 3 has the highest positive opinion about custom topics, except the political questions.

- Other models then LDA → Non-Negative Matrix Factorization.
- Add verbs and adverbs in the analysis of topics.
- Tune VADER or use a different pre-trained model.