7 Navigating Swedish Salafism
Large language model-augmented content detection and topic modeling using BERTopic with YouTube metadata

Jonas Svensson
Linnæus university

The chapter suggests and provides an example of a Large Language Model (LLM)-augmented method for gaining a quick overview of large sets of YouTube videos using metadata collected through the YouTube API. The case chosen is the Swedish Salafist YouTube channel *islam.nu* that houses 1680 videos. An LLM (GPT-4o mini) is given a prompt to guess the content of videos based on information given in their titles and descriptions. These guesses are then used in an LLM-augmented topic modeling process utilizing the Python library BERTopic and the HUMINFRA resource, the Swedish Royal Library's sentence-transformers model "sentence-bert-swedish-cased". The videos thus placed under topics are then again subjected to processing by an LLM, to produce easy-to-read representations of the topics. This method provides a convenient way to quickly understand the content of YouTube video sets and can serve as a first step in a purposive sampling procedure.

1 *Stating the problem: So much data, not enough time*

We live in a world with massive amounts of data potentially useful for humanistic research, given an understanding of humanistic research as exploring and explaining public expressions of the human mind (Bod 2013). The current point in history is exceptional in the extent to which such expressions are being recorded and archived. This development, linked to a larger process of digitalization of human communication, is simultaneously a blessing and a curse for scholars in the humanities. The blessing is that the material for

HUM INFRA Jonas Svensson. 2025. Navigating Swedish Salafism: Large Language Model-augmented content detection and topic modeling Using BERTopic with YouTube metadata. In Gerlof Bouma, Dana Dannélls, Dimitrios Kokkinakis & Elena Volodina (eds.), *Huminfra handbook: Empowering digital and experimental humanities* (NEALT Proceedings Series 59), 183–205. University of Tartu Library. DOI: 10.58009/aere-perennius0176

potential study is unprecedentedly large and diverse. The curse is that it is also *overwhelmingly* so.

In the face of this, there is a need for developing new methods for scholars to gain an overview of a vast material, and based on this select appropriate units for in-depth contextual and qualitative interpretation. Digital methods for collecting, organizing and analyzing large amounts of humanistic data have been available for some time now, but those are often reserved for a select few with specialized competences in how to collect and clean messy data, in transforming that data into computer readable forms, and in selecting, employing or constructing tools for further analysis.

In this article, I argue that the emergence of the technology of generative AI and Large Language Models (henceforth LLMs) opens up for a much wider application of computer-assisted methods in the humanities. The important change is that human-computer interaction increasingly utilizes every day, natural language. Some scholars have already discovered the power of the first-generation generative AI in the form of chatbots (such as ChatGPT) for tasks such as text summarization, image annotation, semantic search, and translation. A few have started to use them to augment their everyday workflow, performing tasks that are otherwise tedious and time-consuming.

2 *The current focus*

In this chapter I will showcase how a simple LLM-augmented process of topic modeling can be used to explore content collected from what is arguably one of the richest, diverse and at the same time overwhelmingly large archives of contemporary expressions of the human mind: the video sharing platform YouTube. While exact statistics are difficult to come by, one estimation from 2025 gives the number of 800 million videos on the platform (but no one really knows), with an additional 500 hours of material being uploaded every minute (Global Media Insight 2025).

From my own perspective, as a scholar within the Study of Religions, particularly interested in contemporary Islamic discourse, YouTube is a goldmine of naturalistic data (Potter & Shaw 2018). Here, individuals and religious groups engage in for example disseminating theology, recording actual religious practices and events, and confronting adversaries, all within a multimodal framework that opens for variety of analytical approaches. Furthermore, the YouTube platform, being an expression of Web 2.0 (O'Reilly 2007), also, to some extent, allows for assessment of impact and responses to the cultural expressions it houses, through statistics on views, likes and not least comments.

Given this, it is somewhat puzzling that YouTube has received relatively little attention within the field of the Study of Religions (Bekkering 2019). One of the reasons, perhaps, is that collecting and processing data through the web interface is cumbersome. YouTube's internal search algorithms are opaque. It is difficult to find appropriate material, and above all to assess how representative or illustrative a particular video is, in light of all videos that potentially *could* have been selected.

In the following, I will showcase how to collect and process metadata via the freely available YouTube API (Application Programming Interface) with the aim of gaining a bird's eye view of a set of videos on the platform to facilitate a purposive sampling of material for more extensive, in-depth analysis. For this I will use the power of generative AI in the form of LLMs. The case is limited to the videos in one YouTube channel that falls within my field of academic interest.

3 *Very short on islam.nu and Salafism*

The YouTube channel *islam.nu* forms part of a larger missionary endeavor of a small group of Muslim activists in a Stockholm suburban area. The group has an extensive online presence, with a website and a podcast. It is present on platforms such as Instagram and Facebook. The orientation is clearly and explicitly Salafist. Salafism, as a "new religious movement" (Meijer 2009) in Islam has been the object of much research worldwide. Generally, Salafism is characterized by a strong focus on a literalist, or fundamentalist, reading of scripture, the Qur'an and even more so the Hadith, i.e. records of (alleged) sayings and doings of the Prophet Muhammad. The word Salafi refers to the pious forefathers, *al-salaf al-salih*, the companions of the Prophet and the following two generations, whose beliefs and practices contemporary Salafis strive to emulate, often in minute detail. This involves shunning almost all religious development after this period in Islamic history as illegitimate innovations, *bida'*. Rejecting religious innovation does not, however, mean rejecting all things modern, including modern technology.

Through its various channels, the group *islam.nu* produces a large amount of material that is potentially available for scholarly, humanistic analysis. Analyses have also been conducted to some extent (Sorgenfrei 2021a,b, Olsson 2020, Olsson et al. 2022, Svensson 2022). The methods used often involve selecting instances of communication (e.g. audio sermons, videos, Instagram postings) and submitting these to in-depth study. So far, however, there is to my knowledge no attempt at providing a systematic overview of the material produced by the group, an overview that could assist in the selection process.

4 *The data*

The *islam.nu* YouTube channel is arguably the most influential Salafi oriented channel in Sweden (Svensson 2021). On September 12, 2024, I gathered metadata for all the 1 680 videos housed on the channel, using the freely available YouTube API (Google 2025), downloading information for each video on:

- Video ID
- Date and time of uploading
- Title
- Description
- Tags
- Number of views
- Number of likes
- Number of comments
- Duration.

While I did this programmatically, using a custom-built Python script,[1] there are several free services on the web offering the same functionality. The information gathered is available via YouTube's web interface, but collecting it manually is not particularly time efficient.

When the metadata was downloaded, the channel had 35 600 subscribers. The 1 680 videos had a grand total of 9.3 million views. The last video was published on September 10, 2024 and the first video in the channel was published nine years before that, on September 3, 2015. The mean number of views for the videos was 5 516, but the median was much lower, 1 019, which indicates a spread around the mean. Indeed, most views for a single video was 1.8 million. This video, along with the 14 other most viewed videos, featured Qur'an recitation. It is only on position 16 that we find the first video featuring material produced "in-house", published on July 2018, with a sermon on *adhan*, or call to prayer. That video has 37 737 views.

The total duration of all videos is approximately 885 hours. This is rather overwhelming as it corresponds to half a year full-time employment for a university researcher in Sweden. To assess the content through watching the videos is hence not viable, which is why a more time-efficient method, using metadata, is needed.

1 The script is available as a Colab from the Linnaeus University Language, Cognition and Culture Lab's resource page: https://moodle.lnu.se/course/view.php?id=65873.

There are three pieces of metadata that could potentially provide information on content: the title of the video, the description and the tags. A researcher could use this information to deduce possible content, and perhaps group the videos into thematic categories, or select a particular set of videos where the metadata points to particularly interesting content. Since the channel in this case consists of "only" 1680 videos, to do such an inventory manually is feasible, but still rather tedious, work. The automated process suggested below can provide some relief.

I have in a previous publication, before the mainstream availability of LLMs, made a computer-assisted attempt at finding thematic patterns in a larger set of Swedish Salafi YouTube videos, utilizing NLP methods of word frequency, word co-occurrence, and semantic network analysis (Svensson 2021). This was a rather complicated process, involving much data cleaning and preprocessing. While the methods applied produced results that did appear reasonable, given my domain knowledge in the field, these results were too general to really justify the effort.

Still, during this attempt, I did discover some characteristics of YouTube metadata that are important for the following. I discovered that overall, the most meaningful piece of metadata to use for computer-assisted processing was the video titles. These were, to a larger extent than the descriptions and tags, unique for each video. Descriptions and tags, and particularly the latter, were often generic for the channel, and to include them in a word frequency or word co-occurrence analysis obscured the results.

I also tried to employ some at the time available methods for topic modeling, but possibly due to the short snippets of text in the video titles, and the overall limited size of the corpus (i.e. all the titles), this did not produce any interpretable results. Such methods rely, ultimately, on word frequency and word co-occurrence metrics, and require larger amounts of text to identify patterns (Golub 2020).

When returning to the same problem a few years later, the technology has advanced. The shortcomings of earlier methods could, at least hypothetically, be overcome through:

1. Providing a label for each video indicating the content, based on *relevant* metadata.

2. Arranging these labels of content into meaningful groups, based on semantics, rather than on word frequencies or word co-occurrences.

For both tasks, I could use the power of LLMs, but in different ways.

5 *Short on BERTopic*

BERTopic is a tool for topic modeling constructed by Martin Grootendorst, and introduced in 2020, but gaining more widespread recognition from 2022 and onwards (Grootendorst 2022). Put in very simple terms, BERTopic takes a corpus of text units (called "documents") as input and uses a pre-trained LLM to arrange these text units in groups based on their *semantic* similarity. These groups, if distinct enough from one another, can be said to relate to a topic or theme in the full corpus. The power of BERTopic lies in the semantic embedding of text units. This embedding does not rely on the corpus itself, but on a huge number of other texts on which the model has been trained. This is like a manual process where the researcher invokes all their previous knowledge in attempting to categorize a particular text unit. Given that it is the semantic content of the document that is used as an arranging principle, the tedious tasks of preprocessing, tokenization, stop-word removal, stemming and/or lemmatization necessary for many previous forms of topic modeling are not a requirement. On the contrary, such manipulation of the text units may obscure the results.

BERTopic has become popular. To date, there are on Google Scholar more than 3 500 scientific papers citing the tool. Some of these contain comparisons with previous methods for topic modeling (Gan et al. 2024, Egger & Yu 2022, Turan et al. 2024), concluding that it is often superior. BERTopic is also, due to a modular structure, highly customizable, not least concerning the main operation, the sentence embedding. Within the framework of HUMINFRA, the KBLab at the Swedish Royal Library has developed a workshop in Google Colab using its own sentence embedder, "sentence-bert-swedish-cased", finetuned for Swedish on the Library's collections (KB Lab 2025). This is also the embedder used in the following.

6 *Preparing texts for BERTopic invoking GPT-4o mini*

For BERTopic to work at its best, the "documents" should then be in a form that makes sense semantically. This is not always the case with YouTube metadata. To give an example from the current set there is the following information on title and description for one video:

Title Surah Al-Alaq - Omar Hisham

Description

Koranen | Kapitel 96 - Al-'Alaq

Recitör: Omar Hisham

https://www.youtube.com/playlist?list=
PLaIDlec9j7J2o2Dy3yiMynYtZnPDIfkfZ

Here, the title alone does provide some information that could be interpreted as relating to the Qur'an, chapter 96. The name "Omar Hisham" does not provide any additional information. Combined with the information in the description, however, a qualified guess could be made. Here, Omar Hisham is also given a specific role, as a reciter (albeit the Swedish word is incorrect. It should be "recitatör" and not "recitör"). For someone with at least basic domain knowledge in Islam, a guess concerning the content would be something like: "A recitation of Sura al-'Alaq (96) in the Qur'an by Omar Hisham". Such a descriptive label could, potentially, facilitate the topic modeling process. Going through all the metadata manually, producing such guesses for all videos that could then be subjected to a topic modeling process would most probably defeat the purpose of using computer-assisted methods, as it could give a fairly good overview of recurring themes. And it would take time.

Summarizing content is an area in which current LLMs excel. Entering the information of title and description into the web application ChatGPT (using GPT-4o) with a prompt to guess the content produces the following:

> This YouTube video likely features a recitation of Surah Al-Alaq (Chapter 96 of the Qur'an), performed by Omar Hisham, who is identified as the reciter. The video may include a traditional Arabic recitation, possibly with subtitles or explanations in another language (like Swedish, given the term "Koranen" for "Qur'an" and the Swedish spelling of "Recitör").

This is impressive, but also contains perhaps a bit too much guessing. Still, entering each piece of metadata into a web based Chatbot is probably more time consuming than guessing the content manually.

To automate the process, I created a simple script in Python where I invoked OpenAI's recent model GPT-4o mini via the OpenAI API. There are of course several LLMs that could have been chosen for the task, including the full GPT-4o model that is the default model under the hood for ChatGPT, but GPT-4o mini is, at the time of writing, widely recognized as capable, fast, reliable and cost-efficient.

The task given to GPT-4o mini was to use the information in the title and description metadata for each video to guess the content. As an extra task, I also prompted the model to extract names of persons possibly featuring in the video, if mentioned (in the case above Omar Hisham). The latter was merely a test of using LLMs for Named Entity Recognition (NER) and it will not be further considered below.

The following was the prompt instruction. It was given in Swedish and has here been translated into English.

> For each object in an array of JSON objects, create two new key/value pairs.
>
> 1. A qualified guess about the content of a YouTube video based on the values of the video's title and description. The guess should be in Swedish, limited to the information in the title and description, without adding information that is not explicitly mentioned.
>
> 2. Information about who is speaking in the video. If the speaker cannot be identified, leave SpeakerGuess as an empty string.
>
> Return only the video Id and the two key/value pairs ContentGuess and SpeakerGuess.
>
> [Here, I included a list of further detailed instructions for the format of the output, and also an example of the desired output, a so-called "one-shot prompting"]
>
> Keep strictly to the example. Do not add the word JSON to the output. Make sure the output is in a strict JSON format. If you cannot guess the content, let the title be your guess.

GPT-4o mini delivered guesses for the content of all the 1680 videos. The titles and descriptions were presented to the model in batches of 15 videos at a time, to fit into the model's context window. The whole process took 11 minutes. For the example given above, the result was: *Recitation av Koranens kapitel 96, Al-'Alaq* (Recitation of the Qur'anic chapter 96, al-'Alaq).

7 *Topic modelling with BERTopic*

The result of the LLM-augmented content guessing was used as the "documents" for BERTopic to work with. The KBLab pretrained model for Swedish is showcased in the Colab workshop available via HUMINFRA (KB Lab 2025). The documents here are Swedish parliament petitions, perhaps an ideal test case, and the results returned are highly impressive. The data

Table 1: Comparison between BERTopic results with different minimal number of items in topic setting.

Min items per topic	Topics	Outliers
5	87	301
10	51	394

for the current chapter, i.e. the LLM-generated labels for YouTube videos might be considered less ideal. They are in Swedish, but also in a particular Swedish "religiolect" (Hary 2011) that mixes Swedish words with specific religious terms with Arabic origin, at times also using archaic Swedish words as well as idiosyncratic neologisms. It is not conceivable that there is enough material in the training data for "sentence-bert-swedish-cased" to cater for this particular "Islamic Swedish". I did try out the default BERTopic model in multilingual mode, but since this article focuses on the use one of the HUMINFRA resources, I will not follow up with a model comparison here.[2]

The output of the BERTopic topic modeling is in the form of a list of topics, based on a procedure that places *semantically* similar text units (i.e. the content guesses) in proximity to one another and identifies clusters of similar texts that are sufficiently distinct from other clusters of similar texts. Unlike other tools for topic modeling, BERTopic does not force all texts into such clusters. Texts that are difficult to assign to a cluster are instead considered to be outliers.

Besides choosing the pre-trained model to use for the embedding of text units, there are several other parameters that could have been adjusted. Here, I used the default settings in the KBLab workshop, except for one. Since the number of videos in the set is small, I decided to set the minimum number of items under each topic to five instead of ten. This was done to reduce the number of outliers (see Table 1)

The topic modeling process took 40 seconds. The result was 87 topics and a total of 301 videos as outliers. It is possible do further processing in BERTopic to reduce the number of outliers (Grootendorst 2025c), forcing them under topics, but for the following, I chose to disregard the outliers altogether.

2 The results from the two different models, using a minimum of five and ten items per topic, are available for comparison at https://github.com/jsnhum/huminfra_cookbook.

8 *What is under the topics?*

The default way in which BERTopic represents the topics is through a list of words particularly significant for the texts under that topic. Significance is calculated using a version of the established metric of TF-IDF or "Term Frequency-Inverse Document Frequency".[3] In simple terms, this procedure attempts at extracting terms significant for a particular text, in relation to all other texts in a corpus to which the text belongs. This is done by identifying terms that occur frequently in the text and weighing this against the number of other texts in the corpus in which the same term occurs. The same principle is used in BERTopic, but with the difference that here it is the frequency of a term in all texts under a topic that is weighted against the number of topics whose underlying texts contain that term. The procedure is referred to as "Class-based TF-IDF" (c-TF-IDF; Grootendorst 2025b).

In Figure 1, there is a hierarchical layout for all 87 topics with representations generated by c-TF-IDF, displaying the three most significant terms. The dendrogram, produced in BERTopic, also indicates calculated similarity between topics.

The results from c-TF-IDF may not be an ideal way to represent the topics. The content guesses are short snippets of text, which skews the metrics of term frequency. Even single words may have large relative frequency. Furthermore, the use of neologism and Arabic terms may also obscure the result.

Still, even in the c-TF-IDF-based representation one does find those that do make sense.

Hence, for example, the representation for topic number 45 is easy to interpret. It has the following representation (using the ten most significant terms): *the atheism*, *critical*, *atheism*, *analysis*, *simplified*, *counter arguments*, *trailer*, *society*, *against*, *focus*. An educated guess would be that this topic covers videos in which atheism is discussed critically.

Deciphering other topics may require some more domain specific knowledge. Number 55 has the representation *wudhu*, *instructions*, *guidelines*, *performance*, *learn*, *video*, *you*, *how*, *knowledge*, *lecture*. This most probably indicates videos with instructions on how to perform *wudu'*, a form of washing to achieve ritual purity.

Yet other topics are more difficult, if not impossible, to infer from the representation, such as topic 79, with the representation: *200*, *the belief dogma* [*trosläran*], *Islamic*, *questions*, *course*, *call to prayer*, *iqamah*, *inclusive*, *11*, *doubt*.

3 For a highly accessible introduction to TF-IDF, see Lavin (2019), who also provides some history for the technique, dating back to the 1970s.

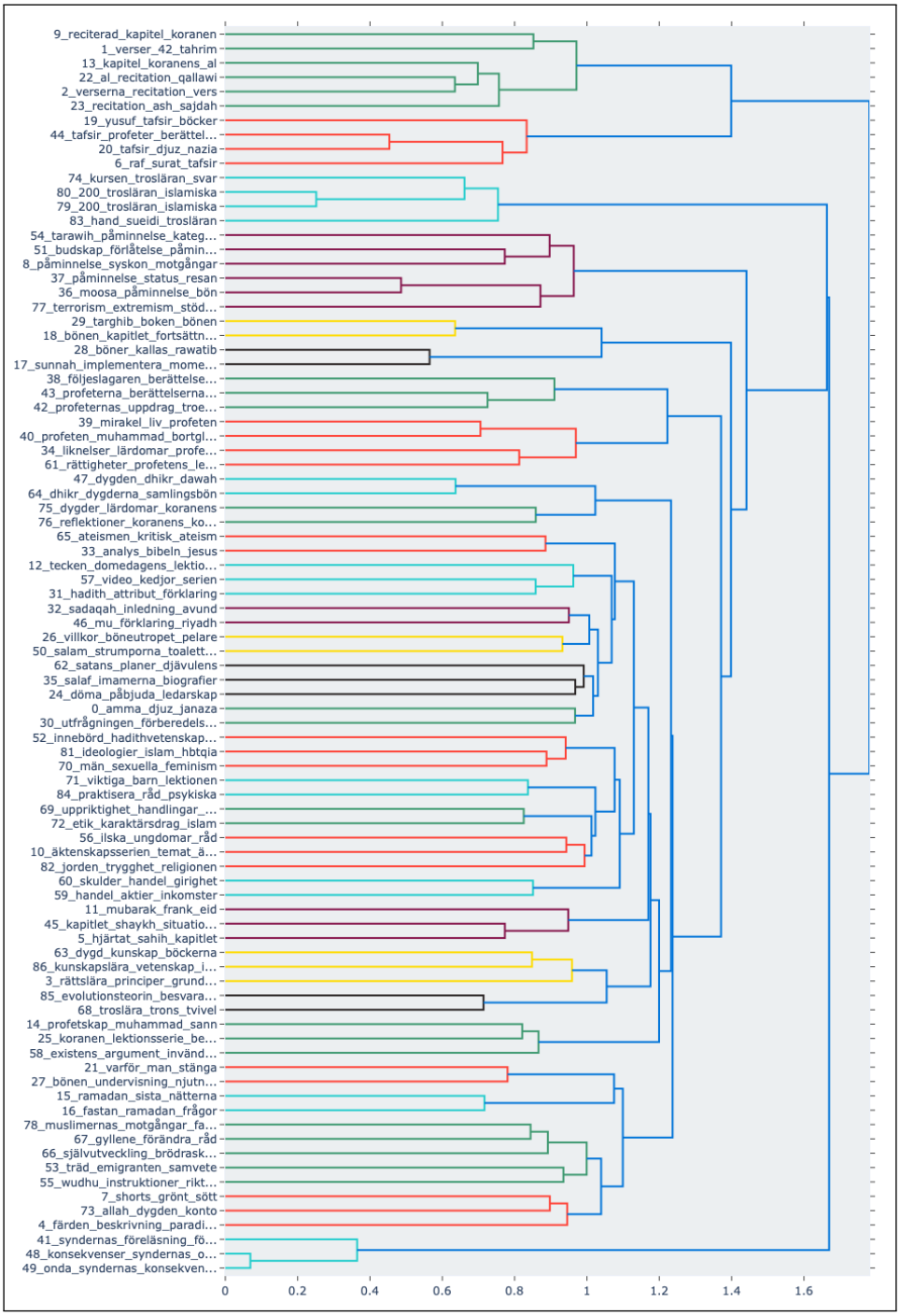

Figure 1: BERTopic hierarchical clustering

It should be remembered, however, that the grouping of videos under topics is done *before* the representation is extracted through c-TF-IDF. The grouping is based on semantics using a pretrained model, not on word prevalence or frequency neither in GPT-4o mini's content guess, nor in the corpus of all such guesses. This means that there are other ways to arrive at useful representations of the topics.

9 *GPT-4o mini again*

One such way is to again invoke the power of LLMs. This is suggested also in the BERTopic documentation, with instructions on how to implement this directly in the process of topic modeling (Grootendorst 2025a). Here, however, given the specific material at hand, I chose to do this representation on the side.

Each topic arrived at by the BERTopic modelling contains a set of short text (the content guesses). I provided GPT-4o mini with these sets with the following prompt (translated from Swedish):

> You are an assistant with great competence in finding summary labels for a list of short descriptions of YouTube videos from a Salafist YouTube channel. You receive a list of descriptions of YouTube videos and provide a summarizing title. The title should be short.
>
> [here follows an example]

The result can be seen in Appendix 1. The original labels in Swedish have here been translated into English.

10 *Interpreting the results*

Now then, we have arrived at a list of 87 topics covering 1 379 of the total 1 680 videos present in the *islam.nu* channel. BERTopic offers much additional functionality beyond identifying topics, in terms of further exploring and arranging those topics. For the purposes in this chapter we can now turn to a qualitative interpretation based on (1) the LLM generated topic representation (2) the hierarchical outline of topics as displayed in the dendrogram in Figure 1 and (3) domain knowledge to achieve what was presented as the goal in the outset, a bird's-eye characterization of the content of the *islam.nu* YouTube channel, and a guide for further purposive sampling.

Islam.nu hosts material that could be expected of a Salafi oriented YouTube channel: a strong focus on basic beliefs and rituals, the scriptures, the prophet

Muhammad and early Islam. Eighty-seven topics may appear a lot, and some of them could probably be merged, especially considering topic similarity as indicated in the dendrogram. One of the more obvious examples of this is the "topic" already mentioned several times: Qur'an recitations. It would appear as if these generally fall under topics 1, 2, 9, 13, 22, 23, all joined together at the top of the dendrogram. As mentioned above, Salafism is highly oriented towards the Islamic scriptures, the Qur'an and the hadith, and historically towards the early period in Islamic history. Both the Qur'an and the hadith recur often in the topics. When history is addressed, it is either the mythological salvation history of Qur'anic figures, or the Prophet Muhammad and his companions. There are few, if any, topics that concern the historical development of Islamic tradition. This is in line with the character of prototypical Salafism as a "deculturalized" form of Islam, where there is a discursive void stretching between the early history of Islam and the here and now.

There are other significant traits of Salafism noticeable in the topic representations. There is a specific topic on *hisba,* i.e. the ritualized practice of "informing" other Muslims about what God expects of them and what He forbids (topic 24), often from a somewhat elevated position of a select few (i.e. Salafis themselves). Such advice to other Muslims appears to be on basic beliefs, ethics and morals, sins and virtues, and obligatory ritual practices. It is striking that many of the topics appear to address what could be seen as "basic" religious beliefs and practices (fasting, almsgiving, beliefs in afterlife, paradise etc.), already known to most Muslims. This mirrors a common view among Salafis that although Muslims may have some knowledge and beliefs about these basics, that knowledge is incomplete or flawed and above all tainted by established cultural understandings and innovations in belief and practice. A general feature of Salafism is the stated need for beliefs and practices to be "purified" from anything that does not have a clear foundation in the scriptures. One can note the apparent strong focus on such a basic practice as prayer stretching over several topics (17–18, 21, 26–29). But "purification" is not only metaphorical. That there is a special topic on cleanliness and hygiene (50) does rhyme well with Salafism in general being strongly (some would claim obsessively) attentive also ritual purity, and whatever substances or actions that can jeopardize such purity.

While they appear rather scarce, there are also topics that goes a bit beyond the issue of correct individual faith and practice. Such include what appears to be comparisons between Islam and other forms of belief systems or systems of thought that challenge it, and I would assume, to the benefit of the former. Those other include atheism (65), science (85), Christianity (33) and political ideologies (81). There appear to be distinct topics that address

some issues on social relations, such as marriage and family relations (10, 56) and gender (70), as well as topics that address economy and commerce (59, 60). One topic could be mentioned in particular, i.e. number 77, represented as "Islam, Extremism, and Terrorism". Previous research on the group behind *islam.nu* has identified them as belonging to a particular category of "puritan" Salafism, not partaking in politics, at least not party politics, and particularly taking public stand against militant forms of Salafism (Ranstorp et al. 2019).

The above summarization provides a bird's-eye view. It gives some very broad outline of the *possible* content of the *islam.nu* YouTube channel. That this outline corresponds to what is already known, and somewhat expected, could be taken as an indication that the topic modeling was successful. Further probing into the LLM generated content guesses for videos housed under each topic would perhaps give a better basis for such evaluation. I cannot do a full-scale evaluation here, and it would defeat the purpose. But I would like to point to a couple of topics that can help illustrate why I deem BERTopic, and its reliance on pre-trained models, useful for the task at hand.

Topic 30 received the LLM-generated representation "Death and the Afterlife ", which could be seen as rather specific. The topic houses six videos. Four of these contain the word "death" in the title and the description, and in the LLM-generated content guess. However, two do not: one bears the title "Reflections from the funeral" and the second "The questioning in the grave", referring to a particular notion of partial punishment/rewards in the grave between death and final resurrection. From a semantic perspective, and judging from the titles, the placement of these videos under topic 30 does make sense.

Another example can be found in the topic "Bible Analyses and the Islamic View of Jesus" (33). Although Jesus, Islam and the Bible are words that occur in the LLM-generated content guesses of four of the eight videos under this topic, all are absent from the remaining four. BERTopic, however, judged (correctly in my judgment) the content guesses "Analysis of the Gospel of Mark", "Analysis of the Gospel of Luke", and "Analysis of the Gospel of John", to belong under this topic, resting on the inbuilt "knowledge" of the underlying LLM of that there is a connection between the words "Jesus", "Bible" and "the Gospels".

It should be stressed, albeit it is not an issue to be pursued further, that the topic modeling of YouTube videos using BERTopic showcased here makes it easy to combine the results with additional pieces of metadata, in order to asses for example if some topics receive more interaction than others (in terms of the number of views, likes and comments) or if certain topics are more prevalent in certain periods of time than in others. Such processing of

the results may be interesting, but they are perhaps not at the core of much humanistic research, with a stronger focus on detail, on careful, culturally contextualizing analysis of discourse and behavior, and a quest to interpret and understand individual expressions of the human mind, i.e. in this case the actual content of the videos.

This, then suggests topic modeling not as an end in itself, but as a means to an end, more specifically the possibility to find, in a large collection of data, specifically representative, or illustrative, cases for in-depth analysis. The task here might be identifying material that can help answer certain predetermined questions (such as *islam.nu* views on e.g. gender roles or terrorism), or to identify new, previously unattended to topics suitable for in-depth study.

11 *Working further*

As much as I would have liked to provide an example of such an in-depth exploration of one of the topics, it is not possible. The videos most certainly contain material where individuals express their religious beliefs. As such it becomes sensitive data which, according to both Swedish and European legislation, is forbidden to collect or process. An exception to this general prohibition could possibly apply in the current case, when the sensitive data has been manifestly made public by the research person. *Islam.nu* is an open missionary channel, reaching out to a general public, Muslim and non-Muslim, and the individuals that feature in videos are publicly professing their beliefs. However, whether this exception applies is not for me to decide. The decision lies with the Swedish national ethical board, to whom a lengthy application must be submitted, together with a fee of 5000 Swedish Crowns. This cumbersome procedure has (rightfully in my opinion) been the object of recent criticism in Swedish academia, particularly from the humanities and the social sciences where publicly available information on expressed beliefs, chosen ethnic affiliations and political stances of e.g. poets, novelists, opinion makers, other academics or religious and political leaders constitute important data. Nevertheless, non-compliance with the demand to have research on sensitive data preapproved is a criminal offence, punishable by fines or imprisonment.

There is, however, nothing hindering me from, as a closing segment, outlining a process for how a further probing into the results from the current topic modeling, for example on the videos categorized under the topic "Critique of Atheism and Arguments Against Atheism" (65). It is interesting, since it belongs to a set of topics in which *islam.nu* appears to direct attention

not only towards Islamic tradition, but also towards the larger society. The topic houses seven videos. A word search on the word stem *ateis-* 'atheis-' in the titles of all videos reveals three additional videos that may be of relevance. This is important, since it shows that there is reason to double check the results of the topic modeling. It does not, however, mean that BERTopic failed. It must be remembered that BERTopic places an item (here a content guess) under one, and only one, topic. Hence, a video guessed to feature an "Interview with a Muslim convert concerning his journey from Atheism to Islam", was placed under topic 84, "Guidance and Advice for Prospective Muslims", in which most videos apparently target conversion to Islam. Another video was labelled "Atheism | Analysis of Western ideologies and their relation to Islam", and placed under topic 81, "Political Ideologies and Beliefs". GPT-4o mini labelled the third and last as "A series of lectures on how to handle doubt and atheism", which BERTopic considered to be an outlier.

Adding these three videos to the sample, the total number of videos for a possible in-depth analysis is ten, with a total length of slightly more than seven hours. Assessing the content by actually watching is perhaps not the most time efficient way to approach the material. Some YouTube videos come with subtitles, and in some cases, the transcript can be accessed from within the YouTube platform. Not all videos have that option though, and the ten videos in this set do not.

Recent amendments to copyright law makes it possible for researchers in Sweden to copy and (securely and temporarily) store copyrighted material that they have legal access to for the purpose of datamining. This would allow for the downloading of the ten videos, as videofiles or soundfiles, and submit them to perhaps one of the more conspicuous advancements in research that the current LLM revolution has to offer: automatic transcription. The state-of-the-art tool for this is OpenAIs Whisper, a free, multilingual transcription (speech-to-text) tool, which, incidentally, is rumored to have been created for the purpose of harvesting data from recorded speech online, and in particularly on YouTube, to use for LLM training (Metz et al. 2024).

Running Whisper on Google Colab transcribes sound at a rate of approximately a third of the recorded time, i.e. a video of one hour transcribes in 20 minutes, with often next to perfect result.[4] To use this method for sensitive data, however, might not be advisable since it means sharing the data with Google. In the case of YouTube videos, however, this may be less of a problem, since they are already, at least partly, owned by Google.[5]

4 For a publicly available Google Colab page for transcription via OpenAI's Whisper (with instructions in Swedish), see Linnaeus University Language, Cognition and Culture Lab's resource page: https://moodle.lnu.se/course/view.php?id=65873.

One hour of speech amounts to around 15 pages of text, single spaced. So, the total number of pages of transcripts of the 10 videos in this example would be around 100.

Transcribed speech is generally not a pleasant read. If I were to assess further the content of the videos, I would probably first of all submit the texts to an LLM for summarization and information retrieval. While not totally reliable, I have found that an iterative process of oscillating between original files, and the LLM summarization facilitates the identification of segments of text relevant in-depth analysis. In the current case, for example, conceivable questions to ask would be for the LLM to extract segments from the texts that provide argumentations against atheism, discussions on how atheism relates to Islam and definitions of atheism.

12 *Concluding remarks*

The example given here of an LLM augmented topic modeling in search of patterns and themes in sets of YouTube videos, using easily accessible metadata, does seem to hold some promise for anyone interested in systematically exploring this enormous, and continuously expanding, archive of expressions of the human mind. The metadata collected through the YouTube API is here limited to one channel and 1680 videos. The method can easily be scaled up, to process information contained in much larger datasets (see e.g. Svensson, in press; Svensson, forthcoming). Here it can be mentioned that the YouTube API also allows for collection of data via searches that can circumvent the search algorithms applied in the web interface.

However, some words of caution are warranted. The use of an LLM to produce more semantically meaningful "guesses" of the content in videos, using titles, descriptions and potentially tags, to gain better results from the BERTopic modeling, is risky. LLMs are still prone to hallucinating, providing responses that are not in line with the expected, in unpredictable ways. Although I have not systematically assessed the quality of GPT-4o mini's guesses for the current article, I have found no examples of hallucinations. I did, however, find that the results generated both in content guesses and in the final LLM-generated representations where not fully consistent over consecutive attempts. There were slight variations in wording and focus.

I have above limited myself to text content in the metadata (and hypo-

5 To cater for the bulk transcription of sensitive data, Linnaeus University has, within the framework of the Huminfra infrastructure, invested in a custom assembled, stand-alone computer to run Whisper. It currently transcribes sound at a speed of approximately 1:1 and is housed in Linnaeus University Language, Cognition and Culture Lab (LiLa).

thetically in transcripts). This is a limitation, given the multimodal character of YouTube data. BERTopic is not (yet) multimodal. However, there are ways in which image material can be processed, again invoking the power of LLMs, or perhaps rather LMMs, "Large Multimodal Models". Besides textual metadata, the YouTube API allows for downloading image material, or more specifically thumbnails of the videos. Such image material can become an additional basis for the content guesses, using a multimodal model that can perform image annotation.

A more accurate topic modeling than the one above would be to transcribe all the videos, alternatively download subtitles when available, and use these texts as the "documents" submitted to BERTopic. However, BERTopic has a limit to the length of texts that it can process, around 300 words. Transcripts would have to be summarized, a process that could be done using LLMs. However, this would incur significant costs, at least if commercial models were to be used. Just as there are LLM-powered techniques for (relatively) quick transcription of spoken content in the videos, there are also methods to summarize the image content, using an LLM with "vision" capabilities (for an example, see OpenAI 2025).

Such expansions to increase the quality of the topic modeling would, however, beat the purpose of the procedure outlined in this chapter. The aim has been to introduce a quick and cost-efficient method to arrive at a bird's eye view of YouTube material and facilitate purposive sampling. The whole process, from downloading the metadata from the YouTube API to arriving at a dataset with videos assigned to a topic with an LLM generated representation, takes around 20 minutes. It is only considering how much time and effort the same procedure would demand if done fully manually that the LLM-augmented method can be evaluated, considering also the loss of the researcher's full control over the processes that it involves. Still, it is a method that does not challenge or replace the specific competences associated with humanistic research, in terms of an informed culturally contextualizing interpretation of expressions of the human mind. It is merely a method for finding such expressions and perhaps justifying that they are made into objects for in-depth analysis in the first place.

References

Bekkering, Denis J. 2019. Studying religion and YouTube. In Alphia Possamai-Inesedy & Alan Nixon (eds.), *The digital social: Religion and belief*, 49–60. Berlin: De Gruyter. DOI: 10.1515/9783110497892-003.

Bod, Rens. 2013. *A new history of the humanities: The search for principles and patterns from antiquity to the present*. Oxford University Press, USA. DOI: 10.1093/acprof:oso/9780199665211.001.0001.

Egger, Roman & Joanne Yu. 2022. A topic modeling comparison between LDA, NMF, Top2Vec, and BERTopic to demystify Twitter posts. *Frontiers in Sociology* 7. DOI: 10.3389/fsoc.2022.886498.

Gan, Lin, Tao Yang, Yifan Huang, Boxiong Yang, Yami Yanwen Luo, Lui Wing Cheung Richard & Dabo Guo. 2024. Experimental comparison of three topic modeling methods with LDA, Top2Vec and BERTopic. In Huimin Lu & Jintong Cai (eds.), *Artificial Intelligence and robotics*, 376–391. Singapore: Springer.

Global Media Insight. 2025. *YouTube statistics 2025* (*demographics, users by country & more*). https://www.globalmediainsight.com/blog/youtube-users-statistics.

Golub, Koraljka. 2020. Automatic identification of topics: Applications and challenges. In Joacim Hansson & Jonas Svensson (eds.), *Doing digital humanities: Concepts, approaches, cases*, 5–26. Växjö: Linnaeus University Press.

Google. 2025. *Add YouTube functionality to your app*. https://developers.google.com/youtube/v3.

Grootendorst, Maarten. 2022. *BERTopic: Neural topic modeling with a class-based tf-idf procedure*. https://arxiv.org/abs/2203.05794.

Grootendorst, Maarten. 2025a. *6B. LLM & generative AI*. https://maartengr.github.io/BERTopic/getting_started/representation/llm.html.

Grootendorst, Maarten. 2025b. *C-TF-IDF*. https://maartengr.github.io/BERTopic/api/ctfidf.html.

Grootendorst, Maarten. 2025c. *Outlier reduction*. https://maartengr.github.io/BERTopic/getting_started/outlier_reduction/outlier_reduction.html.

Hary, Benjamin. 2011. Religiolect. In Deborah D. Moore, Anita Norich & Joshua L. Miller (eds.), *Critical terms in Jewish language studies*, 43–53. Ann Arbour: Frankel Institute for Advanced Judaic Studies.

KB Lab. 2025. *BERTopic workshop: analyzing Swedish parliamentary motions*. https://colab.research.google.com/drive/10kB3wfoHSfZE48vEKmznIw-ff36uR8gs?usp=sharing.

Lavin, Matthew J. 2019. Analysing documents with TF-IDF. *Programming historian* 8. DOI: 10.46430/phen0082.

Meijer, Roel (ed.). 2009. *Global Salafism: Islam's new religious movement*. London: Hurst & Company.

Metz, Cade, Cecilia Kang, Sheera Frenkel, Stuart A. Thompson & Nico Grant. 2024. How tech giants cut corners to harvest data for A.I. *The New York*

Times. https://www.nytimes.com/2024/04/06/technology/tech-giants-harvest-data-artificial-intelligence.html.

O'Reilly, Tim. 2007. What is Web 2.0: Design patterns and business models for the next generation of software. *Communications & Strategies* (65). 17–37.

Olsson, Susanne. 2020. Advising and warning the people: Swedish Salafis on violence, renunciation and life in the suburbs. In Simon Stjernholm & Elisabeth Özdalga (eds.), *Muslim preaching in the Middle East and beyond: Historical and contemporary case studies*. 155–172. Edinburgh: Edinburgh University Press.

Olsson, Susanne, Simon Sorgenfrei & Jonas Svensson. 2022. Puritan Salafis in a liberal democratic context. In Magnus Ranstorp, Linda Ahlerup & Filip Ahlin (eds.), *Salafi-jihadism and digital media*, 92–112. Routledge. DOI: 10.4324/9781003261315.

OpenAI. 2025. *Processing and narrating a video with GPT's visual capabilities and the TTS API*. https://cookbook.openai.com/examples/gpt_with_vision_for_video_understanding.

Potter, Jonathan & Chloe Shaw. 2018. The virtues of naturalistic data. In Uwe Flick (ed.), *The SAGE handbook of qualitative data collection*, 182–199. London: SAGE.

Ranstorp, Magnus, Filip Ahlin, Peder Hyllengren & Magnus Normark. 2019. *Between Salafism and Salafi-Jihadism: Influence and challenges for Swedish society*. Swedish Defence University, Center for Asymmetric Threat Studies. URN: urn:nbn:se:fhs:diva-8534.

Sorgenfrei, Simon. 2021a. "Perhaps we see it in negative terms, but, ultimately, it is positive.": The responses of Swedish Salafis to COVID-19. *Tidskrift for Islamforskning* 15(2). 40–62. DOI: 10.7146/tifo.v15i2.125959.

Sorgenfrei, Simon. 2021b. Crowdfunding Salafism. Crowdfunding as a Salafi missionising method. *Religions* 12(3). 209. DOI: 10.3390/rel12030209.

Svensson, Jonas. 2021. Mönster i svensk YouTube-salafism. Datorassisterad metadataanalys som urvalsmetod i religionsvetenskapliga studier av videomaterial online. *Religionsvidenskabelig tidskrift* 73. 60–83. DOI: 10.7146/rt.vi73.129550.

Svensson, Jonas. 2022. Mönster i rappakaljan: Ett test med datorassisterad, urvalsgenererande fjärrläsning av automatiskt transkriberade salafistiska predikningar på svenska. *Tidsskrift for islamforskning* 16(1). 134–155. DOI: 10.7146/tifo.v16i1.132561.

Svensson, Jonas. In press. "Allah Says Beat Them". An Analysis of Non-Muslim Islams Justifying Domestic Violence. In J. Petersen & A. Ackfeldt (eds.), *Non-muslim islams*. Edinburgh: Edinburgh University Press.

Svensson, Jonas. Forthcoming. Muhammed och monoteisternas fiende. *Chaos*.

Turan, Salih Can, Kazım Yildiz & Büşra Büyüktanir. 2024. Comparison of LDA, NMF and BERTopic topic modeling techniques on amazon product review dataset: A case study. In Fausto Pedro García Márquez, Akhtar Jamil, Isaac Segovia Ramirez, Süleyman Eken & Alaa Ali Hameed (eds.), *Computing, Internet of Things and data analytics*, 23–31. Cham: Springer. DOI: 10.1007/978-3-031-53717-2_3.

Appendix 1 *LLM generated representations*

	Label		Label
0	Djuz 'Amma and Janaza (Funeral)	14	Proofs for Muhammad's Prophethood and God's Existence
1	Quran Recitations	15	Ramadan: Preparation, Practice, and Consequences
2	Quran Recitations of Various Surahs	16	Fasting and Ramadan: Rules, Questions, and Advice
3	Islamic Jurisprudence: Basic Principles	17	Sunnah Prayers and Prayer Rituals
4	Paths to Paradise	18	Chapter on Prayer and the Virtue of the Friday Prayer
5	Sahih al-Bukhari and the State of the Heart	19	Tafsir of Surah Yusuf and Djuz 'Amma
6	Tafsir of Surah al-A'raf	20	Tafsir of Various Surahs and Djuz 'Amma
7	Brief Inspirational Messages #Shorts	21	The Importance of Prayer in Islam
8	Reminders and Teachings on Islamic Virtues and Practice	22	Quran Recitation and Memorization
9	Recitations of Surahs	23	Quran Recitations – Various Surahs
10	Marriage and Family Bonds in Islam	24	Encouraging Good, Forbidding Evil, and Leadership
11	Eid Mubarak, Friday Prayer, and Religious Advice from Shaykh Abdulwadod Frank	25	The Quran as the Revelation of God
12	Signs of the Day of Judgment and Belief in Fate	26	The Importance and Practice of Prayer
13	Quran Recitations and Reflections on Content		

(*Continues on next page*)

(*Continued from previous page*)

	Label
27	Prayer: Focus, Engagement, and Direction
28	Prayer in Islam and Sunnah Prayers
29	Chapter on Prayer
30	Death and the Afterlife
31	Aqidah, Hadith Jibril, and Supplication/Invocation
32	Zakat, Sadaqah, and Islamic Rules
33	Bible Analyses and the Islamic View of Jesus
34	The Prophet's Sunnah and Hadith
35	Biographies of Imams and Theology
36	Advice and Reminders with Moosa Assal
37	Tawhid, the Prophet's Status, and the Quran
38	Stories of the Companions and the Prophets
39	The Life and Qualities of the Prophet Muhammad
40	The Life and Teachings of the Prophet Muhammad ﷺ
41	Sins, Their Consequences, and Forgiveness
42	The Stories and Qualities of the Prophets
43	Stories of the Prophets and Sacred Figures
44	Tafsir of Surah Al-Anbiya and Djuz Tabarak, along with Quran Memorization
45	Sahih al-Bukhari, Biographies, and Reflections on Surahs and Hadith
46	Explanations of Surahs, Hadith Literature, and Islamic Books
47	Virtues and Dawah
49	The Evil Consequences of Sins
50	Salam, Cleanliness, and Hygiene in Islam
51	Reminders on Faith and the Day of Judgment
52	Islamic Values and Rights
53	Motivation and Reminders of Good Deeds
54	Tarawih Reminders and Tawhid
55	Wudhu (Ablution) and Umrah
56	Respect for Parents, Anger, and Social Problems
57	Islamic Teaching and Reflections
58	Arguments and Proofs for God's Existence
59	Islamic Trade and Investments
60	Economy and Financial Ethics in Islam
61	Goodness, Parents, and the Prophet's Rights
62	Satan's Plans and Evil
63	Wisdom and Knowledge
64	Hajj, Dhikr, and Hadith
65	Critique of Atheism and Arguments Against Atheism
66	Personal Development and Relationships
67	Life Guidance and Advice in Islam
68	Faith, Doubt, and Aqidah
69	Faith, Sincerity, and Good Deeds
70	The Role and Status of Men and Women in Islam
71	Muslim Education and Important Lessons
72	Ethics, Morality, and Character in Islam
73	Faith and Love of Allah
74	200 Questions on Islamic Theology
75	Virtues of the Quran and Reflections
76	The Quran: Reflections and Analyses
77	Islam, Extremism, and Terrorism

(*Continues on next page*)

(*Continued from previous page*)

	Label
78	Muslim Behavior, Unity, and Handling of Trials
79	Islamic Theology and Practice
80	200 Questions on Islamic Theology
81	Political Ideologies and Beliefs
82	Islam and Muslims: Fajr Reminders and the Importance of Religion
83	Islamic Theology and Evidential Arguments
84	Guidance and Advice for Prospective Muslims
85	Theology, Science, and Handling Doubts in Islam
86	Islamic Knowledge and Worship

Corresponding author

Jonas Svensson
Department of Cultural Sciences
Linnæus University
jonas.svensson@lnu.se