Data Science and Economics
Universit‘a degli Studi di Milano

In God we trust

Student: Vojimir Ranitovic 963780
Email: vojimir.ranitovic@studenti.unimi.it
Module: Information Retrieval

Introduction

Social media platforms such as Twitter have appear as common communication channels, generating large amounts of unstructured data. Mining meaningful insights from this data presents a considerable challenge, but natural language processing (NLP) techniques typically offer a solution. By analyzing communication patterns, attitudes, and opinions of Twitter users, NLP can help in informing and briefing marketing experts, public relations, and political campaigns etc. The primary goal of this project is to apply NLP techniques to analyze the communication styles of various non-religious Twitter accounts. Specifically, this research aims to identify frequently discussed topics on Twitter, cluster users based on their topic distribution, and analyze sentiment towards specific topics or words. By identifying groups of Twitter users with similar communication styles and examining their attitudes towards specific topics, this project aims to provide valuable insights into the communication styles of Twitter users.

Research question and methodology

By using NLP techniques, the project aims to analyze large amounts of Twitter data to identify both, more or less common topics, the sentiment of these topics, and the communication styles of non-religious users, by clustering similar ones. By doing this, the project aims to provide insights into the interests of accounts, and how they communicate their opinions through Twitter. Finally, the research question is to address what are the common topics, communication styles, and opinions expressed by non-religious Twitter users, and how can natural language processing techniques be used to identify and analyze these characteristics in their tweets.

Data collection and preprocessing

The dataset is scraped from the Twitter using snscrape[1] library. Tweets of all 278 non-religious Twitter accounts were scraped from 2007-01-01 until 2023-01-30. Snscrape did not have any daily restrictions on downloaded tweets which is the case with the official Twitter API. There are 1,224,446 tweets in total. For the sake of faster implementation and knowing more recent events, only the tweets after 2020-01-01 were used, and there are 195,053 tweets after that date.

For the preprocessing phase, standard procedures were used. Stopwords, links, mentions, hashtags, punctuations, and numbers were removed. Also, short words with 3 or fewer letters were deleted and the remaining words were tokenized. After using Gensim's 'ENGLISH_CONNECTOR_WORDS', the most common bigrams were also used in the analysis. Between stemming and lemmatization, the latter is chosen because even though

[1] GitHub repository: https://github.com/JustAnotherArchivist/snscrape

it is slower, it gave more context to topic modeling. During lemmatizing process, only nouns and adjectives are chosen because they provided more information for my specific task, but also other parts of speeches could be used. Below is an example of one tweet before and after preprocessing.

Original:

b"@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'."

Processed and lemmatized:

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

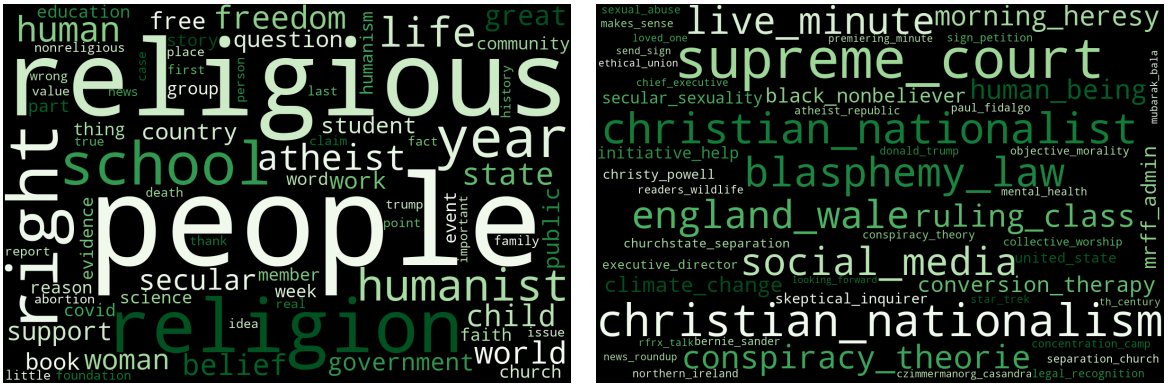

Figure 1: The most common unigrams and bigrams in tweets.

In Figure 1. it can be seen the most common unigrams and bigrams. Those common words give us a better understanding about what could be the possible topics even before using the topic modeling algorithm.

Topic modeling

Gensim's LDA (Latent Dirichlet Allocation) is an unsupervised learning algorithm for analyzing text data and finding topics. It assumes that each document (in our case tweet) is a mixture of a fixed number of topics, and each topic is a probability distribution over the words in the corpus. Many parameters can be tuned in LDA, but two key parameters are alpha and eta. Alpha controls how many topics each document consists of and higher alpha values indicate more diverse topic distributions and lower values indicate more focused topic distributions. The optimal value for alpha depends on the nature

of the corpus. Tweets are short, they are naturally limited to 280 characters, so lower values for alpha would suit better in this case (fewer topics per tweet). Eta controls the sparsity of the topic-word distributions, with higher values, there are more diverse word distributions, and lower values indicate more focused distributions. The optimal value for eta will also depend on the characteristics of the vocabulary, and in the case of tweets, we focused on lower numbers for eta (fewer words per topic).
To evaluate the quality of the topics generated by the LDA, a coherence measure is used, more specifically c_v from Gensim. It 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.

Clustering the Twitter accounts and sentiment analysis
After finding the topics, all the accounts were clustered according to their topic distributions, using the k-means algorithm. To better understand the sentiments and opinions of each cluster VADER (Valence Aware Dictionary and sEntiment Reasoner) was used. VADER uses a lexicon of words and phrases that have been labeled with a sentiment score, and it also takes into account the grammatical rules and structure of the text. This allows it to accurately determine the sentiment of a piece of text, even if the sentiment is expressed in a subtle or complex way. It produces a polarity score ranging from -1 to +1, with -1 indicating extremely negative sentiment and +1 indicating extremely positive sentiment. In our case interval from -0.05 to 0.05 is set to be neutral.

Results

There are a lot of different parameters in Gensims LDA, like chunk size, alpha, eta, passes, number of topics, decay, offset, iterations, gamma_threshold, etc. It would be very difficult to tune all of them, so the most influential ones in the case of my datasets were alpha, eta, and the number of topics. To choose the best model, a random search of these three parameters was conducted. Alpha was in the range (0.001, 1, 0.001), eta in the range (0.001, 0.5, 0.001), and the number of topics in the range (20, 65, 5), where the last number in those tuples shows the steps. Among 300 random iterations, the best model was one with 60 topics, alpha was 0.429 and eta was 0.464. The best coherence score was 0.479.

Topic 30	Topic 31	Topic 32	Topic 33	Topic 34	Topic 35
person(0.041)	virus(0.025)	covid(0.046)	news(0.069)	black(0.037)	health(0.042)
kind(0.027)	city(0.016)	vaccine(0.035)	good(0.027)	kid(0.030)	faith(0.041)
mean(0.017)	homeopathy(0.011)	report(0.034)	discussion(0.023)	conservative(0.030)	point(0.034)
comment(0.013)	likely(0.010)	country(0.029)	possible(0.019)	white(0.029)	course(0.028)
let(0.012)	deadly(0.007)	church(0.028)	amazing(0.017)	speaker(0.022)	need(0.026)
move(0.010)	wish(0.007)	case(0.026)	argument(0.017)	racism(0.009)	care(0.025)
conscience(0.007)	initiative(0.007)	death(0.025)	topic(0.013)	hole(0.008)	well(0.024)
headline(0.007)	friendly(0.005)	family(0.019)	chance(0.012)	audience(0.008)	system(0.021)
sound(0.007)	terrorist(0.004)	pandemic(0.018)	nonsense(0.012)	similar(0.007)	view(0.019)
memory(0.007)	quackery(0.004)	leader(0.013)	believer(0.011)	guy(0.007)	future(0.017)
hypocrite(0.007)	liar(0.004)	blasphemy(0.011)	trust(0.010)	ideology(0.006)	call(0.017)

Figure 2: The part of the LDA's output.

In Figure 2. It can be seen that for example Topic 32 is about Covid19 pandemic, Topic 31 is about the virus and some alternative ways to cure it, and Topic 34 is about racial politics. Out of 60 topics, 54 were named and the rest were unclear about what they were about. Those confusing topics were not that significant and there is just a low amount of tweets that talked about them.

After finding those topics, they are assigned back to each tweet in the dataset, and for each user was made a vector of topic distribution. Then it was filtered so that only topics that are talked about more than 3% per user remained in vectors. Those vectors are then used for clustering, and using K-means and elbow technique there were discovered 4 clusters seen in the Figure 3.

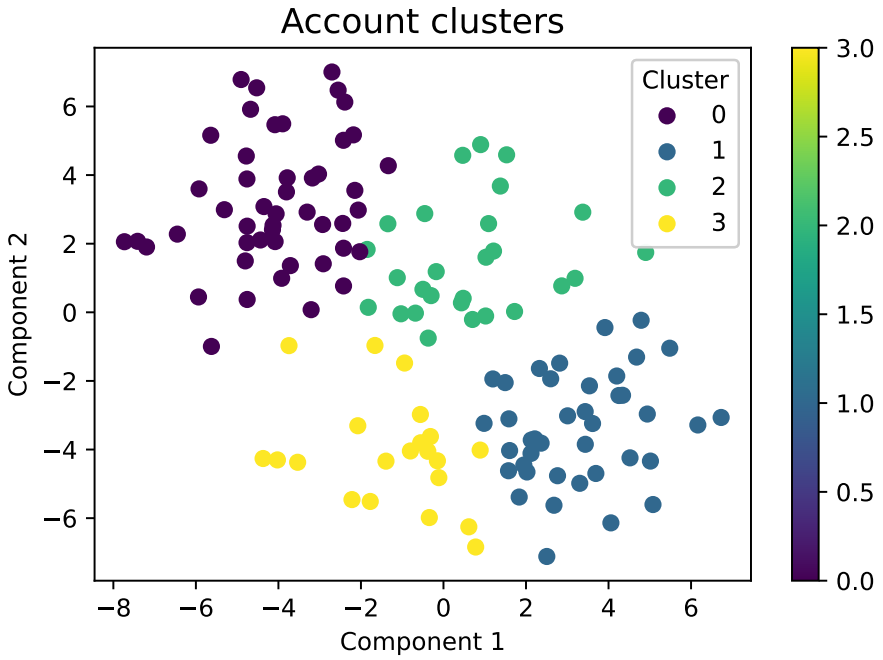

Figure 3: K-means users clustering results.

In Figure 4. we can see topic distribution per each of the 4 clusters. It presents the top 15 topics per cluster, where topics marked with an asterisk are exclusive for the specific cluster (but only in the range of the top 15 topics). So, we can see that those clusters have some topics that are the same, and very important for non-religious people, but those topics are ranged differently. As it can be seen, cluster 0 is more about education and secularism in schools but also about ceremonies like weddings, completely separated from tradition. Cluster 1 values more skeptical narratives, religion, and beliefs. Cluster 2 is all around cluster without an exclusive topic, and cluster 3 is more about science and understanding the world, but also asking the question about the truth and evidences of believers.

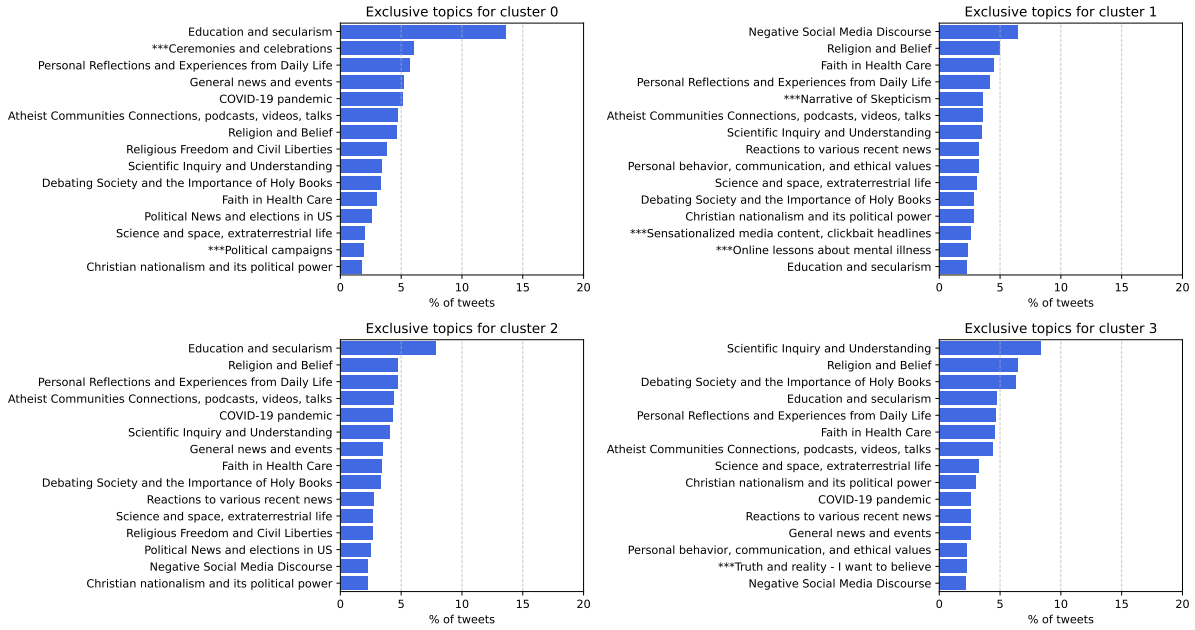

Figure 4: Top 15 topics per cluster.

In the appendix in Figure 7. is shown that the number of tweets per account was not balanced. Accounts like TheCOAPodcast, AtheistsRead and Humanists_Uk are the most influential, and easily can affect opinion mining analysis. Also, an appendix in Figure 8. is shown the top 25 topics overall, and as it can be seen religious nones talk mostly about secularism, beliefs, and science, but also there are everyday topics and general news and Covid-19 pandemic that is one of the major topics in recent years in almost every aspect of life.

After getting all the topics and the clusters, VADER shows that overall sentiments in all tweets are overly positive in Figure 5. The similar situation is also in each cluster as we can see in Figure 9. in appendix. Opinions can be checked also if we investigate cluster and different topics or just keywords of interests. In appendix is shown that all clusters agree that they opinion is overly positive for the topic Ceremonies and celebrations, also overly positive about whole topic talking about Education and secularism, but also there is agreement that COVID-19 pandemic is highly negative topic for each cluster. All this can be seen in Figure 10. until Figure 12. in the appendix.

To extract opinions on specific topics of interest, such as politics, vaccines, or the LGBT community, custom keywords were used, like in Figure 6, and also in the Figure 13. until Figure 19. in the appendix.

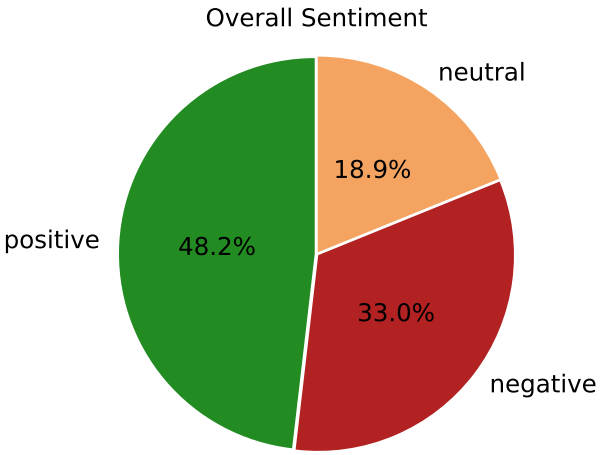

Figure 5: Overall sentiment of all tweets.

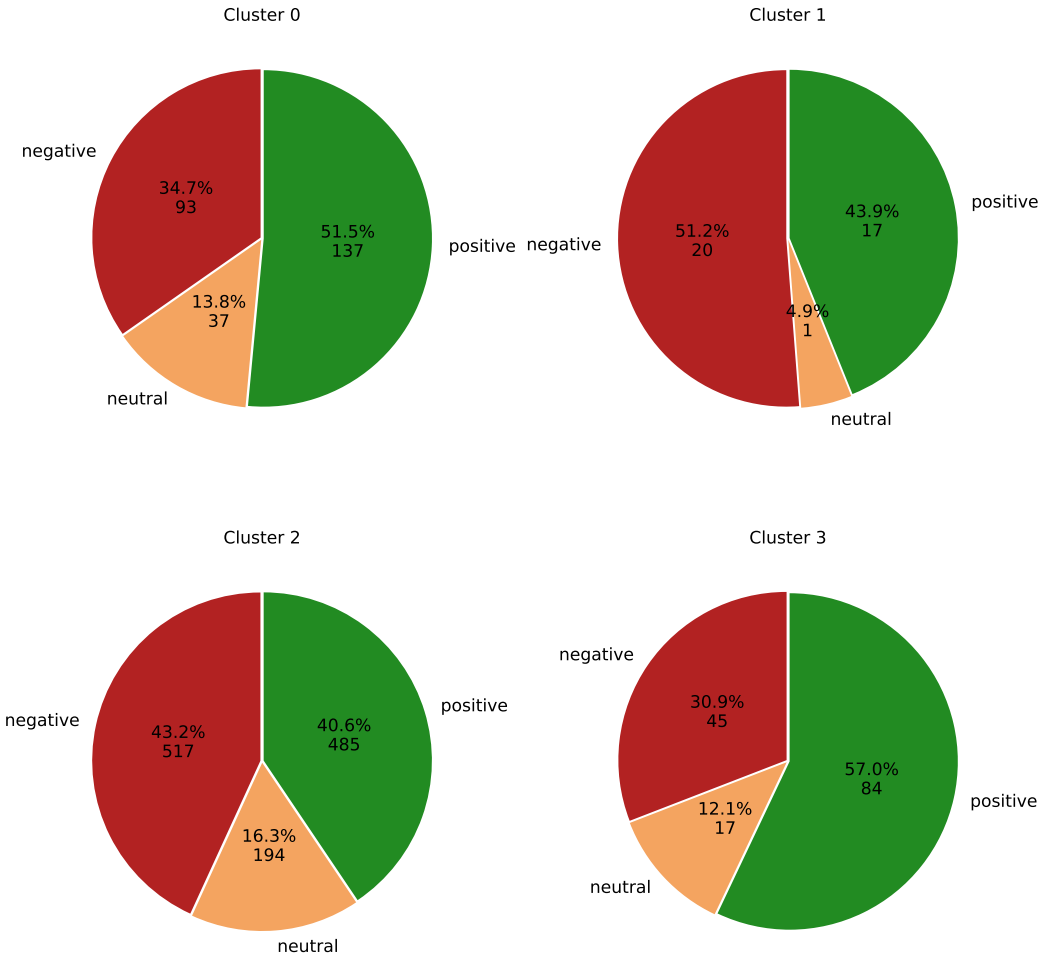

Figure 6: Opionion about **"Vaccines"** per each cluster. Keywords: "vaccine","vaccination", "immunization","pfizer","astrazeneca","moderna","booster".

From this analysis which keywords ask about vaccines, politics, religion, aliens, abortus, and LGBT it can not be seen that there are so many differences (and extreme opinions) in those 4 clusters, there are just slight ones. For example, cluster 0 has a positive opinion for every custom topic that was analyzed. Cluster 1 usually does not have a lot of instances to be considered for opinion, except when it was asked about republicans and democrats, so it is cluster more related to politics. Cluster 2 is uncanny, it has just slightly positive opinions about LGBT and religion, it is only cluster with a negative opinion about vaccines, a negative one about republicans, and a positive opinion about democrats. Cluster 3 is the cluster that is totally against both mentioned politics, and usually, it has the highest percentage of positive opinion about other custom topics.

Conclusion and future work

In this project, it could be seen that LDA could be a powerful tool for topic modeling, but it highly depends on the preprocessing stages and it also demands a lot of human-made decisions. LDA has a lot of parameters that can be tuned, and hyperparameter tuning could be computationally extensive and slow. In the future, instead of LDA, there could be used other, simpler algorithms like non-negative matrix factorization (NNMF), it could be faster and perhaps better for short tweets-like texts. Also, future improvement could be tweaking of VADER model, or using another model for opinion mining and comparing it to VADER's results.

Appendix

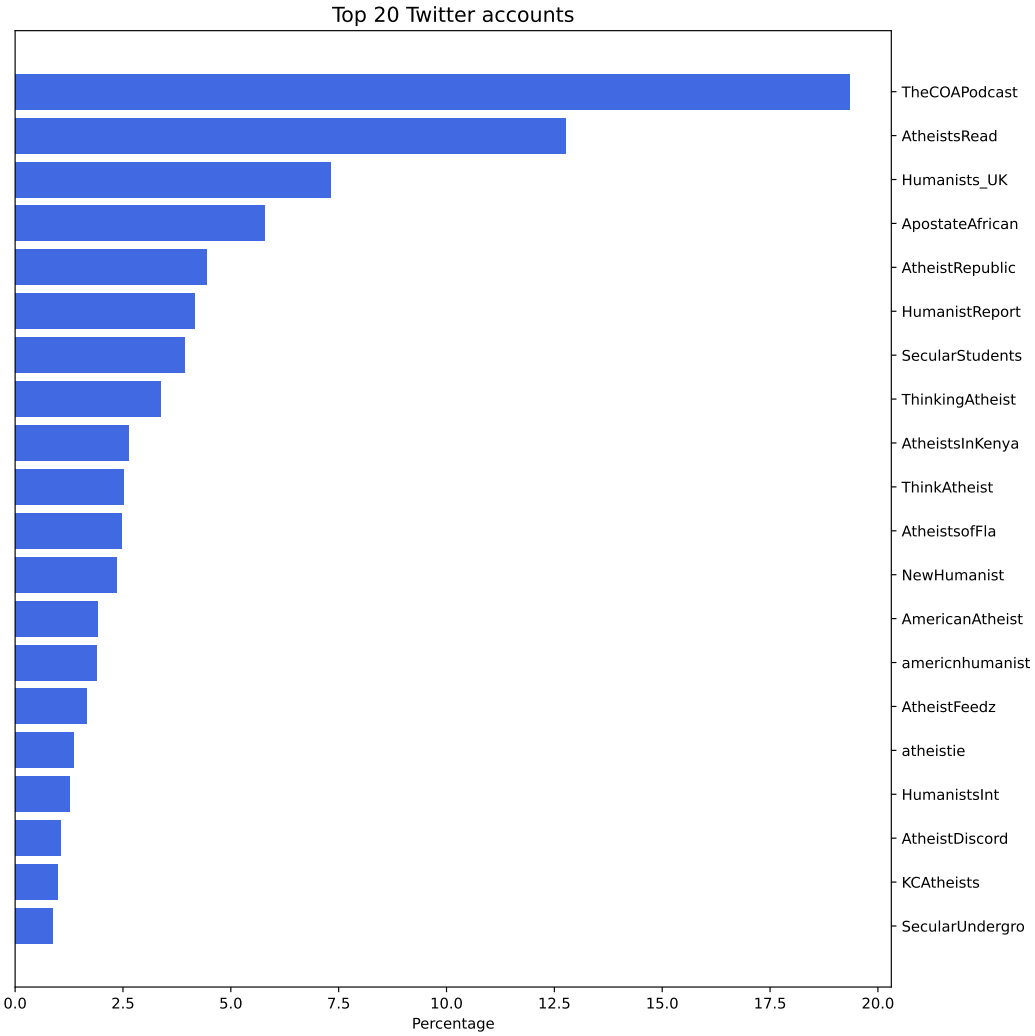

Figure 7: Top 20 Twitter accounts.

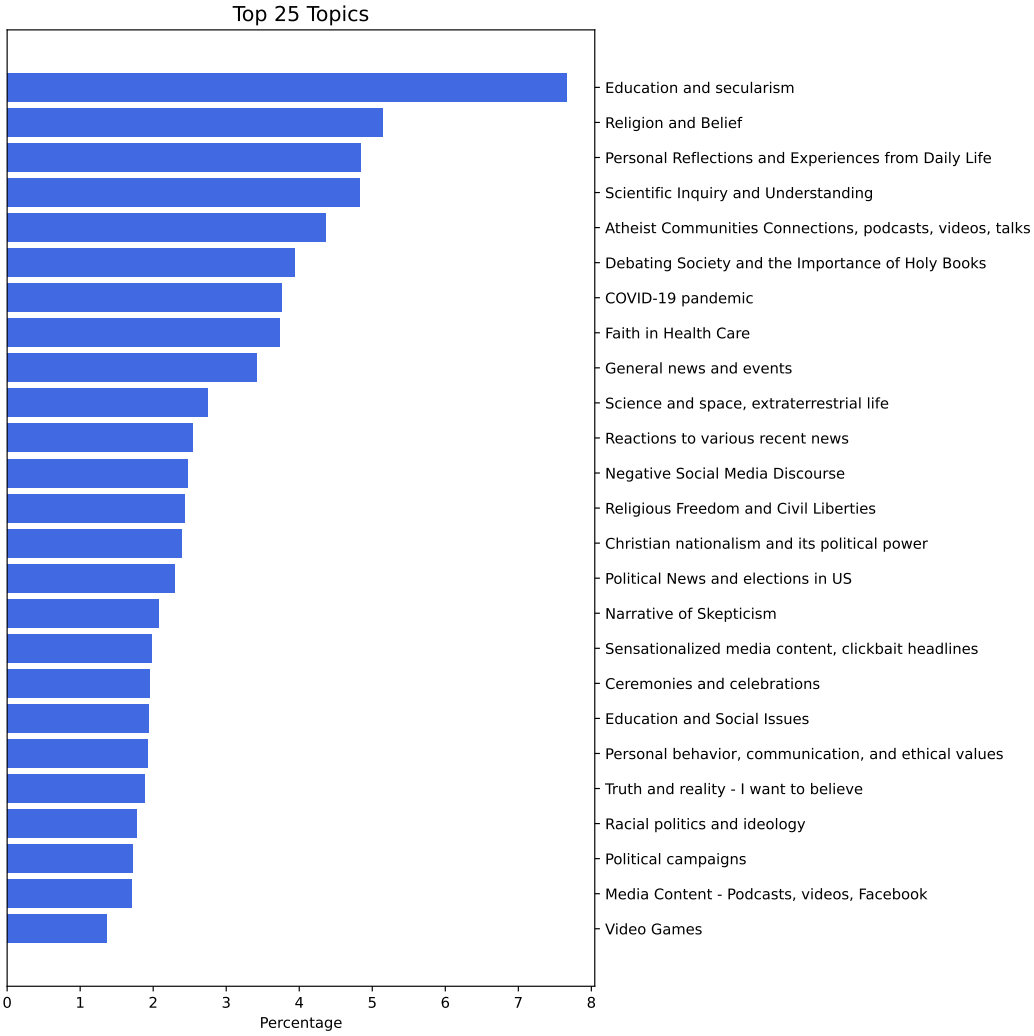

Figure 8: Top 25 topics among religious nones.

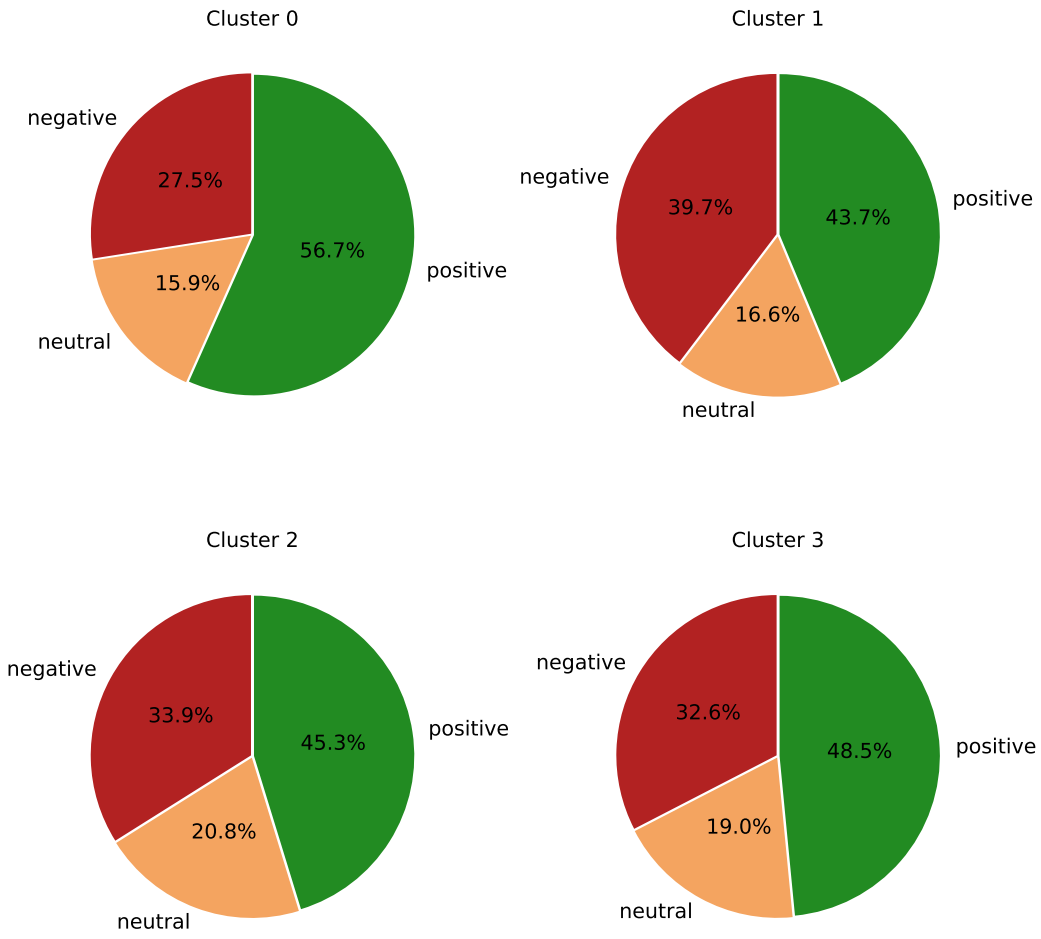

Figure 9: Overall sentiment per each cluster.

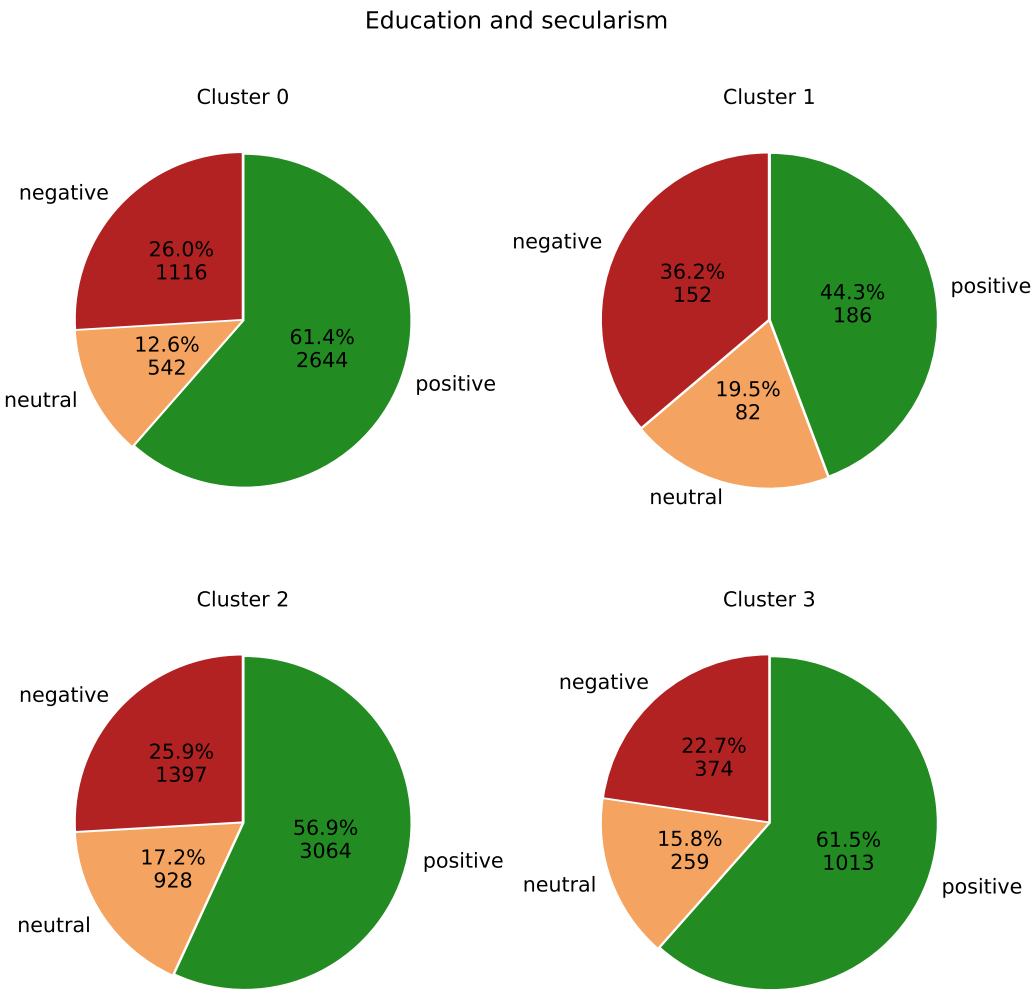

Figure 10: Opinion about "Education and secularism" per each cluster.

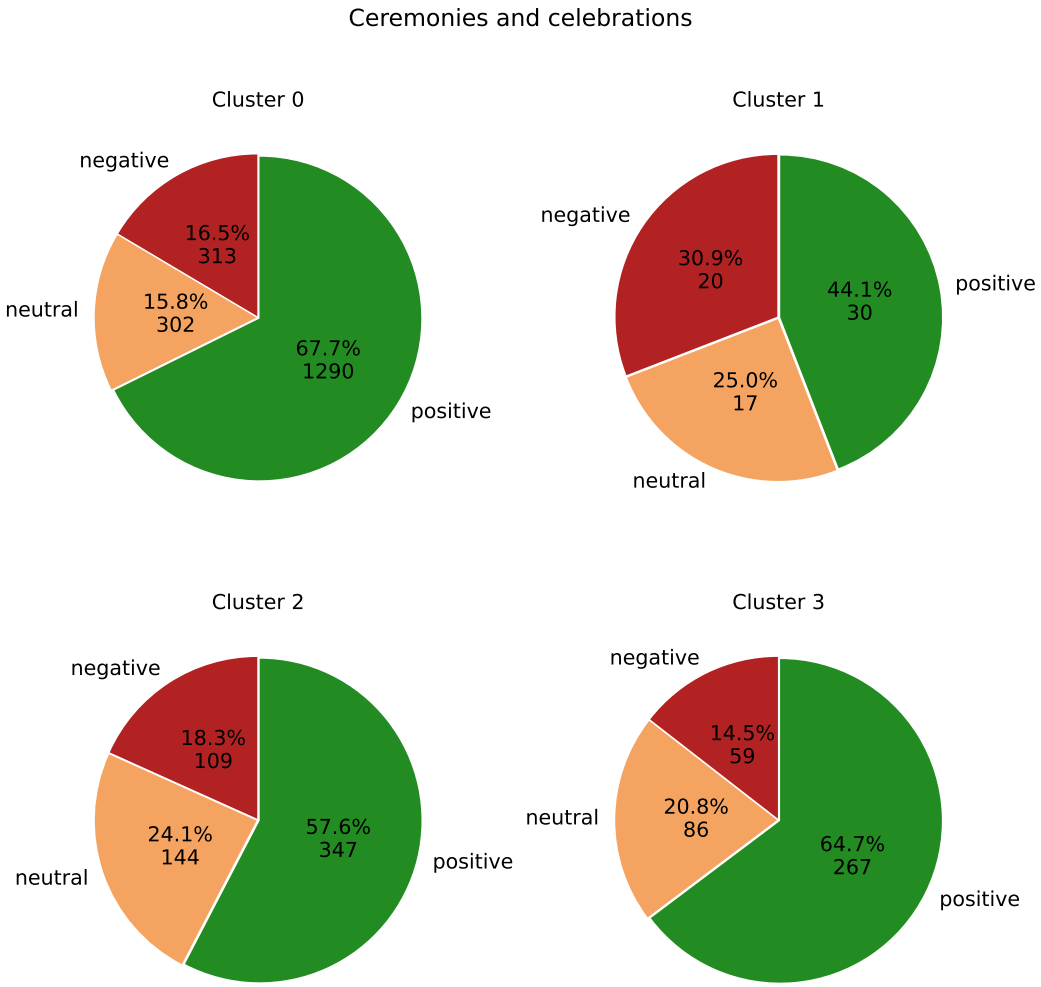

Figure 11: Opinion about "Ceremonies and celebrations" per each cluster.

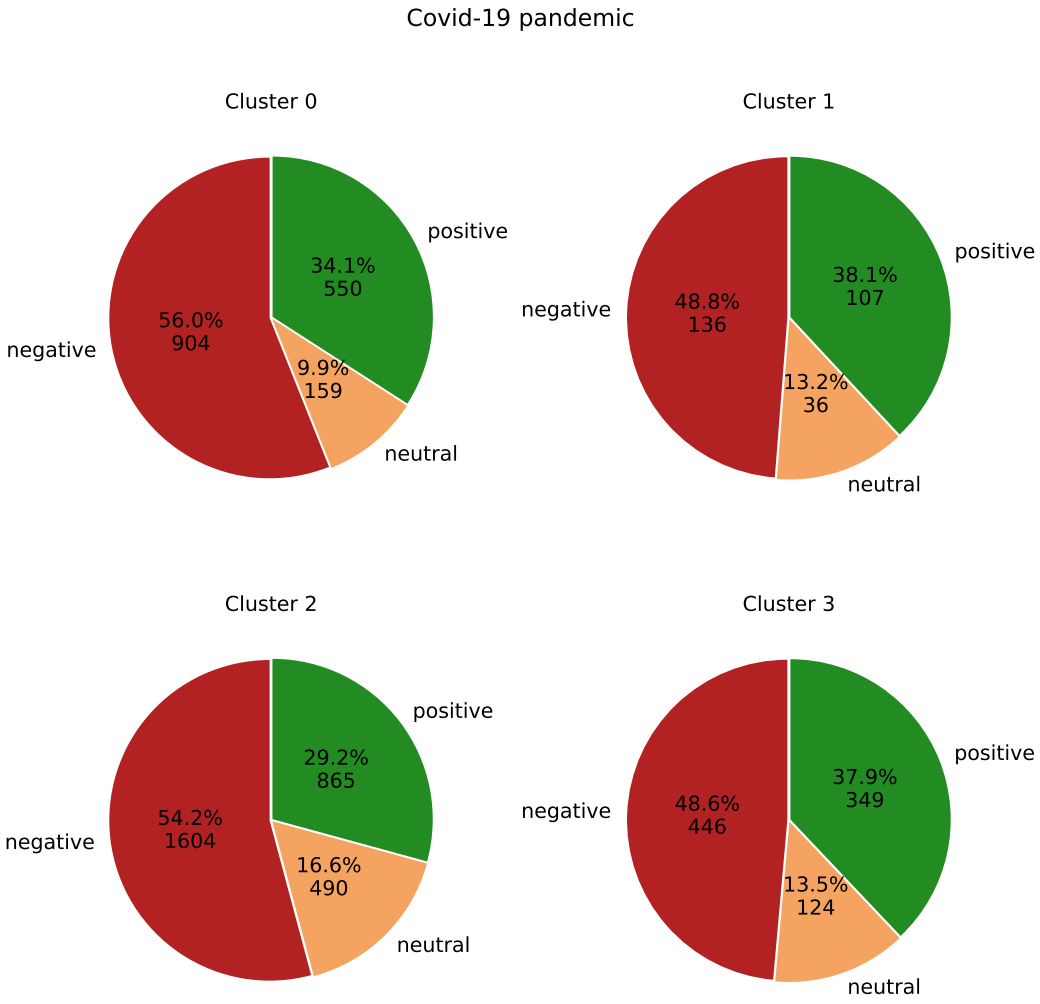

Figure 12: Opinion about "COVID-19 pandemic" per each cluster.

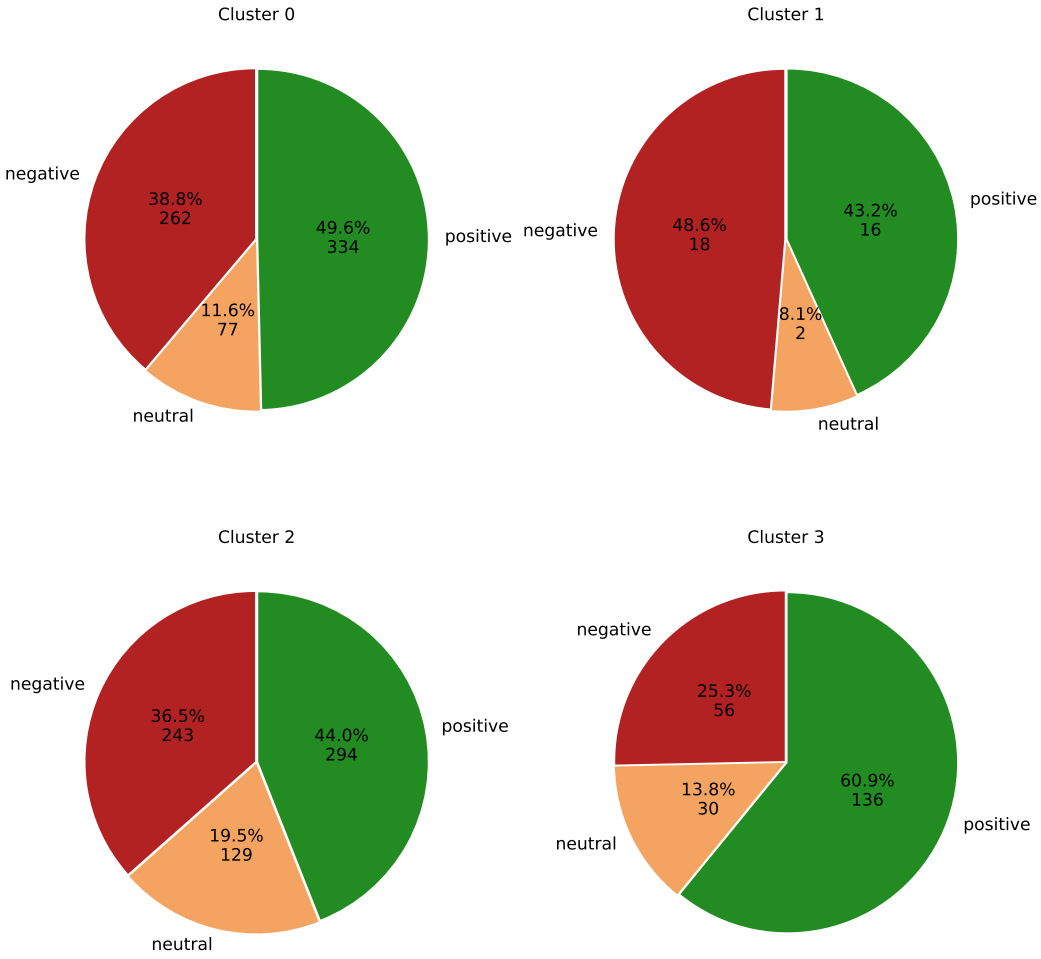

Figure 13: Opinion about **"Abortion"** per each cluster. Keywords: "abortion","abortus".

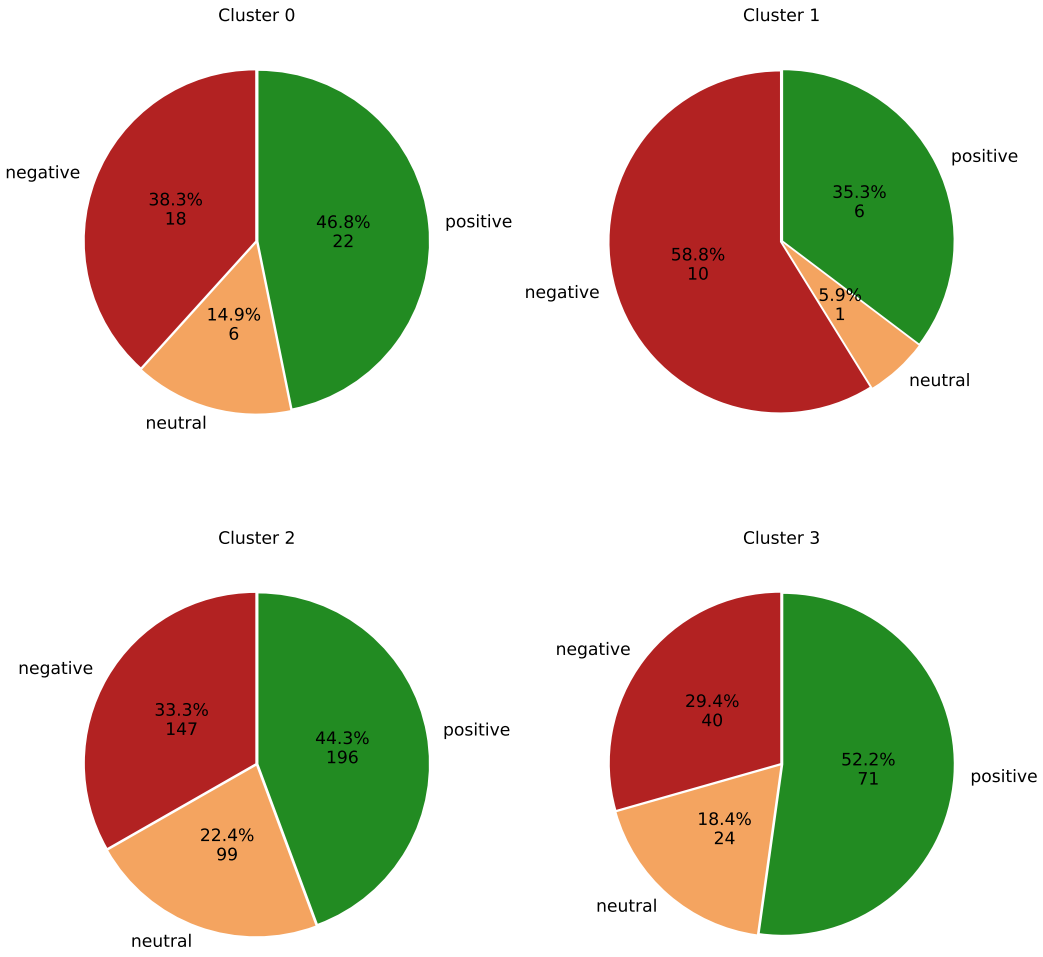

Figure 14: Opinion about **"Aliens"** per each cluster. Keywords: "alien","extraterrestrial","ufo".

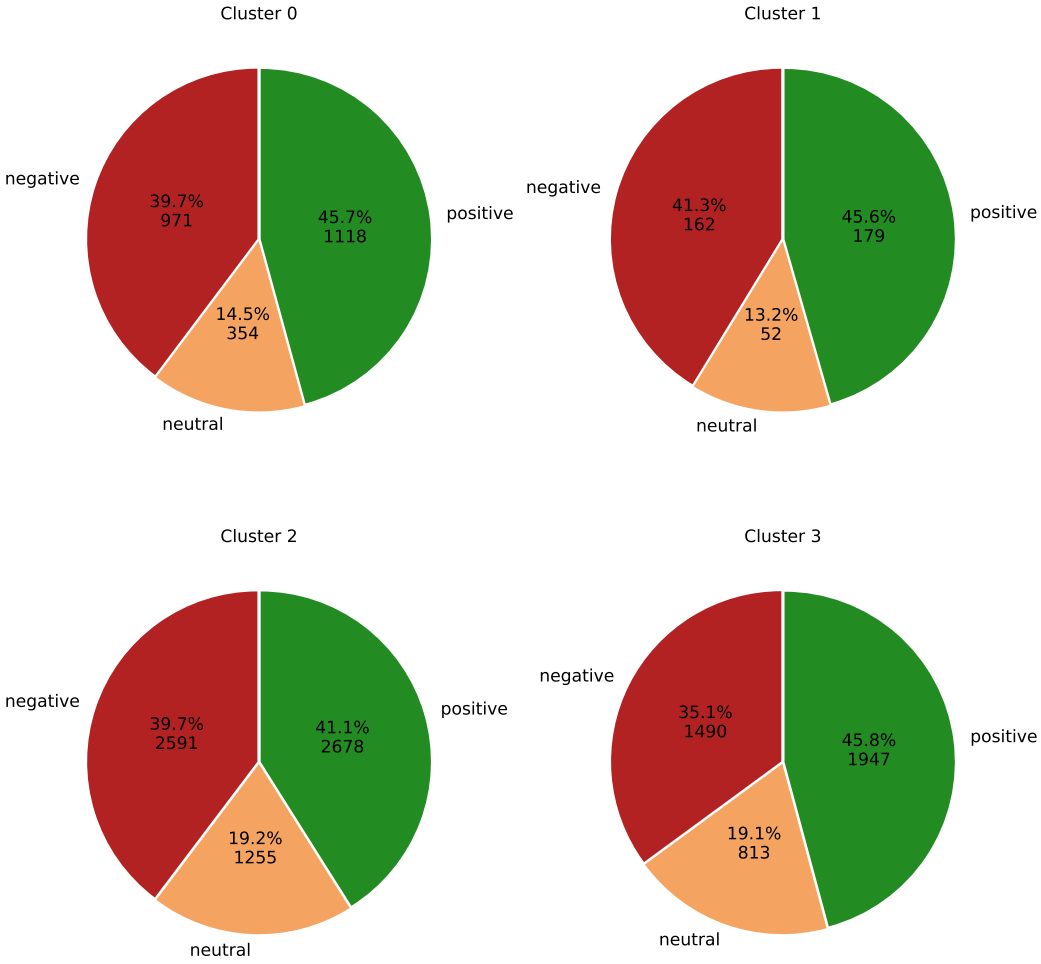

Figure 15: Opinion about **"Religion"** per each cluster. Keywords: "christianity","islam","judaism", "jew","muslim","christian","bible","quran","torah","jesus","yahweh","allah".

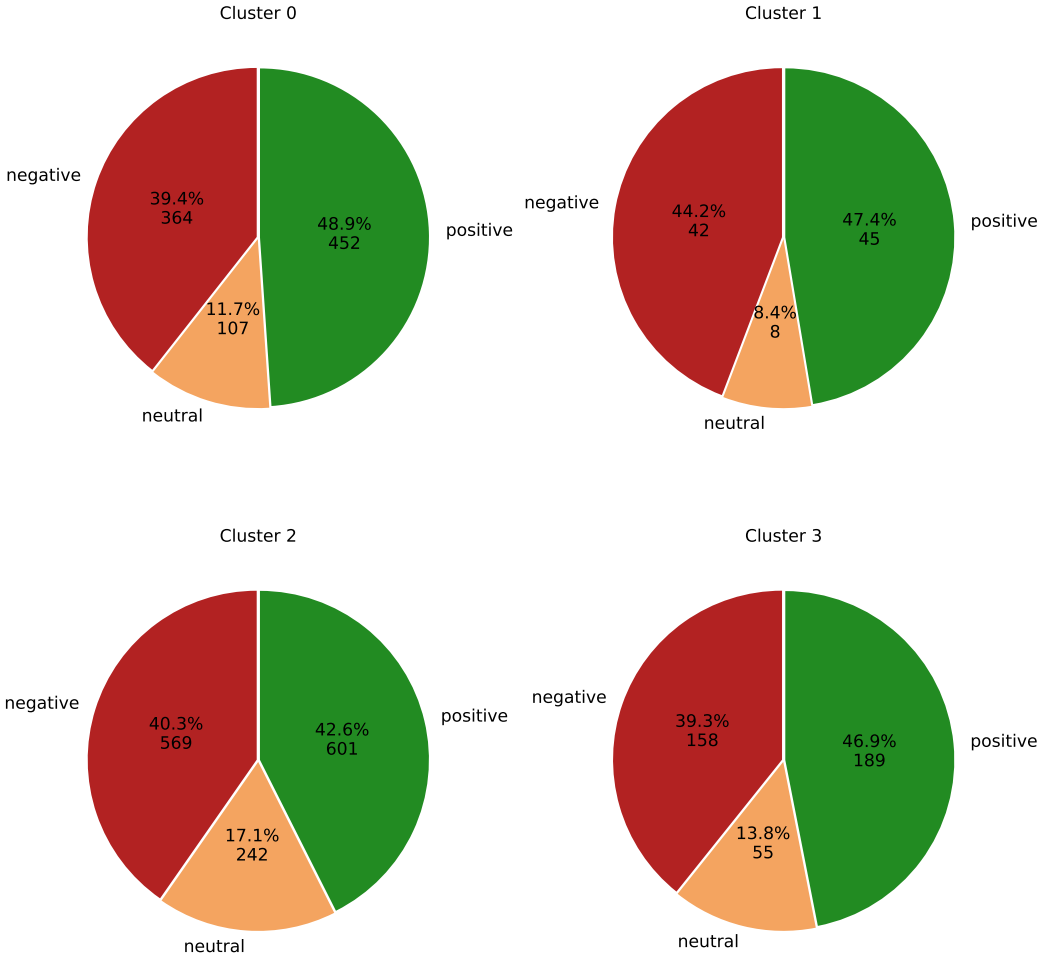

Figure 16: Opinion about **"LGBT"** per each cluster. Keywords: "gay","lgbt","lgbtq+","lesbian", "bisexual","transgender","same sex","same-sex","same-sex","same-sex marriage".

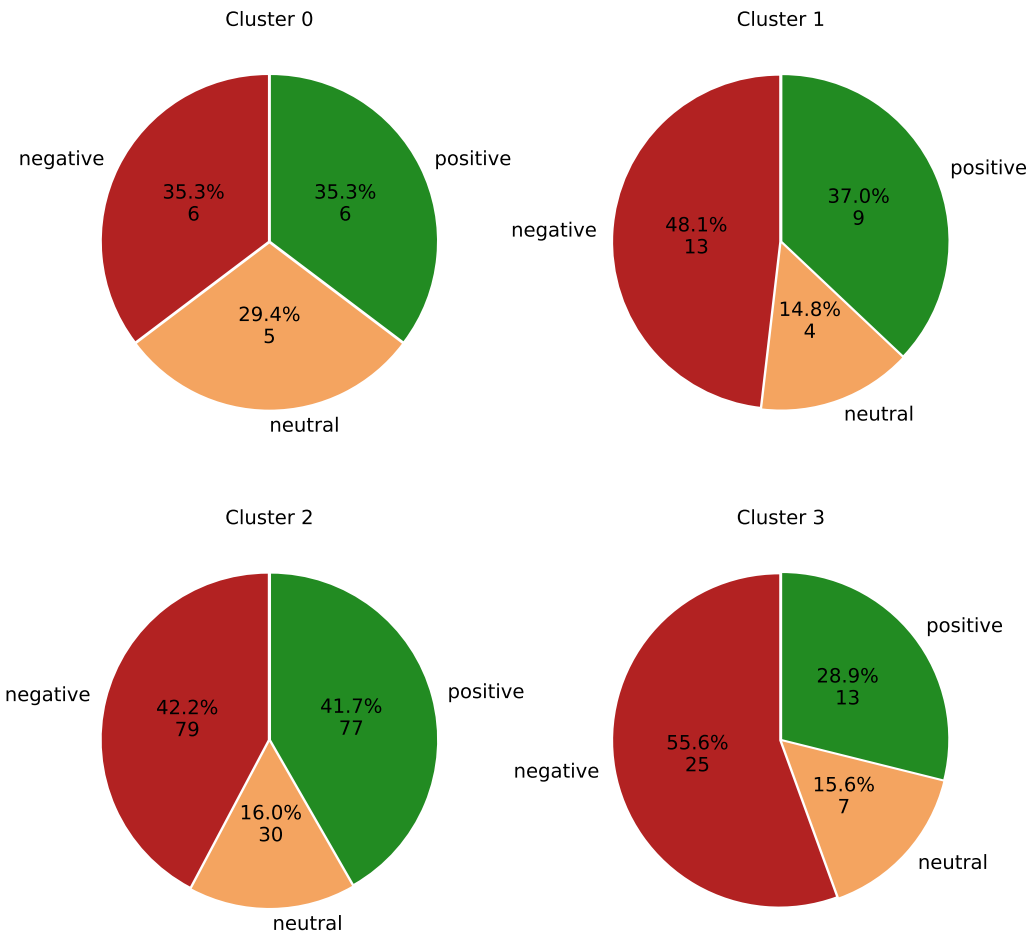

Figure 17: Opinion about **"Woke culture"** per each cluster. Keywords: "woke","woke culture","wokeism".

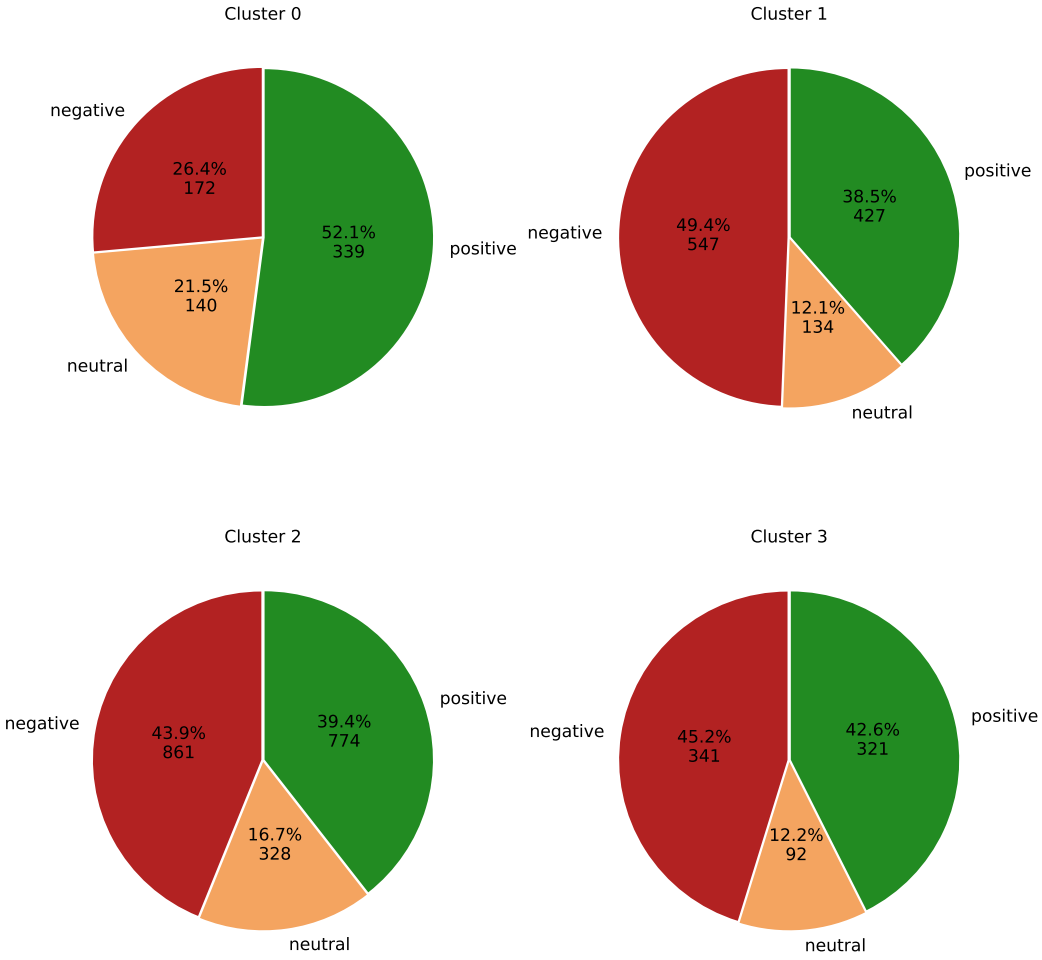

Figure 18: Opinion about **"Donald Trump and Republicans"** per each cluster. Keywords: "trump","republicans","republican","donald trump".

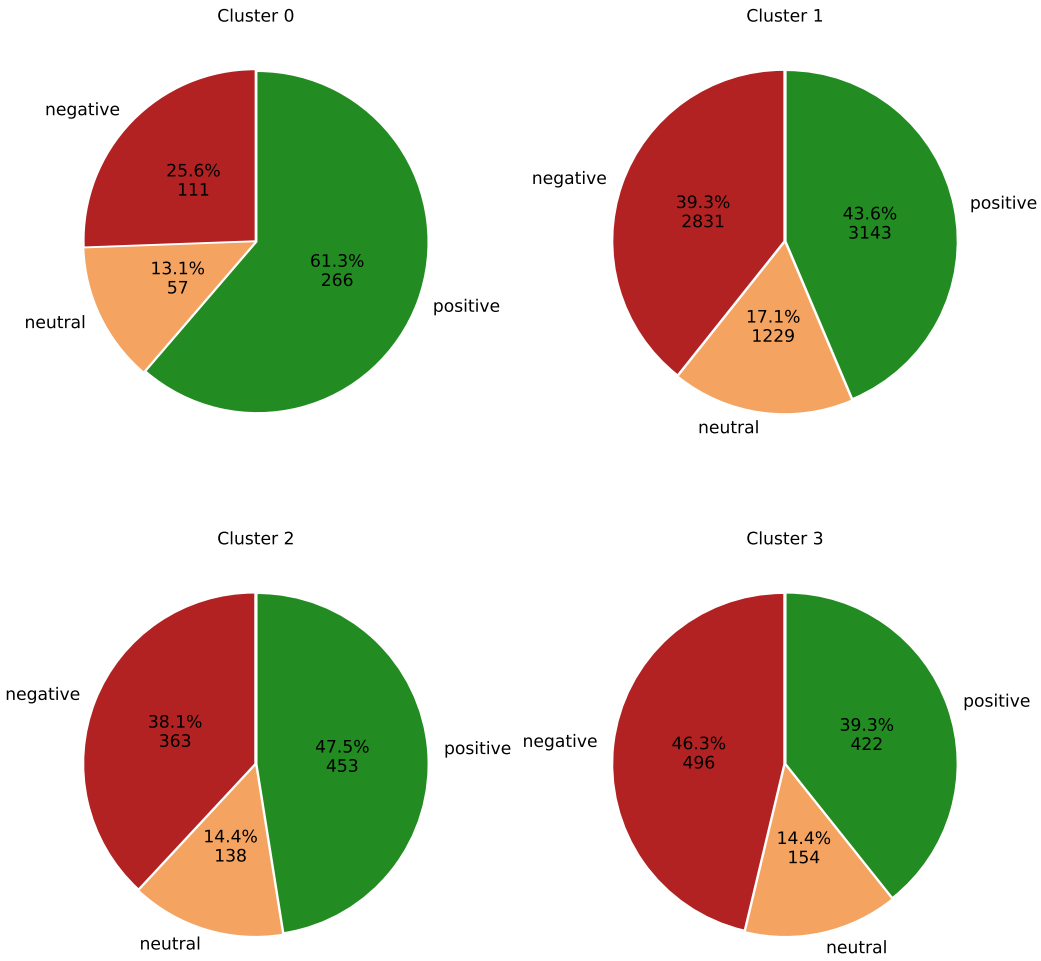

Figure 19: Opinion about **"Joe Biden and Democrats"** per each cluster. Keywords: "joe biden","democrats","biden","democrat".