

Explainer: TikTok Tactics, Far-Right Influence and the Italian Prime Minister Giorgia Meloni

This is an additional explainer accompanying “*TikTok Tactics, Far-Right Influence and the Italian Prime Minister Giorgia Meloni*” outlining some of the methodology that may be useful for other researchers, including data analysis, automated transcription and translation, and topic modelling used to examine Meloni’s TikTok account posts. All the code is [available on GitHub](#).

Open Source Research and TikTok

With more political discourse happening on TikTok, using the platform as a research tool is becoming increasingly attractive.

At the time of writing, TikTok, offers a Research API, which allows academic researchers from non-profit universities in the U.S. and Europe to query a secure application programming interface (API) and retrieve public data about TikTok accounts and content including user profiles, videos and comments. To gain access researchers must submit a clearly defined research proposal to be approved by TikTok.¹

TikTok’s access rules eliminate some journalists, researchers and other investigators from the opportunity to provide transparency on the social and political impacts of TikTok. Journalists and researchers who cannot access the official API therefore have to find alternative routes to access the data.

In August 2023, one alternative route was an open source software scraping tool, David Teather’s TikTok-API framework, which could access data about comments, trending videos, users, as well as data that allows videos to be downloaded. The process of using David Teather’s API and the use of different tools in the analysis of this data are described in detail throughout this article.

As a note, this current API is no longer available due to TikTok implementing enhanced security measures later in 2023. Interested individuals can stay up-to-date on adaptations to the api on the [github page](#) or following other manual research techniques like those [shared by Bellingcat](#) (an investigation organisation specializing in open source intelligence techniques).

¹ For more information, see: <https://developers.tiktok.com/products/research-api/>

The researcher hopes the process outlined in this guide can still be useful for understanding how data from TikTok and other video-based social media can be analysed and visualised in an investigation.

Note: [The Kit by Exposing the Invisible](#) is a good place to start for understanding the basics and beyond of conducting an online investigation.

Workflow: Download, Clean, Analysis

This project followed a standard data journalism workflow: download, clean and then analyse data.

1. Download Data

To download the data you will need a TikTok account.

It's better not to use a personal account for this as it may get flagged and blocked, impacting your personal use of the platform and potentially identifying you and creating risk for you as an investigator. Instead, create a dedicated "shadow account".

Note: the download process described here assumes familiarity with Git and Python toolsets and programming.

First clone the [TikTok-API](#) framework repository, setup a Python environment and install the dependency packages.

Now you can write, or use pre-written, scripts that can query and download metadata (captions, hashtags, play counts, shares, etc.) and videos. I have created an [Influence Industry Italy repository](#) which contains scripts that do this.²

The downloaded data includes the URLs (web addresses) of the videos. From these URLs it is possible to download the MP4 video files using [youtube-dl](#). Audio was extracted using the Python [moviepy tool](#).

Audio was then submitted to AWS (Amazon Web Services) Transcribe for transcription. You will need an AWS account to access the service. AWS provides some [tutorials](#) to support users.

The majority of caption and hashtag text was in Italian. Italian text was translated into English using the deep-translator Python framework to call Google Translate.

²For other examples check out the TikTok-API Examples folder. (as a footnote or tip box) <https://github.com/davidteather/TikTok-API/tree/main/examples>

2. Clean the data

TikTok Post and Transcription Metadata

The transcribed data from AWS Transcribe was verbose including a lot of detailed data such as *start_time*, *end_time*, *alternatives*, *confidence*, *content*, and *type* attributes for each word (as seen in figure X below).

```
{
  "jobName": "tiktoks",
  "TranscribeTikTokAudio7063465536109235461",
  "transcripts": "TranscribeTikTokAudio7063465536109235461",
  "accountId": {
    "items": "994611526795",
    "transcripts": "994611526795",
    "results": {
      "items": [
        {
          "start_time": "0.019",
          "end_time": "0.25",
          "alternatives": [
            {
              "confidence": "0.395",
              "content": "tendere",
              "type": "pronunciation"
            },
            {
              "confidence": "0.259",
              "end_time": "0.419",
              "alternatives": [
                {
                  "confidence": "0.951",
                  "content": "una",
                  "type": "pronunciation"
                },
                {
                  "confidence": "0.991",
                  "content": "legge",
                  "type": "pronunciation"
                },
                {
                  "confidence": "0.689",
                  "end_time": "1.289",
                  "alternatives": [
                    {
                      "confidence": "0.994",
                      "content": "proporzionale",
                      "type": "pronunciation"
                    },
                    {
                      "start_time": "1.44",
                      "end_time": "1.69",
                      "alternatives": [
                        {
                          "confidence": "0.999",
                          "content": "perch\u00e0",
                          "type": "pronunciation"
                        },
                        {
                          "confidence": "0.999",
                          "content": "legge",
                          "type": "pronunciation"
                        }
                      ]
                    },
                    {
                      "start_time": "1.779",
                      "alternatives": [
                        {
                          "confidence": "0.998",
                          "content": "la",
                          "type": "pronunciation"
                        },
                        {
                          "start_time": "1.789",
                      "end_time": "2.039",
                      "alternatives": [
                        {
                          "confidence": "0.999",
                          "content": "legge",
                          "type": "pronunciation"
                        },
                        {
                          "start_time": "2.049",
                      "end_time": "2.609",
                      "alternatives": [
                        {
                          "confidence": "0.996",
                          "content": "proporzionale",
                          "type": "pronunciation"
                        },
                        {
                          "start_time": "2.619",
                      "end_time": "3.019",
                      "alternatives": [
                        {
                          "confidence": "0.998",
                          "content": "serve",
                          "type": "pronunciation"
                        },
                        {
                          "start_time": "3.029",
                      "end_time": "3.22",
                      "alternatives": [
                        {
                          "confidence": "0.994",
                          "content": "a",
                          "type": "pronunciation"
                        },
                        {
                          "start_time": "3.23",
                      "end_time": "3.71",
                      "alternatives": [
                        {
                          "confidence": "0.999",
                          "content": "impedire",
                          "type": "pronunciation"
                        },
                        {
                          "start_time": "3.72",
                      "end_time": "3.759",
                      "alternatives": [
                        {
                          "confidence": "0.998",
                          "content": "al",
                          "type": "pronunciation"
                        },
                        {
                          "start_time": "3.769",
                      "end_time": "4.25",
                      "alternatives": [
                        {
                          "confidence": "0.397",
                          "content": "centro",
                          "type": "pronunciation"
                        },
                        {
                          "start_time": "4.26",
                      "end_time": "4.309",
                      "alternatives": [
                        {
                          "confidence": "0.997",
                          "content": "di",
                          "type": "pronunciation"
                        },
                        {
                          "start_time": "4.32",
                      "end_time": "4.579",
                      "alternatives": [
                        {
                          "confidence": "0.995",
                          "content": "vincere",
                          "type": "pronunciation"
                        },
                        {
                          "alternatives": [
                            {
                              "confidence": "0.0",
                              "content": ":",
                              "type": "punctuation"
                            },
                            {
                              "start_time": "4.59",
                              "end_time": "4.71",
                              "alternatives": [
                                {
                                  "confidence": "0.998",
                                  "content": "ma",
                                  "type": "pronunciation"
                                },
                                {
                                  "start_time": "4.719",
                              "end_time": "4.789",
                              "alternatives": [
                                {
                                  "confidence": "0.996",
                                  "content": "le",
                                  "type": "pronunciation"
                                },
                                {
                                  "start_time": "4.8",
                              "end_time": "5.23",
                              "alternatives": [
                                {
                                  "confidence": "0.999",
                                  "content": "dico",
                                  "type": "pronunciation"
                                },
                                {
                                  "start_time": "5.239",
                              "end_time": "5.599",
                              "alternatives": [
                                {
                                  "confidence": "0.999",
                                  "content": "se",
                                  "type": "pronunciation"
                                },
                                {
                                  "start_time": "5.76",
                              "end_time": "6.01",
                              "alternatives": [
                                {
                                  "confidence": "0.993",
                                  "content": "anche",
                                  "type": "pronunciation"
                                },
                                {
                                  "start_time": "6.019",
                              "end_time": "6.13",
                              "alternatives": [
                                {
                                  "confidence": "0.999",
                                  "content": "fosse",
                                  "type": "pronunciation"
                                },
                                {
                                  "start_time": "6.389",
                              "end_time": "6.5",
                              "alternatives": [
                                {
                                  "confidence": "0.999",
                                  "content": "una",
                                  "type": "pronunciation"
                                },
                                {
                                  "start_time": "6.51",
                              "end_time": "6.73",
                              "alternatives": [
                                {
                                  "confidence": "0.999",
                                  "content": "legge",
                                  "type": "pronunciation"
                                },
                                {
                                  "start_time": "6.739",
                              "end_time": "7.55",
                              "alternatives": [
                                {
                                  "confidence": "0.998",
                                  "content": "proporzionale",
                                  "type": "pronunciation"
                                },
                                {
                                  "start_time": "8.18",
                              "end_time": "8.369",
                              "alternatives": [
                                {
                                  "confidence": "0.999",
                                  "content": "con",
                                  "type": "pronunciation"
                                },
                                {
                                  "start_time": "8.38",
                              "end_time": "8.39",
                              "alternatives": [
                                {
                                  "confidence": "0.998",
                                  "content": "i",
                                  "type": "pronunciation"
                                },
                                {
                                  "start_time": "8.399",
                              "end_time": "8.789",
                              "alternatives": [
                                {
                                  "confidence": "0.999",
                                  "content": "numer",
                                  "type": "pronunciation"
                                },
                                {
                                  "start_time": "8.8",
                              "end_time": "8.88",
                              "alternatives": [
                                {
                                  "confidence": "0.999",
                                  "content": "di",
                                  "type": "pronunciation"
                                },
                                {
                                  "start_time": "8.89",
                              "end_time": "9.43",
                              "alternatives": [
                                {
                                  "confidence": "0.999",
                                  "content": "oggi",
                                  "type": "pronunciation"
                                },
                                {
                                  "start_time": "9.439",
                              "end_time": "9.88",
                              "alternatives": [
                                {
                                  "confidence": "0.998",
                                  "content": "scusi",
                                  "type": "pronunciation"
                                },
                                {
                                  "start_time": "9.89",
                              "end_time": "10.51",
                              "alternatives": [
                                {
                                  "confidence": "0.997",
                                  "content": "gilet",
                                  "type": "pronunciation"
                                },
                                {
                                  "alternatives": [
                                    {
                                      "confidence": "0.0",
                                      "content": ":",
                                      "type": "punctuation"
                                    },
                                    {
                                      "start_time": "10.519",
                                      "end_time": "10.81",
                                      "alternatives": [
                                        {
                                          "confidence": "0.999",
                                          "content": "poi",
                                          "type": "pronunciation"
                                        },
                                        {
                                          "start_time": "11.26",
                                          "alternatives": [
                                            {
                                              "confidence": "0.999",
                                              "content": "quando",
                                              "type": "pronunciation"
                                            },
                                            {
                                              "start_time": "11.489",
                                          "end_time": "11.569",
                                          "alternatives": [
                                            {
                                              "confidence": "0.998",
                                              "content": "si",
                                              "type": "pronunciation"
                                            },
                                            {
                                              "start_time": "11.579",
                                          "end_time": "11.64",
                                          "alternatives": [
                                            {
                                              "confidence": "0.998",
                                              "content": "va",
                                              "type": "pronunciation"
                                            },
                                            {
                                              "start_time": "11.649",
                                          "end_time": "11.699",
                                          "alternatives": [
                                            {
                                              "confidence": "0.999",
                                              "content": "a",
                                              "type": "pronunciation"
                                            },
                                            {
                                              "start_time": "11.71",
                                          "end_time": "12.06",
                                          "alternatives": [
                                            {
                                              "confidence": "0.999",
                                              "content": "votare",
                                              "type": "pronunciation"
                                            },
                                            {
                                              "alternatives": [
                                                {
                                                  "confidence": "0.0",
                                                  "content": ":",
                                                  "type": "punctuation"
                                                },
                                                {
                                                  "start_time": "12.069",
                                                  "end_time": "12.3",
                                                  "alternatives": [
                                                    {
                                                      "confidence": "0.996",
                                                      "content": "magari",
                                                      "type": "pronunciation"
                                                    },
                                                    {
                                                      "start_time": "12.31",
                                                  "end_time": "12.64",
                                                  "alternatives": [
                                                    {
                                                      "confidence": "0.999",
                                                      "content": "fratelli",
                                                      "type": "pronunciation"
                                                    }
                                                  ]
                                                }
                                              ]
                                            }
                                          ]
                                        }
                                      ]
                                    }
                                  ]
                                }
                              ]
                            }
                          ]
                        }
                      ]
                    }
                  ]
                }
              ]
            }
          ]
        }
      ]
    }
  }
}
```

Image 1: Screenshot of transcribed data from AWS Transcribe

The cleaning process extracted all of the individual words and combined them into a series of sentences under a single transcript attribute that looked like this:

```
{
  "createTime": {
    "0": "1644520033000"
  },
  "id": {
    "0": "7063159758185827590"
  },
  "transcript": {
    "0": "huge increases in costs for citizens and businesses, electricity and gas bills tripled, thousands of businesses exhausted and without concrete help will be forced to permanently close, millions of workers at risk. In this disastrous situation, what will be the concern of the government press and Italian politics will be to give answers to families and businesses that have been on their knees for days now, the most discussed topic by the media and politics. You know who Giorgia Meloni is, for a change, who, when asked if she will vaccinate her five-year-old daughter, dared to answer no, having evaluated, like all the other Italian mothers and fathers, the relationship between risks and benefits. A free choice and I agree that the government allows investigations, newspaper headlines, politicians and members of the intelligentsia, all concentrated on this theme, all concentrated on painting and building the monster. Why? To divert citizens' attention from the reality they are experiencing, the entire mainstream is deployed to prevent them from protesting against a government that is objectively a little late to avoid yet another blow that is about to hit Italians. To try, while we're at it, to criminalize the only opposition to the current government, they even went so far as to use my daughter to attack me because these people have no scruples in a scruple because they are in disarray. So I turn to free Italians, be aware of the fact that this game that is being played on my skin is actually being played to cheat you, because until we return to elections they will do everything to tarnish us, to prevent people from opposing the policies of this government, to prevent people from voting the way they don't want. But while they will continue to try to pit citizens against each other and divert attention from their failures, from their incompetence, with these means we will not give respite. We in the classrooms and in the squares will continue to fight to represent all those Italians who ask for answers and who do not give up on this reality. I'm sorry, don't stop us. "
```

Image 2: Screenshot of AWS Transcribe data that has undergone the cleaning process

Further, the TikTok-API posts contained large quantities of meta-data. The main JSON file used in this investigation had more than 40 000 lines with 24 attributes at the top level some of which contained multiple child attributes which themselves contained up to 5 levels of nested attributes. The 24 attributes at the top level of the hierarchy produced by the [TikTok-API](#) tool included things like *createTime*, *desc*, *music*, *video*, *collectCount*, *commentCount*, *playCount*, etc. looked like this:

```

{
  > "author": {-
  > },
  > "challenges": {-
  > },
  > "collected": {-
  > },
  > "contents": {-
  > },
  > "createTime": {-
  > },
  > "desc": {-
  > },
  > "digged": {-
  > },
  > "duetDisplay": {-
  > },
  > "duetEnabled": {-
  > },
  > "forFriend": {-
  > },
  > "id": {-
  > },
  > "itemCommentStatus": {-
  > },
  > "music": {-
  > },
  > "officialItem": {-
  > },
  > "originalItem": {-
  > },
  > "privateItem": {-
  > },
  > "secret": {-
  > },
  > "shareEnabled": {-
  > },
  > "stats": {-
  > },
  > "stitchDisplay": {-
  > },
  > "stitchEnabled": {-
  > },
  > "video": {-
  > },
  > "textExtra": {-
  > },
  > "warnInfo": {-
  > }
}

```

Image 3: The 24 attributes produced by the TikTok API tool

With densely nested child attributes like 'author' that contain child attributes such as avatarLarger, id, signature, nickname, uniqueId:

```

{
  "author": {
    "q": {
      "avatarLarger": "https://p16-sign-va.tiktokcdn.com/tos-maliva-avt-0068/0bec253cf0dadba9322e33da9de26918-c5_1080x1080.jpeg?x-expires=1694883606&x-signature=cU3DG5u4e7inzul7LH2wYhbLE20k3D",
      "avatarMedium": "https://p16-sign-va.tiktokcdn.com/tos-maliva-avt-0068/0bec253cf0dadba9322e33da9de26918-c5_720x720.jpeg?x-expires=1694883606&x-signature=3FgXh2Fkxh5d9HsGDvRv0qYuwkL0k3D",
      "avatarThumb": "https://p16-sign-va.tiktokcdn.com/tos-maliva-avt-0068/0bec253cf0dadba9322e33da9de26918-c5_1080x1080.jpeg?x-expires=1694883606&x-signature=3FatgHwJ9FrpI28g10zFmP7mW10k3D",
      "commentSetting": 0,
      "downloadSetting": 0,
      "duetSetting": 0,
      "ftc": false,
      "id": "7057902765381534725",
      "isADVirtual": false,
      "isEmbedBanned": false,
      "nickname": "Giorgia Meloni",
      "openFavorite": false,
      "privateAccount": false,
      "relation": 0,
      "secUid": "MS4wLjABAAAACw7rCwtXwxzNy9XotKY2Kor66nES25d91t0d7NmU-tYvTdEvG1b14jwjhaYqs5q",
      "secret": false,
      "signature": "Presidente del Consiglio dei Ministri della Repubblica Italiana",
      "stitchSetting": 0,
      "uniqueId": "giorgiameloni_ufficiale",
      "verified": true
    },
    "1": {-
  > },
  > "2": {-
  > },
  > "3": {-
  > },
  > "4": {-
  > },
  > "5": {-
  > },
  > "6": {-
  > },
  > "7": {-
  > },
  > "8": {-
  > },
  > }
}

```

Image 4: An example of nested child attributes produced by the TikTok API tool

The quantity of child attributes can make it slow and cumbersome to process, and many of the attributes at each level were not necessary for this investigation. Therefore, the data was 'cleaned' by removing the unnecessary points such as *author*, *challenges*, *duet related attributes*, *music*, *forFriend*.

Hashtags were parsed into clear singular phrases. The hierarchy was 'flattened' which allows the nested attributes to be associated directly to the top level post which is useful to the analysis process later described.

‘Engagement Ratio’

The figures which are provided by TikTok associated with overall engagement can be skewed as videos are initially featured to users on TikTok’s “For You” page, which is the default landing page that greets users when they open the app. Due to this set up, posts that aren’t actively engaged with still count as “viewed” (Cheng and Li, 2023). Therefore, the algorithm on this page isn’t necessarily based on a user’s active interaction (Medina Serrano et al., 2020) and so the video view count may be misleading.

Therefore, The engagement counts (*Like, Comment, and Share*) were converted into a single calculated Engagement Ratio. The Engagement Ratio formula (Cheng, Z. and Li, Y., 2023) divides the number of Likes, Comments and Shares by the number of views This Engagement Ratio was used as the basis for evaluating engagement with content in this investigation.

Cleaning the Automated Transcription and Translation

There were some issues with the AWS transcription service, for example, the transcription service didn’t recognise acronyms and LGBT was parsed out as individual letters. Furthermore, some of the audio also contained more than one speaker which isn’t automatically reflected in the data. Although it is possible to configure for more than one speaker within AWS Transcribe, in this case content was the focus of investigation rather than the speaker so this was not used. Background music and sound can also be an issue as they may interfere with whether the speaker or background song lyrics are transcribed. It is worth considering how much of an impact this may be for an investigation, in this case it was considered low-risk.

Italian hashtags sometimes produced two English words which required conversion into a hyphenated phrase to easily track the term. As a quality check, several of the translations were checked by fluent Italian speakers. The translation service was not entirely accurate sometimes creating misleading context and gender (sentences with ‘she’ were misgendered as ‘he’). For this investigation, these smaller errors were not a big problem as the intention was to understand the topics discussed.

Further steps to prepare the data for processing included:

- Convert each transcript post to a series of sentences
- Remove spoken word fillers like ‘um’, ‘ah’, ‘huh?’
- Label every sentence with the TikTok post date and time
- Replace ‘one hundred and ninety four’ with ‘194’ (Law one hundred and ninety four guarantees women’s right to abortion in Italy).³

³ This was done as BERTopic can’t interpret the phrase ‘one hundred and ninety four’ consistently as a phrase, instead it sees these as individual words and clusters them individually. Using 194 guarantees the clustering

- Transcripts with less than twenty seven words were excluded because some of these contained ‘noise’. Examples of this include posts where Meloni greets her audience from the stage whilst on the campaign trail saying, ‘Ah, good evening everyone. And for this presence, thank you.’. This text was printed out and manually verified to ensure it wasn’t significant for topic analysis.

3. Analyse and visualise the data

Python Jupyter Notebooks were used to run analytical algorithms including a search for repeating word counts in content and hashtags, producing engagement counts (likes, shares, comments) and Topic Modelling to generate visualisations in the form of word clouds, bar charts and BERTopic specific data visualisations.

In addition to the automated processes, the video content was manually analysed to identify aspects like setting, identity of speaker, and number of speakers.

Visualisations were used within the analysis process. For example, Python Matplot was used to visualise engagement, and hashtag counts, as well as both of these compared to a timeline (seen in Image 5):

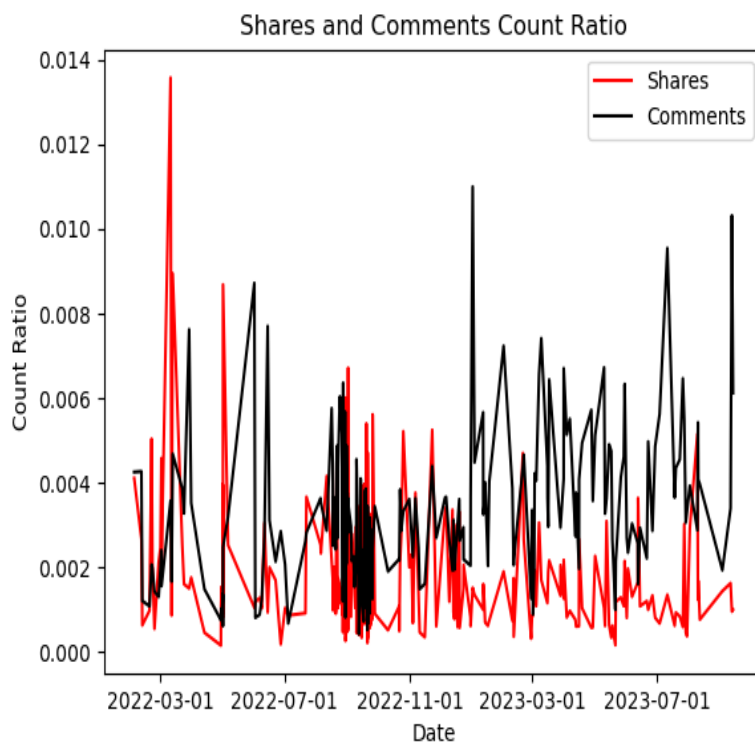


Image 5: Graph showing the engagement count ration of shares and comments over time

and ensures where the translation has output 194 instead of ‘one hundred and ninety four’ there is consistency.

The Python word cloud framework was used to visualise clusters of hashtags. The more often a word appears the larger it appears in the word cloud visualisation (seen in Image 6).



Image 6: A word cloud demonstrating the frequency of certain words found in the TikTok posts

Guided Topic Modelling

Topic modelling is an automated technique for scanning a large set of documents to detect patterns of words and phrases and cluster them in groups that characterise the overall content (Pascual, 2019). To do this, this investigation used BERTopic a Python framework that uses a Natural Language Processing algorithm to support topic modelling (an explainer and support can be found [here](#)). Topics are a cluster of words that relate to each other. A good example of this would be the Abortion topic where *194, the 194, an abortion*, can be seen as part of the Abortion topic cluster.

[Guided Topic Modelling](#) (otherwise known as Seeded Topic Modelling) “guides” the topic model by setting seed topics which the model will converge on. In this investigation labels relevant to the political debate, initial topics provided by BERTopic, and topics related to right-wing tactics (drawn from Jason Stanley’s *How Fascism Works*,⁴ Eviane Leidig’s *The Women of the Far Right*,⁵ and Ico Maly’s

4 Stanley, Jason. *How Fascism Works : The Politics of Us and Them*. Random House, 2020

5 Leidig, Eviane. *The Women of the Far Right*. Columbia University Press, 2023.

“Metapolitical New Right Influencers: The Case of Brittany Pettibone”⁶) such as migrants, abortion, fake news were provided as a topic list. Guided Seed topics were used to refine the topics, until ultimately the model was limited to twenty-nine topics. Some raw topics were renamed e.g. green pass > COVID Mandates for clarity and less common or relevant outliers were removed.

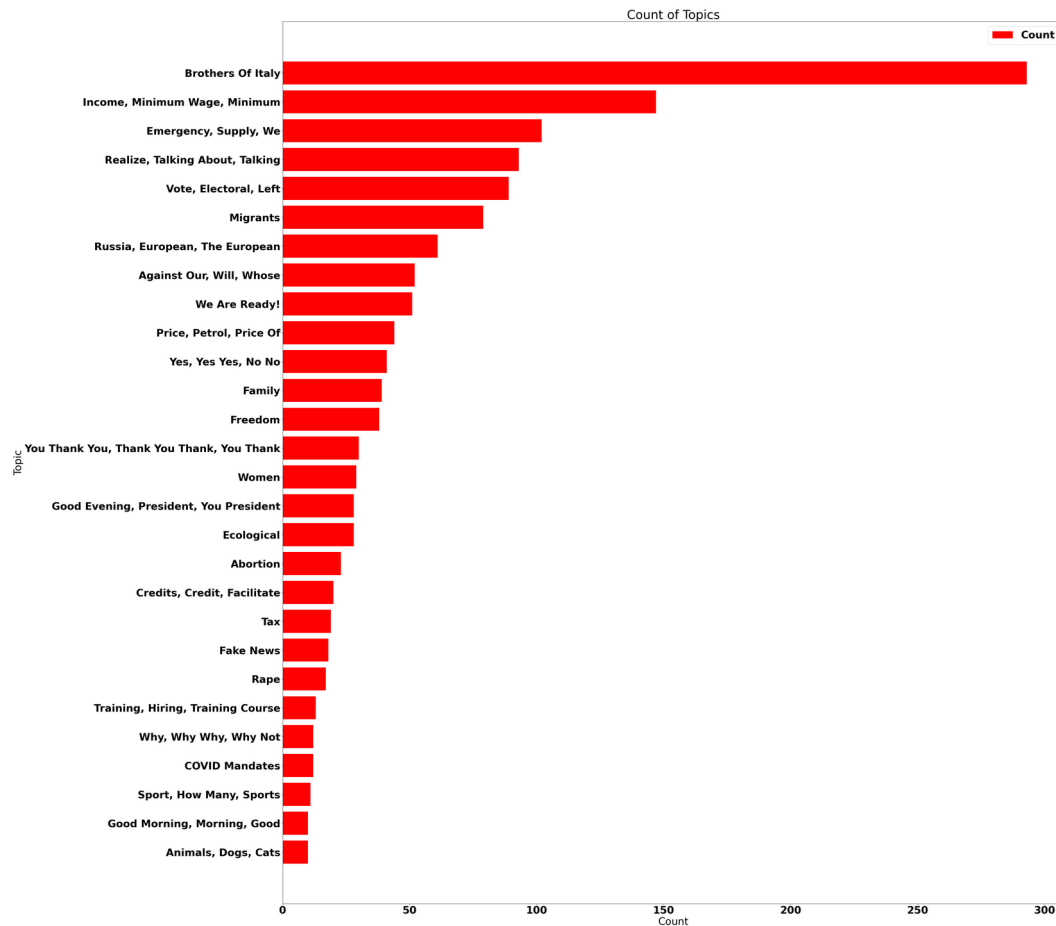


Image 7: The frequency of topics mentioned in the TikTok posts according to guided topic modeling approach

Visualisations

BERTopic provides multiple ways of visualising topics which can make identifying unusual distributions of data, patterns, clusters, gaps and outliers; in short a picture is worth a thousand words.⁷ For this investigation multiple BERTopic ways of visualising topics over time were used and helped especially with identifying topics prevalent during moments throughout the election campaign and after.

6 Maly, Ico. “Metapolitical New Right Influencers: The Case of Brittany Pettibone.” *Social Sciences*, vol. 9, no. 7, July 2020, p. 113, <https://doi.org/10.3390/socsci9070113>.

7 Unwin, A. (2020) ‘Why Is Data Visualization Important? What Is Important in Data Visualization?’, *Harvard Data Science Review*, 2(1). Available at: <https://doi.org/10.1162/99608f92.8ae4d525>.

Bar chart

This was used as a primary form of visualisation as it gives an indication of how often a specific topic is mentioned in comparison with other topics and provides a visual ranking (such as image 6 above).

Topic Word Scores

This topic visualisation serves a couple of roles, firstly it lets us verify that the BERTopic framework is working accurately and it also helps in identifying the topic cluster. (See full sized image on page 14)

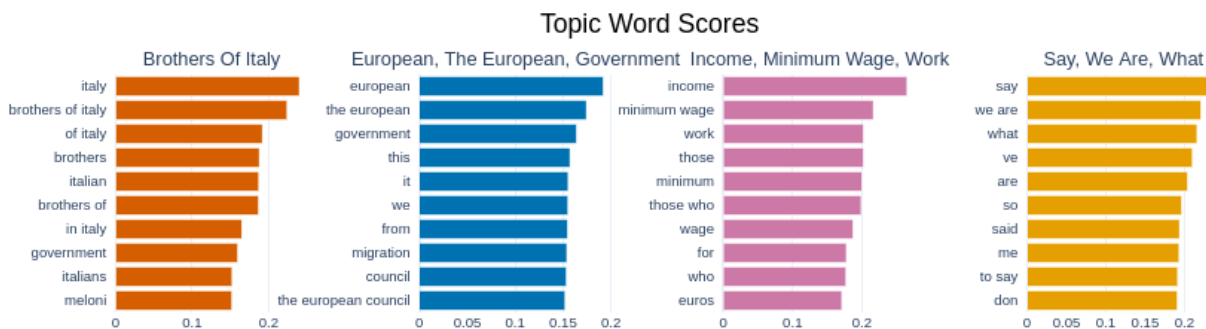


Image 8: Frequency of words mentioned in a topic cluster from the TikTok posts

<https://s3.amazonaws.com/frasercrichton.com.influence.industry/topic-word-scores.html>

Similarity Matrix

The Similarity Matrix indicates how similar certain topics are to each other and is useful for identifying topics that are closely aligned. It's possible in BERTopic to merge these topics to make the Topic Model simpler to understand.

Similarity Matrix

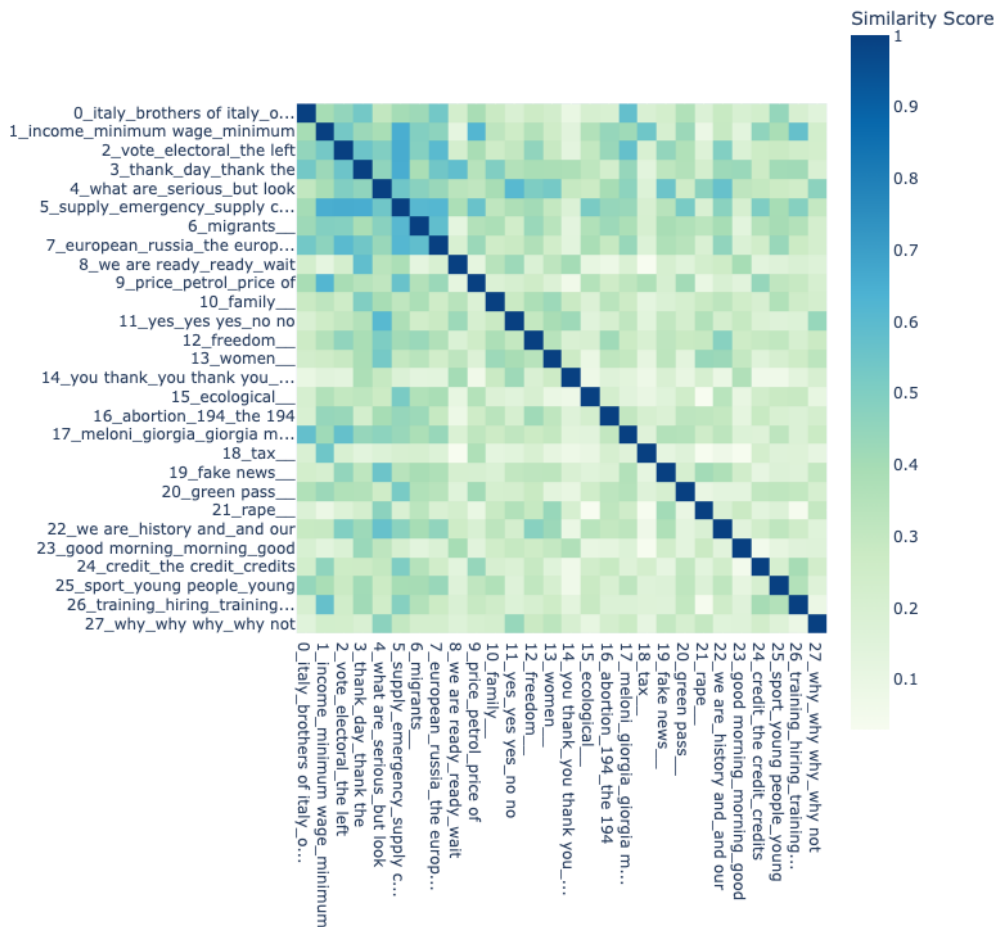


Image 9: Frequency of words mentioned in a topic cluster from the TikTok posts

<https://s3.amazonaws.com/frasercrichton.com.influence.industry/similarity-matrix.html>

Intertopic distance map

A intertopic distance map shows the proximity of use of words of one cluster of topics to another. The area of the circles represents the number of words that belong to each topic. Topics that are closer together have more words in common.⁸

For instance, in the lower left quadrant Women and Rape are close to each other. In the opposite quadrant, Brothers of Italy and Russia, European, The European are close.

⁸ *Getting to the Point with Topic Modeling | Part 3 - Interpreting the Visualization* (2020) Maveryx Community. Available at: <https://community.alteryx.com/t5/Data-Science/Getting-to-the-Point-with-Topic-Modeling-Part-3-Interpreting-the/ba-p/614992> (Accessed: 11 March 2024).



Image 10: Intertopic distance map showing the proximity of word frequency

<https://s3.amazonaws.com/frasercrichton.com/influence.industry/inter-topic-distance-map.html>

Topics over Time

Visualising Topics over Time is a way of seeing on what date a topic is most used. BERTopic can generate an interactive HTML visualisation that can be embedded in a web page.

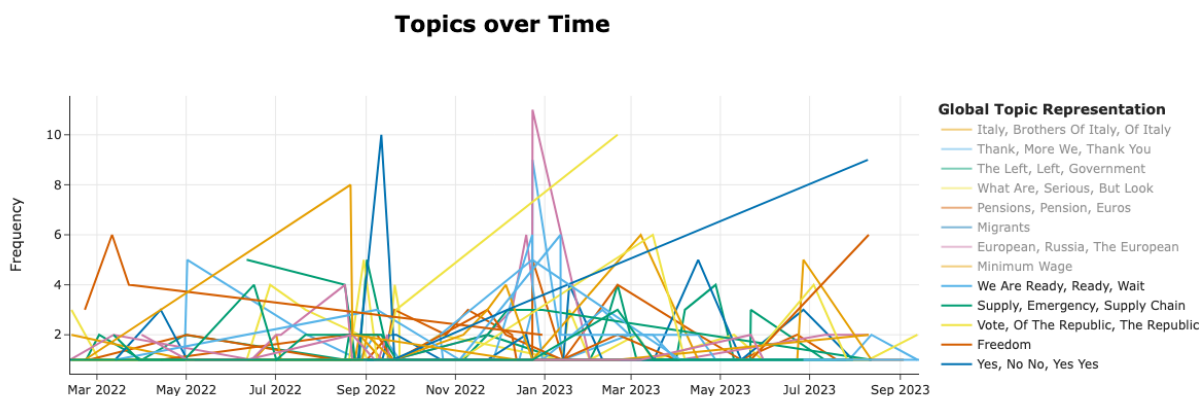


Image 11: Graph showing the frequency of topics between March 2022 - September 2023

<https://s3.amazonaws.com/frasercrichton.com/influence.industry/topics-over-time.html>

Researcher Reflections

Firstly, as this was an initial scoping exercise, the findings should be double-coded by a second researcher before being used further. Unfortunately, as TikTok changed access and the scraping tool isn't working, it is not possible to easily compare this dataset with a similar one for another Italian politician or another right-wing candidate elsewhere.

Secondly, given that the researcher is not fluent in Italian, the language of the subject content and the errors or biases that automated translation and transcription services may have introduced should be considered in the findings. The researcher manually verified a small selection of posts to ensure accurate translations with a fluent Italian speaker.

The tools (TikTok-API scraper, AWS Transcription and Google Translation services, BERTopic) saved time over manual analysis, brought new insights and helped visualise the underlying data. BERTopic produced powerful interactive visualisations and accurately identified seed topics. The flexibility that the framework has to support multiple different strategies for extracting data and topics is, perhaps, its weakness. It makes for complex configuration where you need to be able to understand some of the minutiae of large language models to be sure you are getting good results. Topic modelling is also [limited](#) on short text formats but still helped derive insights that might have taken longer in a manual approach. In the future, it might be interesting to use other LLMs like [ChatGPT](#) to discover narratives present in the posts, cluster them and visualise them over time or to do fact checking and verification.

Social media platforms are still finding a balance between privacy and transparency but the public data in these platforms is created by citizens and yet it the corporate entities both profit from and control access to that data. These platforms play a huge role in political campaigns and journalists need access to these resources to identify disinformation and hold politicians to account.

Read the case study:

[TikTok Tactics, Far-Right Influence and The Italian Prime Minister Giorgia Meloni](#)

About the Author

Fraser Crichton is a freelance researcher and visual artist. His project The Moral Drift examined abuse in state care in Aotearoa New Zealand and the State's failure to provide redress and accountability to survivors. This project have been exhibited at Toi Tauranga Tauranga Art Gallery and Te Pātaka Toi Adam Art Gallery. His writing has appeared in Open Democracy and he attended Tactical Tech's Investigating the Influence Industry: Summer School Lite. Connect with him on Bluesky: [@frasercrichton.com](https://bsky.app/profile/frasercrichton.com)

The open source code supporting this investigation is available on GitHub at: <https://github.com/frasercrichton/investigation-influence-industry-italy>

Topic Word Scores

