MASTER DEGREE IN
CYBERSECURITY

DIPARTIMENTO
MATEMATICA

LEVERAGING GDPR DATA PORTABILITY:

A USER-CENTRIC APPROACH TO SOCIAL MEDIA ANALYTICS

Prof. Mauro Conti
Prof. Fabio De Gaspari
PhD Luca Pajola

Michele Gusella
2122861

BACKGROUND

Online Social Networks

Number of social media identities

5.66 billion

Year-wise growth rate

+4.8%
+259 million

Weekly activity time

18H 36M

Data updated to October 2025
Source: https://datareportal.com/social-media-users

BACKGROUND

GDPR and Data portability

Art. 20 GDPR

Right to data portability

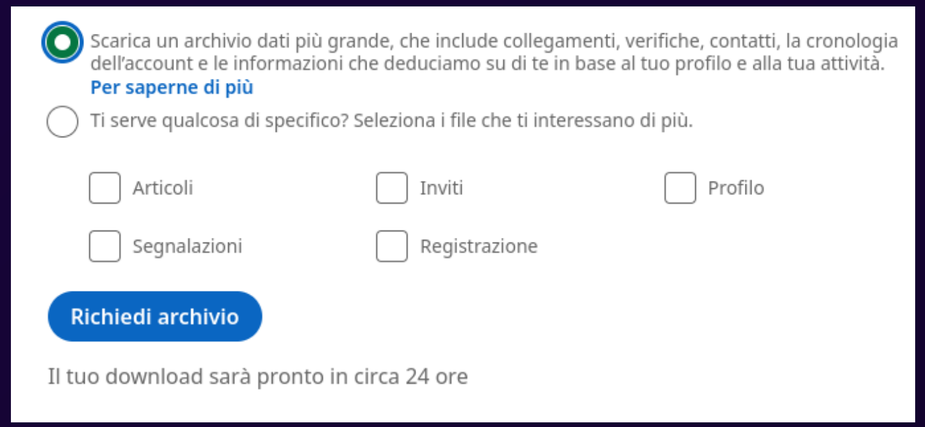
The data subject shall have the right to receive the personal data concerning him or her, which he or she has provided to a controller, in a **structured**, **commonly used** and **machine-readable format** and have the right to transmit those data to another controller without hindrance from the controller to which the personal data have been provided [...]

Data Download Packages (DDPs)

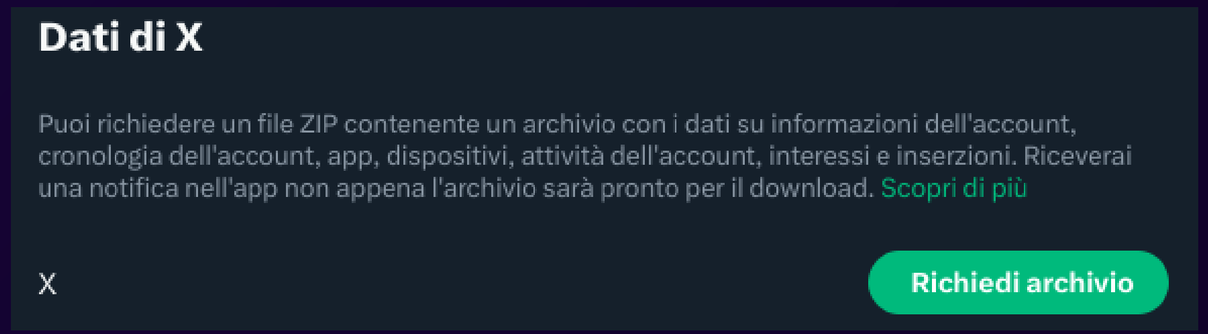
Source: https://gdpr-info.eu/art-20-gdpr/

BACKGROUND

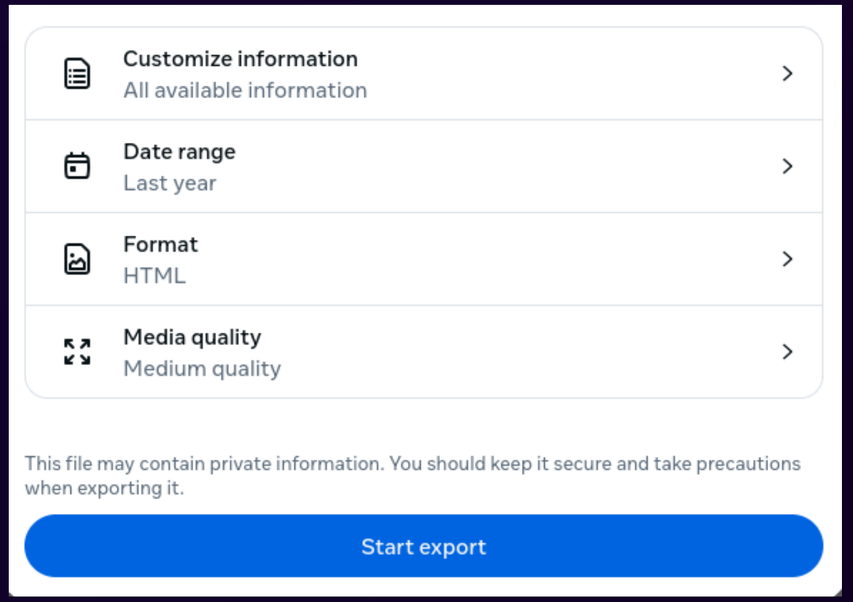
Data Download Packages



LinkedIn DDP request page



Twitter DDP request page



Instagram DDP request page

BACKGROUND

DDPs

Problems:

- Raw data (JSON, CSV)
- Heterogeneous formats
- Hard to interpret


```
{
  "impressions_history_app_message": [
    {
      "string_map_data": {
        "In-app message name": {
          "value": "In-app Message 687152328703306"
        },
        "Tipo di clic": {
          "value": "Secondary Click"
        },
        "Numero": {
          "value": "1"
        }
      }
    },
```



```
{
  "Comment": {
    "Comments": {
      "App": 1,
      "CommentsList": [ ... ]
    }
  },
  "Direct Message": {
    "Direct Messages": {
      "ChatHistory": {}
    }
  },
  "Income+ Wallet": {
    "Transaction History": {
      "TransactionsList": null
    }
  },
```



```
Member Age,Buyer Groups,Company Names,Company Names,Company Follower of,Company Names,Company Category,Company Size,Degrees,degreeClass,Member Schools,Company Growth Rate,Fields of Study,Company Connections,Function By Size,Job Functions,Member Gender,Graduation Year,Member Groups,Company Industries,Member Interests,Interface Locales,interfaceLocale,Member Traits,Profile Locations,Company Revenue,Job Seniorities,Member Skills,Job Titles,Job Titles,Job Titles,Years of Experience
```

BACKGROUND

Research Gap

Existing work:

- API/Scraping data collection
- Large scale analysis

Missing:

- Analysis of personal DDPs
- User-centric perspective

Beyond Twitter: Exploring Alternative API Sources for Social Media Analytics

Alina Campan[a] and Noah Holtke
School of Computing and Analytics, Northern Kentucky University, Nunn Drive, Highland Heights, U.S.A.

Open Access Article

Use and Abuse of Personal Information, Part I: Design of a Scalable OSINT Collection Engine

by Elliott Rheault, Mary Nerayo, Jaden Leonard, Jack Kolenbrander, Christopher Henshaw, Madison Boswell and Alan J. Michaels *

Social Media Web Scraping using Social Media Developers API and Regex

Lusiana Citra Dewi [a] ✉, Meiliana [a], Alvin Chandra [a]

PROPOSED SOLUTION

Contributions

- **User-centric framework**
  → Directly analyzes personal data

- **Unified pipeline**
  → Handles multiple platforms, formats, and data types

- **Integration of multiple analysis modules**
  → Topic modeling, sentiment analysis, ...

- **Clear and interactive output**
  → Designed for non-expert users

PROPOSED SOLUTION

Preliminary DDP analysis

- Download procedures
- Data restrictions
- Classification of files
- Ontology

Macro-Category	Micro-Categories
User Profile	Personal data Personal interests Settings & Preferences
Social Graph	Connections Following/Follower Unidirectional
Content Creation	Publications Activity
Social Relationship	Feedbacks Comments Tags General Interactions
Messaging	History Notifications
Marketplace	Orders Payments Subscriptions
Algorithm & Browsing	Searches Content consumed Saved elements
Profiling/Advertising	Interests Interactions
Telemetry	Device metrics Navigation & Interface

Ontology macro-categories and micro-categories

PROPOSED SOLUTION

Design goals

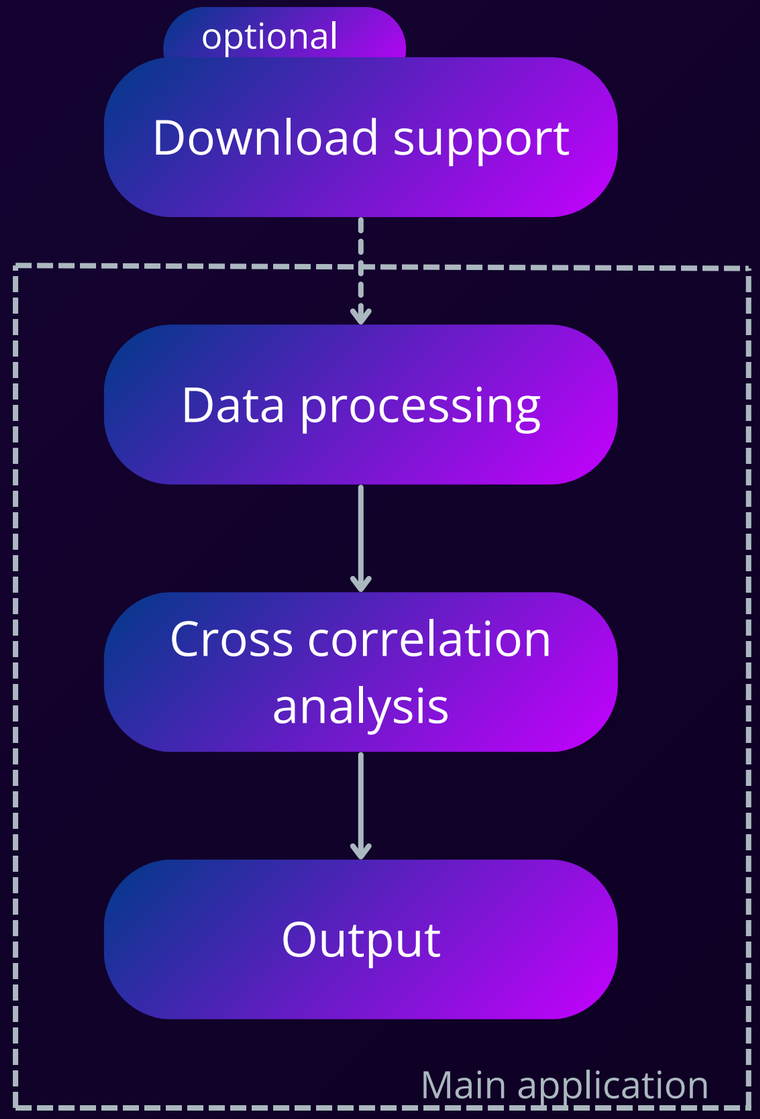
- User-centered system
- Handle multiple DDPs
- Privacy aware
- Modular

Main application structure

parsers/

analyzers/

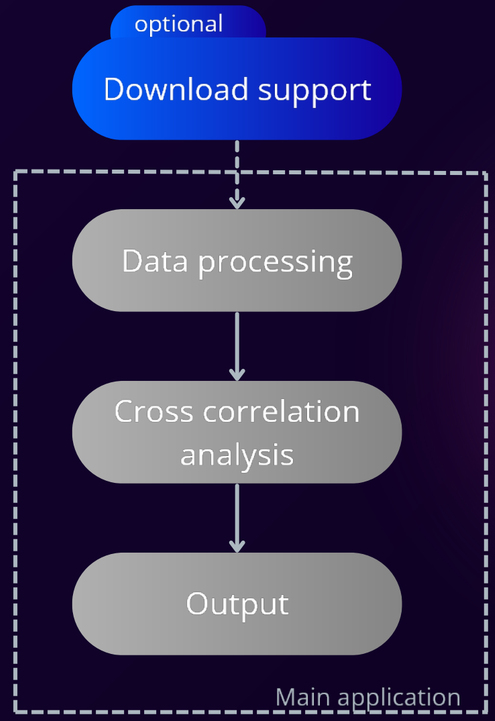
app.py

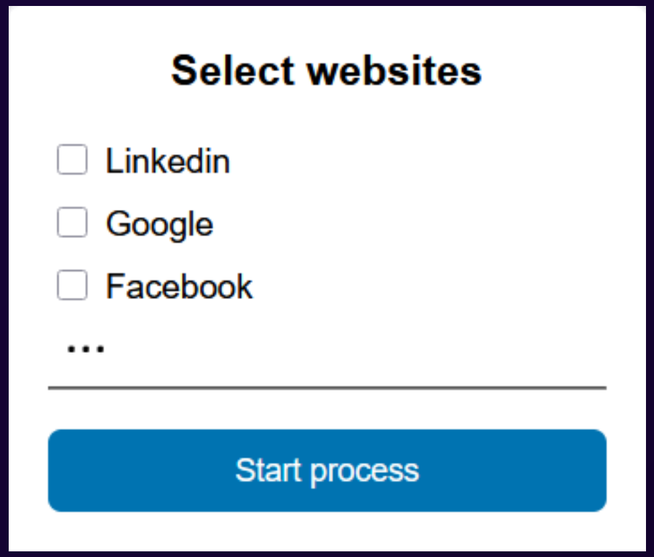


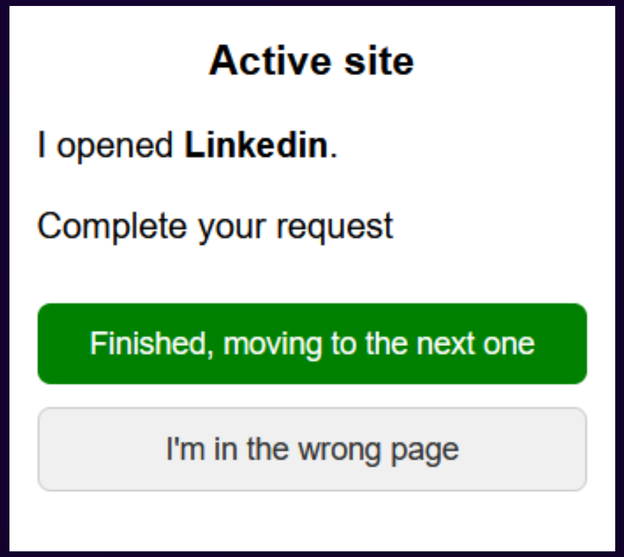
PROPOSED SOLUTION

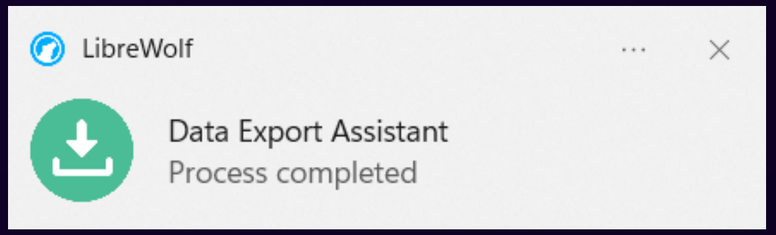
DDP download support

Guide users to DDP download in a direct and efficient way.





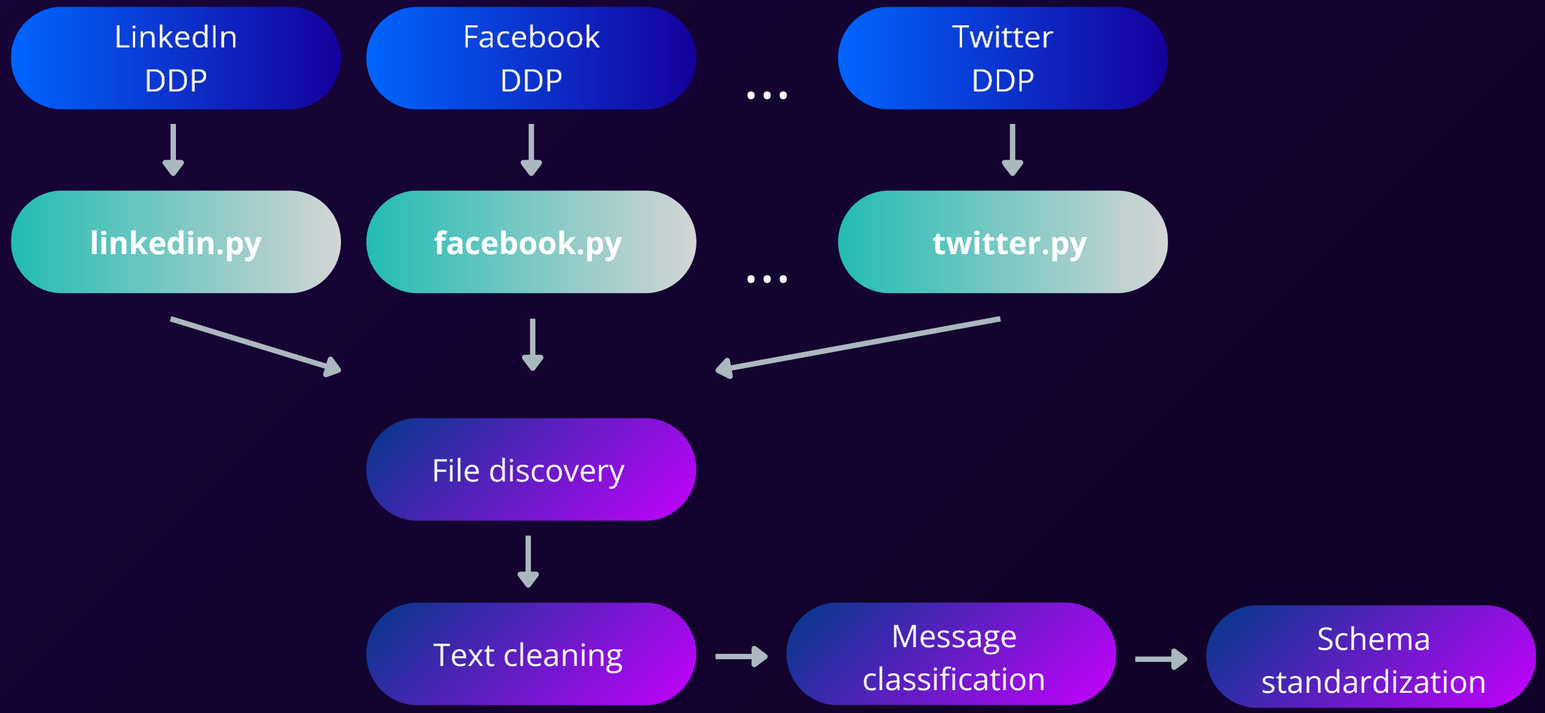


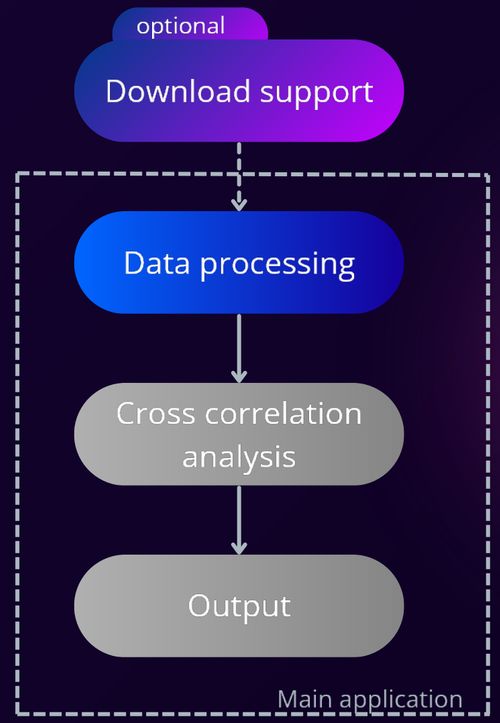


PROPOSED SOLUTION

Data processing

Data is processed by each parser.

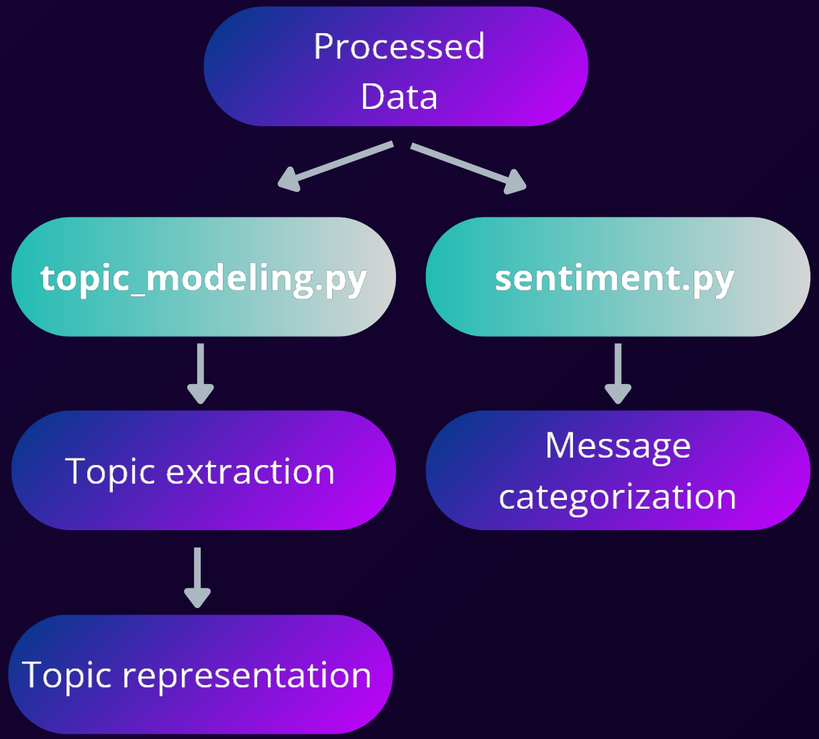


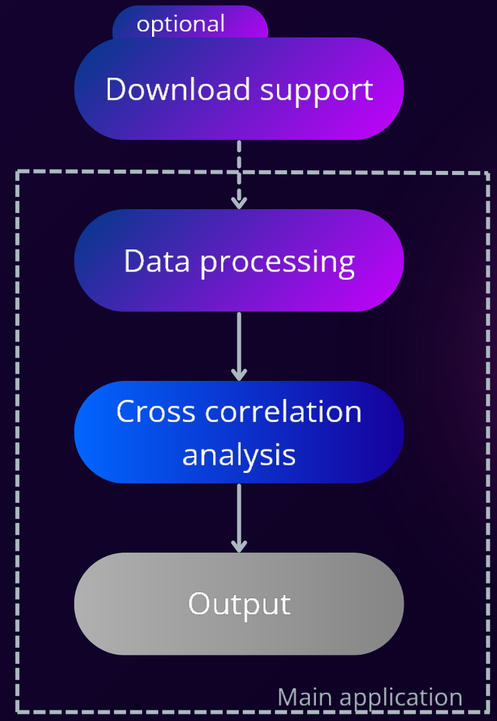


PROPOSED SOLUTION

Data analysis

There are two main analysis modules.





PROPOSED SOLUTION

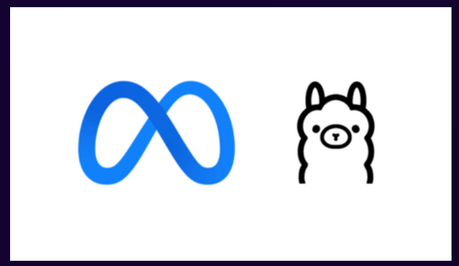
Data analysis

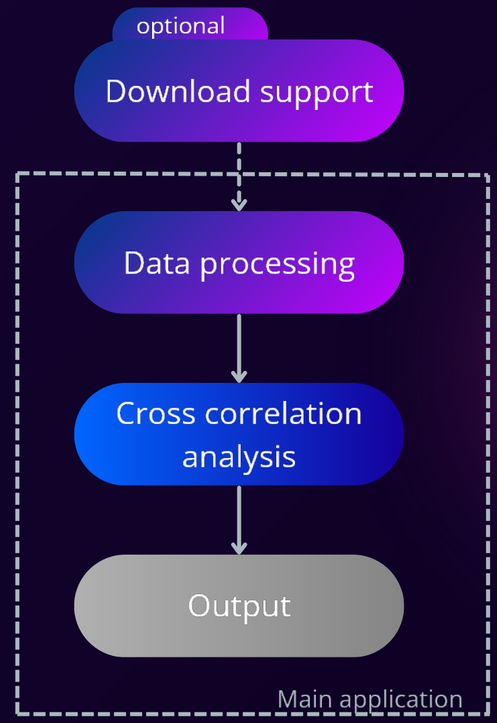
Topic modeling is performed using BERTopic.
Hyperparameters:

- minimum cluster size
- maximum number of topics

Topic representation can be done using:

- llama2 LLM
- KeyBERT



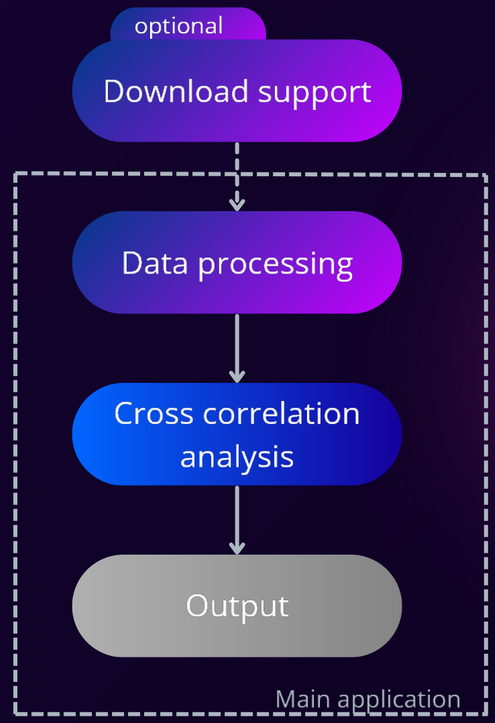


PROPOSED SOLUTION

Data analysis

Sentiment analysis is performed using:

- Transformer model
- VADER (lightweight rule-based algorithm)

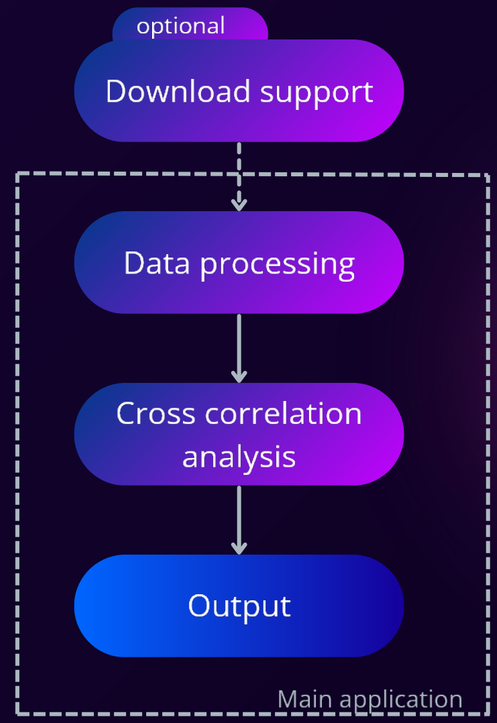


Representative examples

Positive	Neutral	Negative
Però complimenti l'hai fatta bene, mi piace tantissimo	Farei	Cattivo
Bellissima sta foto!	anche i miei amici me lo hanno mandato	Si era comportata male
Capodanno quest'anno sarà stupendo 😍	Nugget	Che tristezza

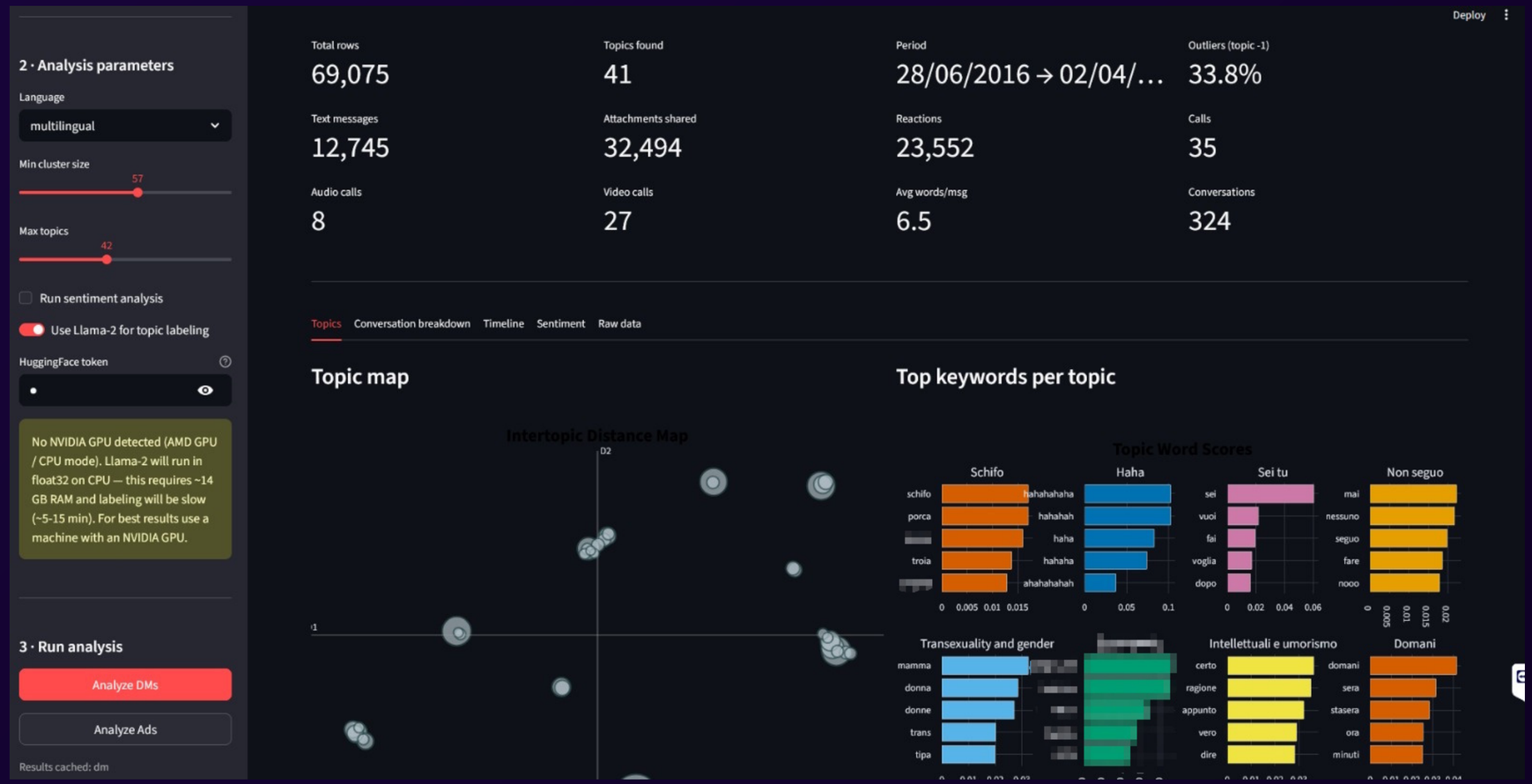
PROPOSED SOLUTION

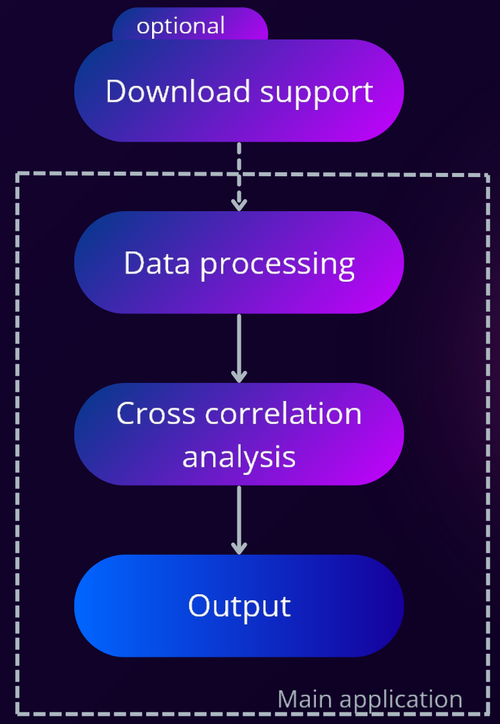
Output visualization



PROPOSED SOLUTION

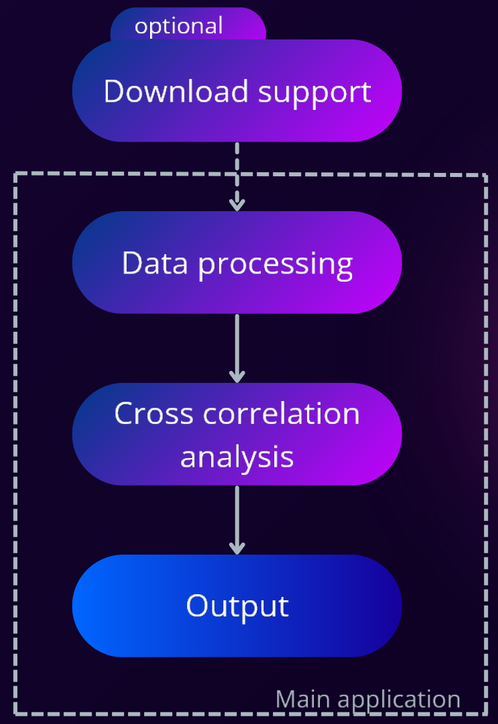
Output visualization





PROPOSED SOLUTION

Output visualization

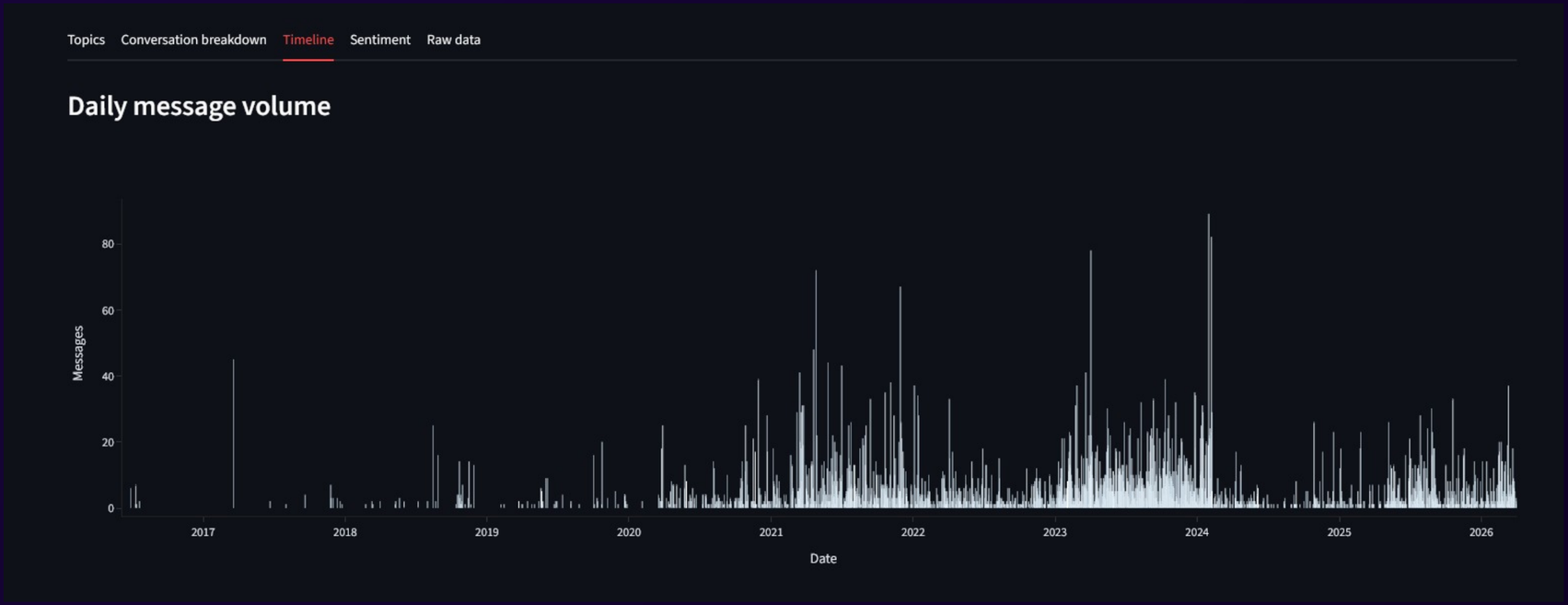


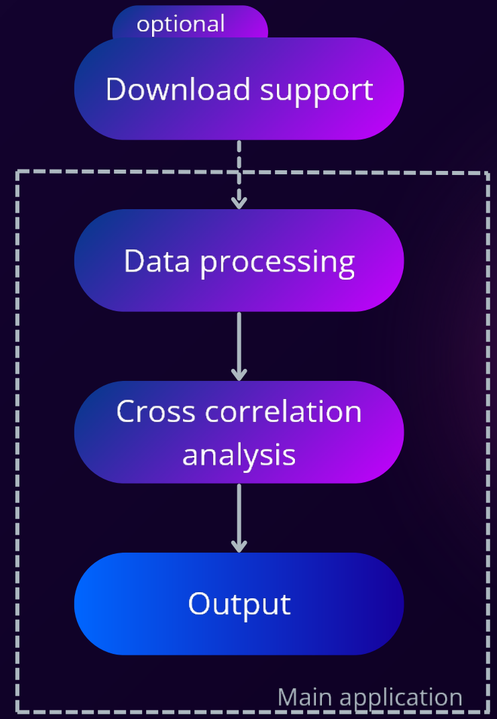
Topic detail

Topic	Count	Name	Representation	CustomName
19	80	19_magazine_fashion_versace_collezione versace	magazine, fashion, versace, collezione versace, gucci, inverno 2025, nuova collez	Versace Autunno-Inverno 2025
20	79	20_adventzagreb_adventzagreb coming_coming_zagreb	adventzagreb, adventzagreb coming, coming, zagreb, zagreb obviously, obviously a	Zagreb Christmas
21	68	21_go ad_includes access_free premium_supergrok	go ad, includes access, free premium, supergrok, access supergrok, ad free, prem	Ad-free X with Premium+
22	66	22_asl_dai piccoli_asl roma_piccoli	asl, dai piccoli, asl roma, piccoli, piccoli grandi, grandi proteggi, gratuiti, gratuit	Vaccini gratuiti ASL Roma
23	65	23_tile_game_tile explorer_explorer	tile, game, tile explorer, explorer, free game, tile free, free, tile game, freefon	FreeForm Challenge
24	63	24_music_sony music_sony_clips	music, sony music, sony, clips, universal music, universal, australia, spotify	Podcast to clip
25	61	25_omd_tumor_global_chile	omd, tumor, global, chile, gbp, clients, gedi digital, geotech, clients gwagon	Brain Tumor Study
26	60	26_guess_why guess_guess know_know why	guess, why guess, guess know, know why, know, why, ciò, sei, ciò senti, a	Knowledge vs Guessing
27	53	27_server_data_gt_99	server, data, gt, 99, advance, server advance, servizi, walrus, sistema, adva	Server Advance
28	53	28_daily_afterwards_daily payouts_funded daily	daily, afterwards, daily payouts, funded daily, after getting, day after, payouts aft	Daily Payouts

PROPOSED SOLUTION

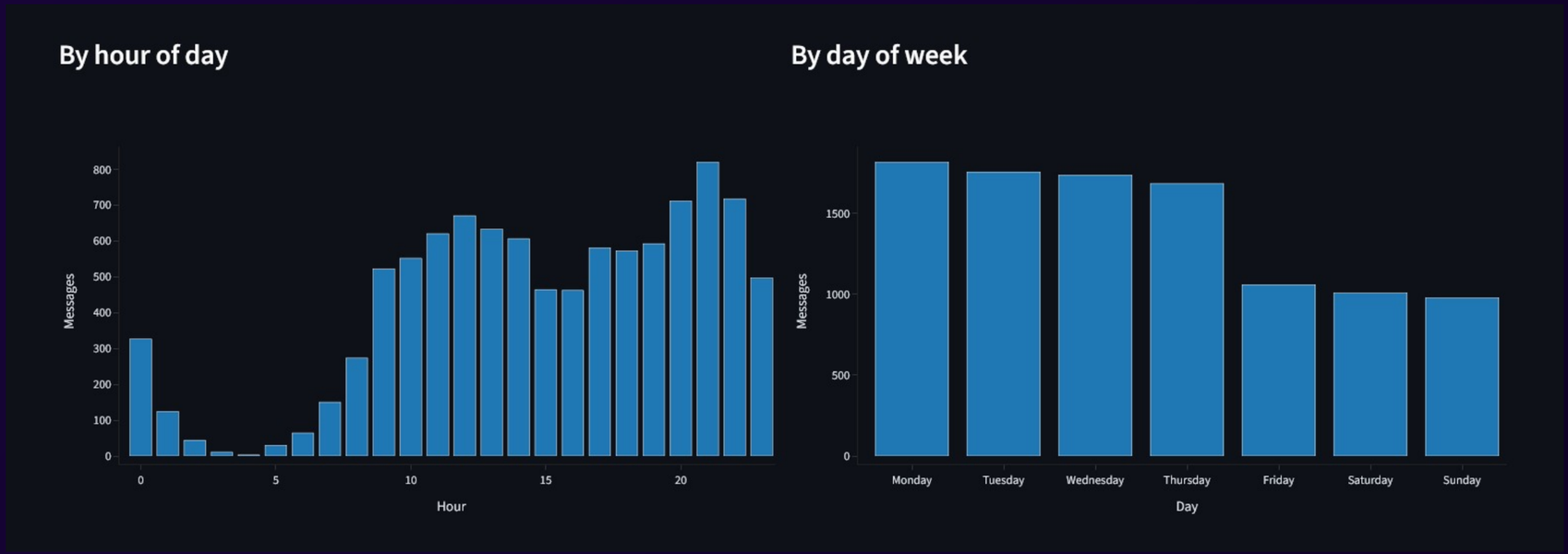
Output visualization

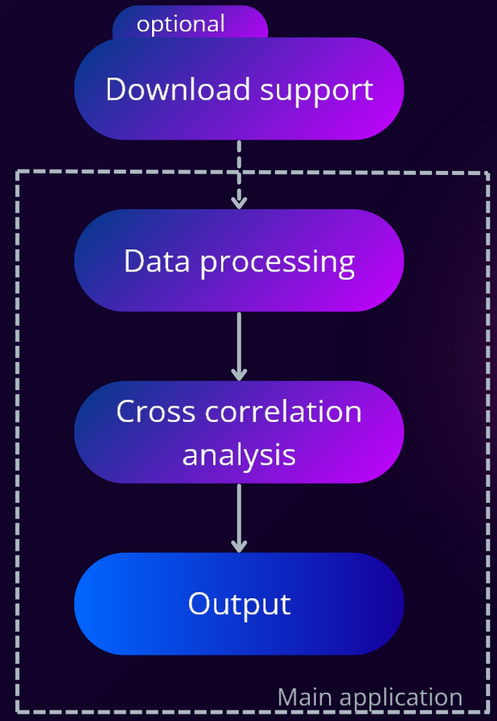




PROPOSED SOLUTION

Output visualization





RESULTS

Quantitative analysis

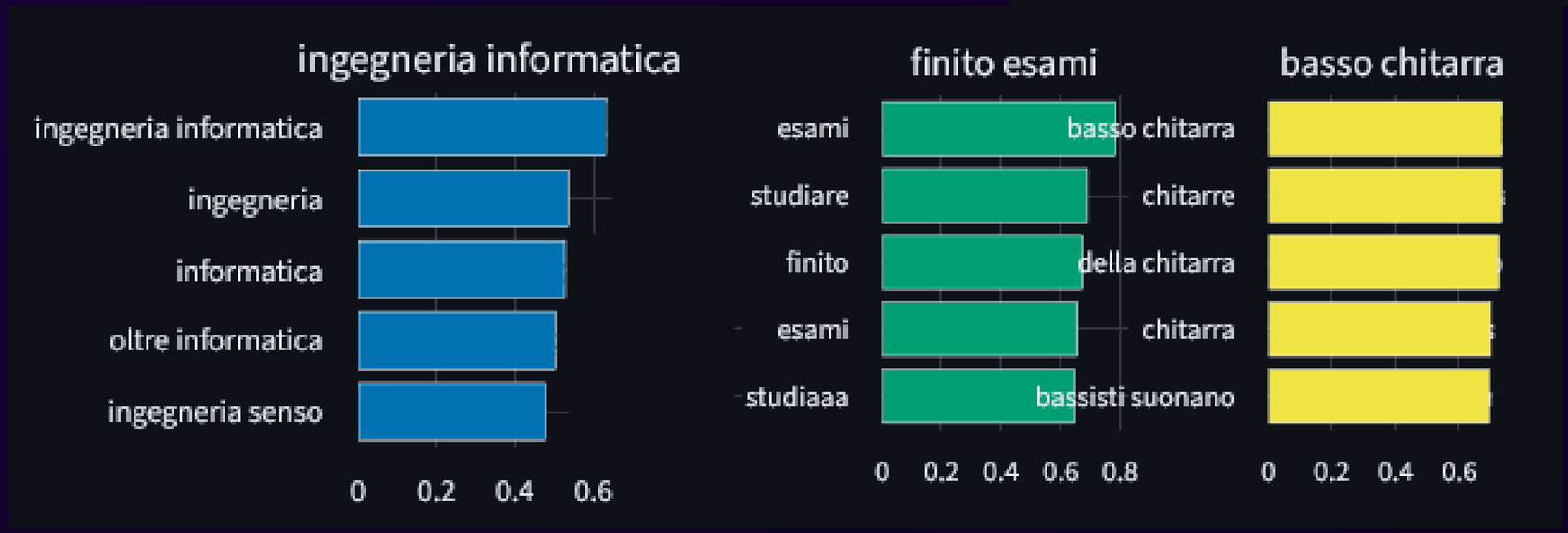
Min cluster size	Outliers rate	Execution time
10-15	10%-25%	Higher
50-60	30%-45%	Lower

- Max number of topics introduces a trade-off between interpretability and coverage
- The effect of clustering parameters is strongly influenced by the nature of the data

RESULTS

Qualitative analysis

- Presence of Low-Informative Topics
- Noise and Semantic Ambiguity
  (e.g. spelling errors)
- Still, relevant topics are extracted



RESULTS

Findings

- Ability to extract meaningful and coherent topics
- Topic modeling highly depends on both configuration of parameters and the nature of the data
- LLM-based representation models are slower but produce a better descriptive label

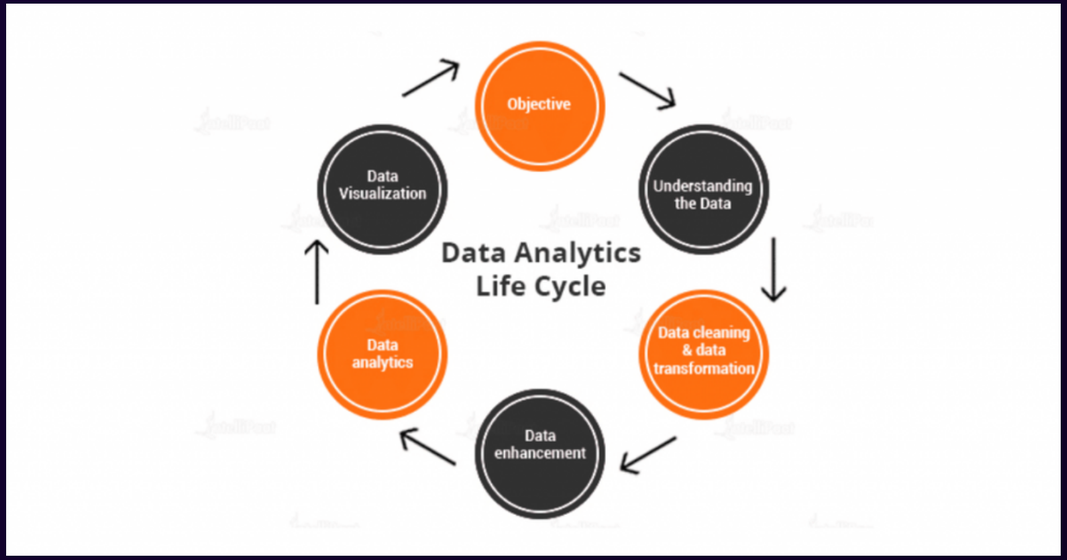
KeyBERT	llama2-7b
bass guitar	Musical instruments
pizza eat	Food and flavors
fast internet online	Starlink

CONCLUSION

Future work

This framework introduces the analysis of personal data, but there is still room for improvements:

- Introducing context to textual data
- Analyzing more data sources such as photos/videos
- Automate the entire pipeline
- Enrich DDP using external metadata

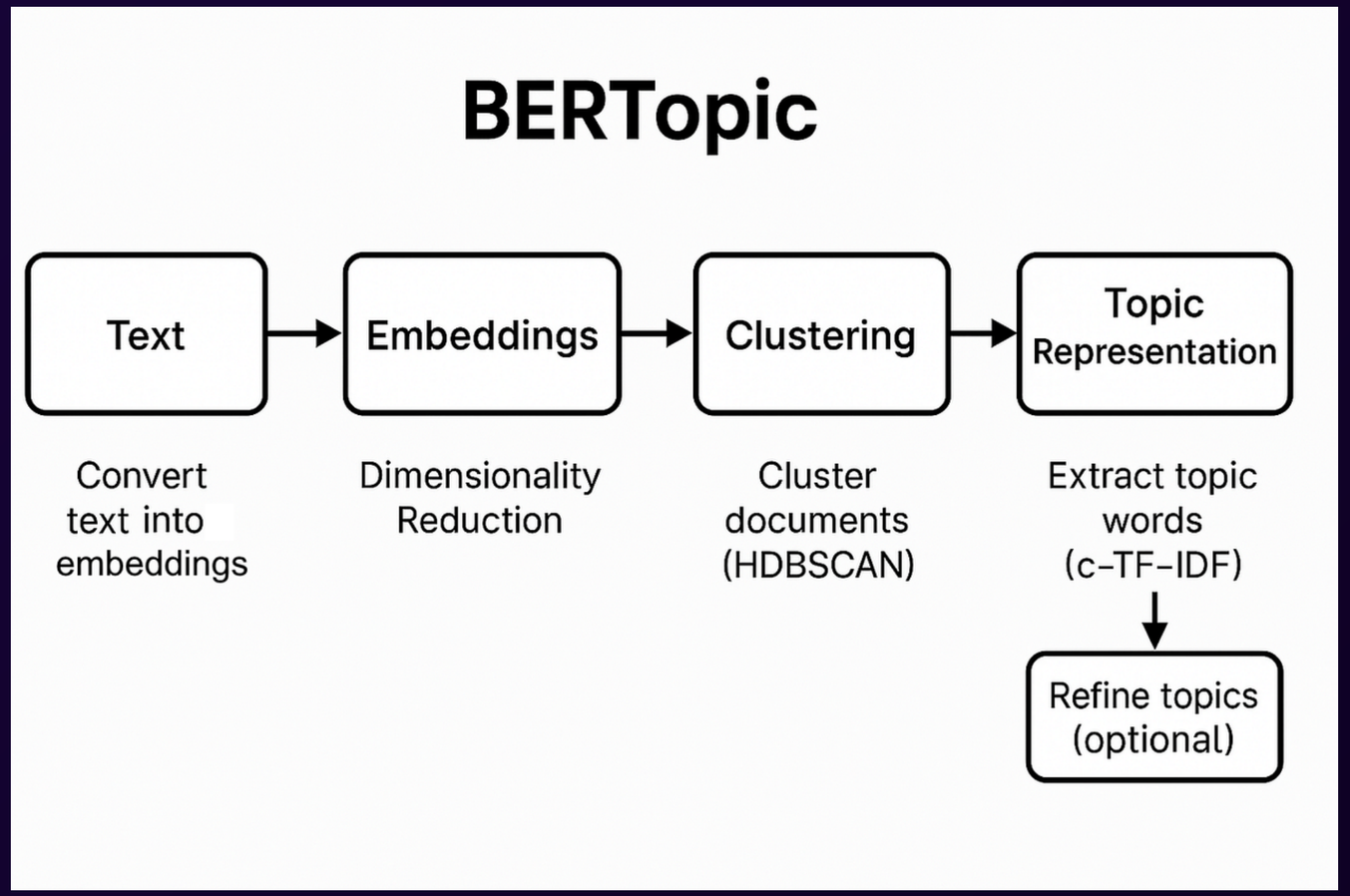


THANKS FOR YOUR ATTENTION

ANY QUESTIONS?

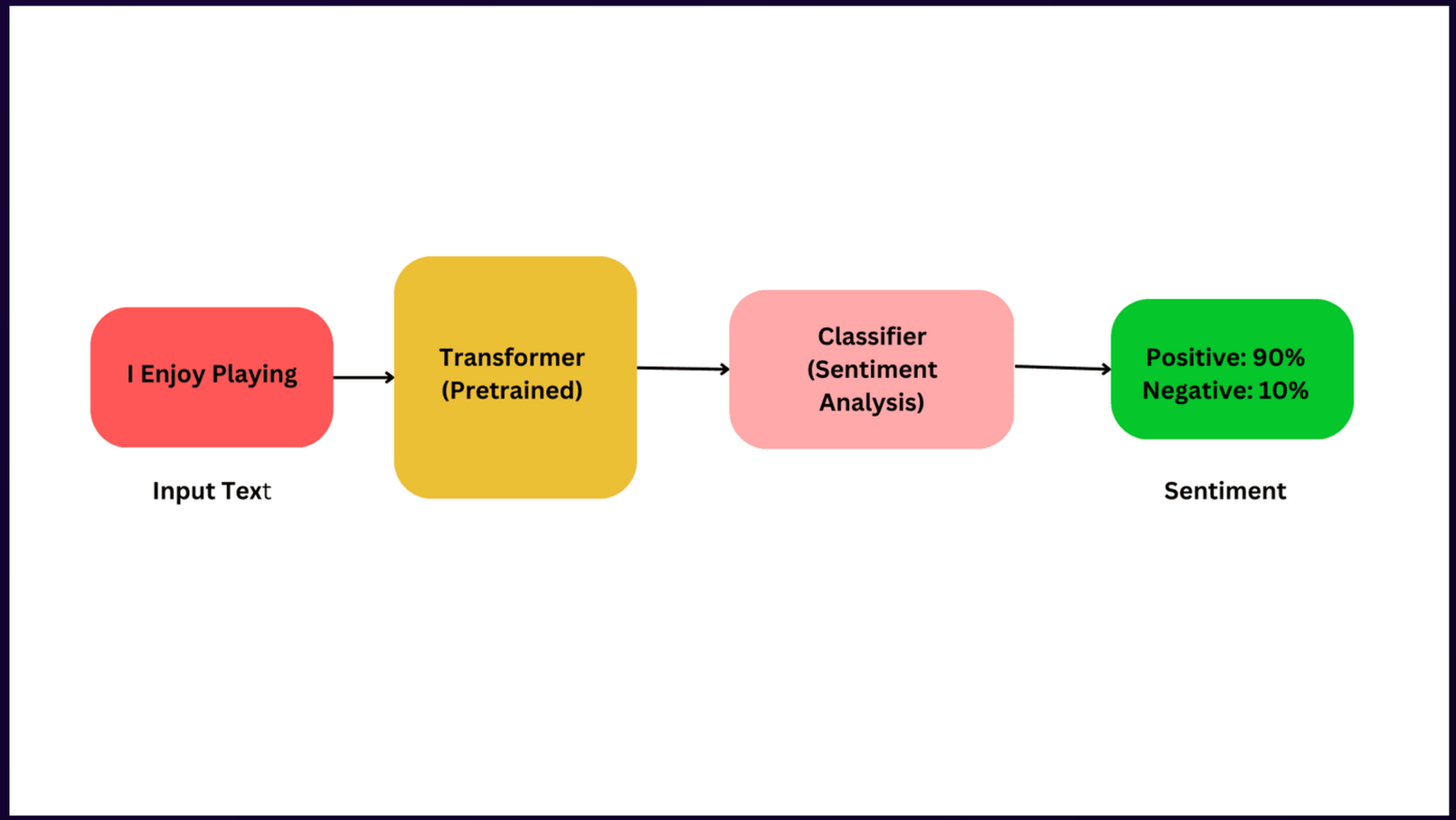
PROPOSED SOLUTION

Topic modeling



PROPOSED SOLUTION

Sentiment analysis



PROPOSED SOLUTION

Output visualization