2R Project design protocol for assignment 2

OBJECTIVES

The current project aims to investigate how attitudes of trust and distrust toward artificial intelligence (AI) are expressed in public YouTube videos. Specifically, it intents to identify overall trust attitudes towards AI, sentiment polarity and thematic patterns that characterise these attitudes, using a combination of computational and language-based approaches.

DATA COLLECTION AND ANALYSIS PROTOCOL

The current project follows these steps:

1. Defines keyword synonyms for the two groups (trust vs distrust) using the Merriam-Webster Thesaurus to minimise keyword bias. Initially, I tested the Python lexicon libraries such as WordNet and the NRC-Emotion Lexicon, but their terms didn't seem as appropriate for the purpose of this project – not as relevant results during a pilot test.
2. Retrieves YouTube video data through YouTube's API, and store them locally in a json format for easier retrieval at a later point, without having to repeat API calls – saves on efficiency and costs. Opted in using YouTube's API instead of scraping to align more closely with the platform's policies.
3. Transcribes the audio from the matching videos using OpenAI's whisper model through their API, and store the transcripts locally in .txt files.
4. Cleans and preprocesses the transcripts, including decoupling contractions, removing punctuation and stopwords (for NLTK but not LLMs and DistilBERT to not detract from meaning). Applies tokenisation for NLTK only, because the rest of the models have an internal / built-in tokenisation process.
5. Identifies overall attitudes by counting word frequencies across all transcripts using a bag-of-words document-feature matrix (DFM) and visualising a set of highest-frequency words using a bar plot and a word cloud.
6. Estimates trust and sentiment polarity (positive/trustworthy, negative/distrustful, neutral) through the use of sentiment models, such as NLTK Vader, DistilBERT sentiment pipeline, and LLM-based classifiers. Their outputs will be displayed in the console as well as plotted. I decided to use multiple different models for sentiment analysis as another way of validating the results by comparing findings through model pairings and calculating their agreement percentage in the absence of true labels.
7. Identifies the most dominant thematic patterns with their matching set of keywords with FASTopic for topic modelling, and results will be plotted, and mentioned later within this report too. I will first extract the top 20 topics and go sequentially down to 5 topics to observe any variations and patterns.

ETHICAL ASSESSMENT

Data format and information sources

This project will collect publicly available data from YouTube using their official API (API key obtained but not shared with this submission) and respecting their access rules. YouTube content will include transcribed text extracted from the audio files of publicly accessible videos through OpenAI's Whisper model. Although the data is public, raw data won't be shared to comply with the platforms' T&Cs and to minimise any identifiable information about the users when uploading the project files on GitHub.

Models to be used
OpenAI's Whisper (speech-to-text) model will be used to transcribe YouTube's audio content, using OpenAI's API key. Using APIs means that the data I gather and analyse are not fully handled locally, and automated transcription, such as the Whisper model, might not perform as well under potentially different accents and styles of speech, reflecting their training bias.

Morever, four classifiers will be used for sentiment/trust labelling: NLTK Vader (SentimentIntensityAnalyzer), DistilBERT (distilbert-base-uncased-finetuned-sst-2-english) with a sentiment pipeline, and two LLM-based classifiers (gpt-4o-mini, and gemini-2.0-flash-lite). Their sentiment labels output (positive, negative, neutral) will be mapped to trust, distrust, neutral, and will be compared through a model-pair agreement percentage and word-error-rate for robustness/validation.

Project dissemination
The Python analysis source code alongside any results reports and Sphinx documentation will be shared on GitHub, supporting open science and reproducibility. The Sphinx documentation, although not requested in the assignment, I always like to include it as it helps explaining further the code structure and functions purpose. Similarly, I will also include a requirements textfile for the Python libraries used for reproducibility purposes, and a README file to briefly introduce the project alongside this project protocol document. I know I might have gone a bit overboard (I also included stylistic message functions, etc) but I enjoyed it, been nerding out!

The raw data (e.g., original Youtube content stored as json files, and raw transcription text) will not be shared due to their potentially sensitive and identifiable content, and the platforms' T&Cs. Instead, a sample dataset (cleaned_transcript_records.csv inside the data folder) will be provided in order to allow for testing of the rest of the analysis pipeline in the project's codebase. Nothing to change in the code for that, the checks have already been set to True in the main.py for:

json_files_exist = True
transcriptions_completed = True
cleaned_data_csv_exists = True

So you can run the main.py (hopefully) without any hiccups to check the analyses. Remember first to add your own API keys in the config.py custom_packages module.

Bias potential
In terms of retrieval bias through YouTube search query terms, I tried to minimise that by using the established thesaurus by Merriam-Webster and documenting the exact terms used ("trustworthy" and "untrustworthy") to find more representative synonyms, and query date (August 2025) in the main.py file. WordNet and NRC were piloted but then dropped as they didn't seem to be as relevant in their keywords.

There is potentially some transcription/algorithmic bias, considering that model performance, in this case OpenAI's Whisper, depends on what data the model has been trained on. A subset of transcripts will be manually checked, and any unusable files will be excluded. The same goes for the sentiment analysis models, which is why I do a model-pair comparison on their trust label agreement percentage, and in terms of LLMs, I retrieved and manually compared their results across different time-points.

SYSTEM VALIDATION

Search query keywords are compared across repeated runs. Manual validation happens right after the YouTube video data are stored as json files, by checking if the titles, and a subset of the video content is representative of the search query terms. Unrelated videos may be manually logged separately in one of the raw files just to keep track.

Similarly, a subset of transcriptions are checked manually and compare them with the actual video content. Invalid ones (e.g., missing language pack, missing segments, etc) are logged and excluded.

Since I don't have true labels for the sentiment/trust analysis, I simply do a model-pair comparison in terms of agreement percentage. Additionally, for LLMs I run them at different datetime points and compare again.

MY EXPORTED MODEL COMPARISON RESULTS, AND TOPICS

The exported plots can be found in the "figures" folder.

Model	Trust label count	Distrust label count	Neutral label count
NLTK Vader	32	4	-
DistilBERT	12	24	-
GPT-4o-mini	10	21	5
Gemini-2.0-flash-lite	7	17	12

Reference model 1	Compared model 2	Agreement rate (%)
NLTK Vader	DistilBERT	44.44
NLTK Vader	GPT-4o-mini	33.33
NLTK Vader	Gemini-2.0-flash-lite	27.78
DistilBERT	GPT-4o-mini	55.56
DistilBERT	Gemini-2.0-flash-lite	47.22
GPT-4o-mini	Gemini-2.0-flash-lite	80.56

DOMINANT TOPICS – Shortest version (5 topics):

Topic 1: augmented, businesses, segment, hate, individual, moderation, miss, speakers, aversion, precision, implicit, vision, spam, financial, insights

Topic 2: disinformation, score, win, diplomacy, cicero, deceive, gemini, similarity, deepfake, deepfakes, ninja, article, opus, capable, plagiarism

Topic 3: echo, roger, droid, zombie, hardcase, apollo, fives, ugh, hell, ethics, eliza, newsworthy, wait, dead, captain

Topic 4: people, know, going, right, lot, think, things, kind, actually, way, want, like, news, said, sure

Topic 5: kiro, pausing, cybersecurity, pause, anthropic, institutions, agent, attackers, methods, objectives, similarly, app, goals, students, cognitive

DOMINANT TOPICS – Short version (10 topics):

Topic 1: quote, players, humans, playing, chatbot, question, point, deception, scientific, bubble, companies, people, big, started, workers

Topic 2: financial, segment, vision, pausing, pause, cybersecurity, security, personalized, attackers, traffic, exciting, autonomous, healthcare, threats, transforming

Topic 3: agency, keto, grief, podcast, gloves, worried, net, maltodextrin, carbs, medium, shirt, sort, element, love, nature

Topic 4: augmented, similarity, ninja, ghost, plagiarism, humanized, score, turnitin, humanizer, stealth, reduced, reduce, cognitive, adjust, creativity

Topic 5: hidden, apollo, intentionally, anthropic, developers, intentional, objectives, methods, goals, deceiving, internal, marks, evaluation, hide, responsibly

Topic 6: businesses, hate, moderation, implicit, aversion, gender, algorithm, bias, labelled, women, differences, participants, biases, posts, employees

Topic 7: roger, droid, property, zombie, hardcase, condo, ugh, fives, wait, gen, hell, shit, ship, buildings, fuck

Topic 8: newsworthy, cigarettes, packages, fusion, poll, movie, saudi, ford, fbi, esports, spark, devices, says, existing, apparently

Topic 9: kiro, coalition, betrayals, sneeze, app, opus, alliances, snowden, nsa, spec, cursor, purpose, chance, moves, plan

Topic 10: disinformation, deepfakes, deepfake, campaigns, manipulation, chambers, efforts, spam, misinformation, politician, narratives, languages, audio, accounts, biden

DOMINANT TOPICS – Longest version (20 topics):

Topic 1: misinformation, image, matter, listen, media, congress, photo, reporting, deepfake, biden, voice, policy, states, chatbots, social

Topic 2: apollo, aversion, businesses, employees, feedback, unsupervised, implicit, responsibly, movers, frontier, kids, insights, overcome, gut, somatic

Topic 3: keto, gloves, carbs, maltodextrin, recipes, element, tortilla, recipe, ingredients, net, pictures, hands, nutritional, flavor, sort

Topic 4: similarity, ninja, plagiarism, humanized, ghost, score, turnitin, humanizer, stealth, reduced, documents, stealthwriter, humanize, interface, reduce

Topic 5: roger, droid, hardcase, zombie, ugh, fives, hell, captain, wait, fuck, shit, gross, ass, echo, sex

Topic 6: moderation, hate, adjust, differences, existing, state, energy, participants, mainstream, creativity, renewable, attributes, sci, robots, individual

Topic 7: scientific, scientist, spam, agency, messages, speakers, agentic, blocking, whatsapp, gmail, jones, blocks, polyglot, joy, languages

Topic 8: win, diplomacy, gemini, cicero, opus, game, powers, strength, pro, claude, play, meta, moves, coalition, allies

Topic 9: augmented, grief, cognitive, students, anxiety, class, student, mental, miss, loop, closure, sudden, brains, independently, autonomy

Topic 10: segment, vision, sneeze, cursor, vehicles, doctors, mode, healthcare, entertainment, facial, precision, apps, transforming, identify, list

Topic 11: jobs, bubble, dot, com, hype, workers, biggest, companies, amazon, coldfusion, episode, cycle, aura, layoffs, washing

Topic 12: attackers, cloud, threats, logs, defenses, attacker, security, cybersecurity, network, response, autonomous, anomalies, alerts, intervention, traffic

Topic 13: objectives, methods, anthropic, hidden, unintended, marks, evan, accountability, explained, hubinger, goals, researchers, hide, proper, samuel

Topic 14: poll, skepticism, respondents, majority, favor, charge, news, americans, politics, china, interesting, article, india, podcast, expressed

Topic 15: property, condo, buildings, fusion, growth, active, management, market, infrastructure, tribe, visibility, footprint, managers, rental, aaron

Topic 16: gen, gender, women, white, shocking, labelled, men, chat, generation, society, demographic, soap, tick, samsungs, embeddings

Topic 17: newsworthy, cigarettes, packages, fbi, saudi, esports, snowden, nsa, ford, says, critics, nakasone, vaping, director, vote

Topic 18: pausing, pause, financial, pushing, narrative, awareness, finance, enforce, opportunity, nations, tailored, regulatory, advocating, measures, institutions

Topic 19: kiro, app, conversation, ask, eliza, perry, chatbot, program, code, design, step, paranoid, external, spec, prompts

Topic 20: disinformation, deepfakes, campaigns, misleading, false, deceive, manipulate, chambers, manipulation, spread, narratives, harder, audiences, beliefs, alphastar

The data visualisation plots can be found in the “figures” folder, as they were quite a lot to just dump here.

FUTURE DIRECTIONS AND LIMITATIONS

- In the future, I could further polish the search terms used and expand on them to find potential commonalities across platforms.
- I could have manually transcribed a number of YouTube videos to use as the ground truth when evaluating OpenAI's Whisper model.
- The lack of ground truth labels for sentiment/trust analysis, prevents me from using the classic sklearn.metrics such as F1, accuracy score, etc. In the future, I could potentially, manually create a small human-annotated set to train and test against.
- Currently limited to just YouTube video content. In the future I could examine YouTube comments, and expand on other platforms such as Reddit, news websites, etc for broader applicability.
- For potentially richer topic insights and validation, I could compare FASTopic with BERTopic, and do some fine-tuning.