iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** III **Month of publication:** March 2026

DOI: https://doi.org/10.22214/ijraset.2026.78087

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue III Mar 2026- Available at www.ijraset.com

Training Generative AI on Social Media Data: Implications and Outputs - A Worked Out Example

Raabia Riaz[1], Muhammad Areeb Chatni[2], Dr. Tanzeel Ur Rehman[3]
[1]*Director Tech, RabsXpress*
[2]*Kingston University London*
[3]*University of Gloucestershire*

Abstract: ***This chapter presents a follow-along methodological exercise that shows how social media data can be used to study generative AI training and conditioning. It is written to be reproducible by both technical and non-technical readers and does not assume access to large commercial training pipelines. Using a simulation-based experimental approach, the chapter walks through collecting text from Reddit and Twitter (X), constructing two matched datasets of 2,500 texts per platform, and preparing those datasets for comparison through careful cleaning and documentation. Analysis is carried out in Orange Data Mining, using visual, screenshot-led steps to explore sentiment, basic linguistic patterns, and framing differences across platforms. The chapter then demonstrates prompt-based conditioning with a small open-source generative model, keeping prompts and generation settings fixed while varying only the platform-specific input text. The emphasis is on methodological clarity, transparency, and cautious interpretation rather than substantive claims about public opinion.***

I. INTRODUCTION

In the field of artificial intelligence (AI), systems based on generative technologies, particularly large language models (LLMs) and related text-generation architectures, have become central to contemporary AI development. These systems are increasingly deployed in client-facing applications, where user experience (UX/UI) design encourages engagement and continued use, as well as in organisational and public-sector contexts. In these settings, generative systems produce text that appears fluent, persuasive, and contextually adaptive. However, such outputs are not the result of "general intelligence" or neutral statistical computation. Rather, they reflect the data on which generative systems are trained or conditioned, the curation and filtering processes applied to those data, and the alignment strategies used to adapt models for specific tasks (Bender *et al.*, 2021; Bommasani *et al.*, 2021). For this reason, understanding generative AI requires methodological approaches that attend to data provenance and to the ways in which different data environments shape model behaviour.

Social media platforms are widely recognised as important sources of training material for contemporary AI systems. Even when direct training on social media data is not explicitly disclosed, social media discourse may enter training mixtures through web-scale corpora, content aggregation pipelines, and datasets derived from publicly available online text. Social media platforms are particularly attractive for this purpose because they generate large volumes of up-to-date content and reflect everyday language use. They capture how people discuss emerging topics and express opinions about socially salient issues in real time. At the same time, social media platforms are not neutral or uniform linguistic spaces. Each platform shapes communication through technical affordances, such as post length or reply structure, community norms, such as subreddit conventions, and moderation or ranking mechanisms (Gillespie, 2018). As a result, training or conditioning generative systems on social media data has clear implications for the tone, style, and framing that these systems reproduce.

Despite the central role of training data, research access to training pipelines remains limited. Commercial model development is often opaque, proprietary datasets are inaccessible, and the computational resources required for full model training are beyond the reach of most researchers. Consequently, much existing research focuses on model evaluation, bias identification, or broader societal impacts, rather than on the training stages and methodological steps through which data are translated into model behaviour (Bender *et al.*, 2021; Mitchell *et al.*, 2019). This gap is particularly relevant for methods handbooks. Handbooks are not primarily intended to report new empirical findings; rather, they aim to teach readers how to conduct research, document methodological decisions, and interpret outputs responsibly. Accordingly, the purpose of this chapter is to provide a transparent, step-by-step workflow demonstrating how researchers can collect social media text, construct matched corpora, analyse those corpora using widely available software, and use them to condition a generative model within a controlled simulation.

©IJRASET: All Rights are Reserved

The chapter uses a single worked example throughout: cost-of-living discourse in the United Kingdom during the year 2025, compared across Reddit and Twitter (X). This example is intentionally narrow and illustrative. The aim is not to claim representativeness of UK discourse or to generalise about platform behaviour. Instead, the example is designed to show readers how to move from platform access to data extraction, from corpus cleaning to exploratory analysis, and from conditioning to the examination of generated outputs. Orange Data Mining is used as the primary analytical tool because it provides transparent, visual workflows that are well suited to instructional contexts and support screenshot-based explanations of each methodological step (Demšar *et al.*, 2013). In addition, Orange offers a user-friendly environment for researchers who are new to data-driven analysis.

A further design choice in this chapter is the use of prompt-based conditioning rather than full model fine-tuning. Full training or fine-tuning would require substantial computational resources and would shift the focus of the chapter toward engineering concerns, which falls outside the scope of this methodological illustration. Prompt-based conditioning allows the relationship between exposure to different text corpora and variation in generated outputs to be examined while keeping the model fixed and the methodological logic visible. This approach can therefore be framed as a simulation-based experimental demonstration in which one variable, the platform-specific corpus, is varied while other conditions, including prompts, model choice, and generation parameters, are held constant.

The chapter is organised sequentially in line with the workflow it presents. Section 4.6.2 outlines the methodological logic of simulation-based illustration and explains its relevance for studying generative AI training processes. Section 4.6.3 describes the data sources and the rationale for comparing Reddit and Twitter as distinct discourse environments. Section 4.6.4 details platform access and data extraction, including practical steps for logging into APIs, specifying filters, and constructing matched corpora of 2,500 texts per platform. Section 4.6.5 addresses corpus preparation and standardisation, with emphasis on how cleaning decisions affect analysis and interpretation. Section 4.6.6 demonstrates exploratory analysis in Orange, including sentiment categorisation and linguistic indicators. Section 4.6.7 describes prompt-based conditioning using a small open-source model and illustrates how outputs can be compared across conditions. Finally, Sections 4.6.8 and 4.6.9 discuss implications, limitations, and ways in which the method can be adapted to other topics, platforms, and research contexts.

II. METHODOLOGICAL FRAMEWORK: SIMULATION-BASED ILLUSTRATION OF TRAINING

A common challenge in research on generative AI is the gap between how these models are developed in practice and what most researchers can study. Training modern language models requires large amounts of computing power and complex data pipelines that are usually not accessible or reproducible outside commercial settings. As a result, these systems can have wide social and institutional influence while remaining difficult to examine in detail. Simulation-based experimental research provides a practical way to address this problem by using simplified and controlled setups to explore how specific factors, such as training data, affect model outputs (Gilbert and Troitzsch, 2005; Epstein, 2006).

Simulation-based approaches are well established in computational social science, where researchers build models that intentionally simplify complex phenomena in order to examine mechanisms, test sensitivities to assumptions, or explore the effects of parameter variation (Epstein, 2006). The purpose of such simulations is not to claim that the model represents the phenomenon itself. Rather, simulations create an analytical context that makes certain relationships observable. Applied to generative AI, simulation-based illustration can be used to examine how exposure to different data environments influences generated outputs, without asserting that the simulation reproduces commercial training pipelines.

In this chapter, the simulation is designed around a simple experimental logic. A fixed base generative model is used under two conditions. In Condition A, the model is exposed to a curated set of Reddit texts about UK cost-of-living issues. In Condition B, the same model is exposed to a curated set of Twitter texts on the same topic. In both conditions, the prompts used for generation are identical, as are the generation settings. The only difference between the two conditions is the corpus used for conditioning. This control-and-variation design is central to experimental inference, even outside a laboratory setting, as it allows outputs to be interpreted as a function of a single, clearly defined variable.

Prompt-based conditioning can be understood as a practical way of exposing a model to specific types of text under controlled conditions. While it is possible to update model weights through fine-tuning when working with training data, this approach requires substantial computational resources and shifts the focus toward model engineering. In contrast, prompt-based conditioning involves providing the model with representative examples from a corpus as contextual input before generation. In current language model practice, prompting is not simply a user-facing interface but a means of influencing how responses are framed and how generated text is shaped (Wei et al., 2022).

For the purposes of methodological illustration, prompt-based conditioning is preferred because it is transparent and easy to reproduce. It allows readers to see how exposure to different data contexts can influence model outputs without encouraging claims about model performance or optimisation. At the same time, this approach has clear limitations. Prompt-based exposure is not equivalent to full model training, and it is therefore important to state explicitly that the simulation illustrates relationships between data context and generated outputs rather than replicating large-scale training processes.

A further reason simulation-based illustration is well suited to a methods handbook is that it keeps the focus on research process rather than on results. The aim is to show how to document design decisions, structure datasets, select and justify analytical measures, and interpret outputs carefully, rather than to optimise model performance or produce predictive claims. For readers, the main value lies in learning how to build a replicable workflow and how to avoid common pitfalls, such as over-interpreting sentiment scores, confusing platform features with user intent, or treating generated outputs as direct evidence of underlying social reality.

III. DATA SOURCES AND PLATFORM SELECTION: WHY REDDIT AND TWITTER (X)?

Social media research often treats social media as a single category, even though platforms differ considerably in how communication is structured. Each major platform emerged at a different point in time and under different social and technological conditions. As a result, platforms have evolved in distinct ways, shaped by design choices, user practices, and market pressures. These differences matter methodologically because platform design can influence text length, vocabulary use, emotional tone, and framing (Boyd and Ellison, 2007; Bruns and Stieglitz, 2013). Platform features also change over time through repeated iterations in response to user behaviour and competition within the social media landscape. For this reason, platform selection in a generative AI training simulation should be treated as a deliberate methodological decision rather than a neutral one.

Reddit and Twitter (X) offer a useful comparison because they support different forms of public communication, despite both having achieved widespread adoption and visibility. At the same time, both platforms have faced criticism and backlash related to moderation practices and the enforcement of community standards, including the suspension or banning of users for policy violations. Reddit is organised around subreddits that function as topical and community boundaries, encouraging longer posts, narrative accounts, and peer-to-peer discussion. Comment threads allow extended exchange and clarification, which often results in contextual, experience-based, and first-person discourse. Twitter, by contrast, was initially designed around short-form posting and continues to favour rapid responses and declarative statements, even as character limits and interaction features have expanded over time. Visibility on Twitter is strongly shaped by engagement metrics and algorithmic ranking, which tends to promote compressed expression, evaluative language, and framing that references broader political or policy contexts.

The worked example in this chapter focuses on cost-of-living discourse in the UK during 2025. This topic was selected for both practical and methodological reasons. It generates substantial discussion across platforms and combines personal experience with broader economic and political commentary, making it well suited for examining sentiment and framing. At the same time, it avoids higher-risk categories of data, such as sensitive health information or highly personal disclosures, while still providing meaningful social context. The UK focus is operationalised through subreddit selection, keyword filtering, and contextual cues within the text. While platform content alone cannot reliably establish geographic location, this limitation is acknowledged, and UK relevance is treated as a filtering criterion rather than a definitive classification.

Ethical and policy considerations also play an important role in shaping the data collection process. Public availability does not automatically make social media data free from ethical responsibility. Researchers must consider platform terms of service, user expectations, and established ethical guidance relating to consent, anonymisation, and potential harm (Townsend and Wallace, 2016; Fiesler and Proferes, 2018). For this reason, the workflow presented in this chapter collects text content only and excludes user identifiers. Analysis is conducted at an aggregate and illustrative level, and generated outputs are presented as methodological examples rather than as claims about individual users or communities.

IV. ACCESSING REDDIT AND TWITTER (X) APIS AND EXTRACTING TWO MATCHED CORPORA

Although this section is presented as an instructional walkthrough, it also serves as a practical use case that demonstrates how two platforms can be compared using APIs. The intention is to show how both technical and non-technical users can approach data collection in a structured and reproducible way. Readers are expected to be able to follow the steps and replicate the process, substituting their own topic, platform, or time window if required. The worked example focuses on a single topic and year, cost-of-living discourse in the UK during 2025 and aims to extract 2,500 texts from each platform.

©IJRASET: All Rights are Reserved

1) Step 1: Define inclusion criteria before collecting data

Before logging in to any API, define inclusion criteria. Doing this upfront reduces post-hoc bias and ensures the corpus is not shaped entirely by what is easiest to retrieve.

For this example, the inclusion criteria are:

- Platform: Reddit and Twitter (X)
- Year: 2025 (January–December)
- Language: English
- Topic: Cost of living in the UK
- Unit of analysis: One "text" per row (a Reddit post or comment; a tweet/post)
- Sample size: 2,500 texts per platform
- Fields stored: platform label, year, subreddit name (Reddit only), text content

2) Step 2: Reddit API access (developer app + authentication)

a) Create a Reddit developer application

To access Reddit data via API, create a developer app (script-type). In practice, this requires logging into Reddit, navigating to the application management page, and registering a script. The app will provide credentials (client ID and secret) used for authentication.

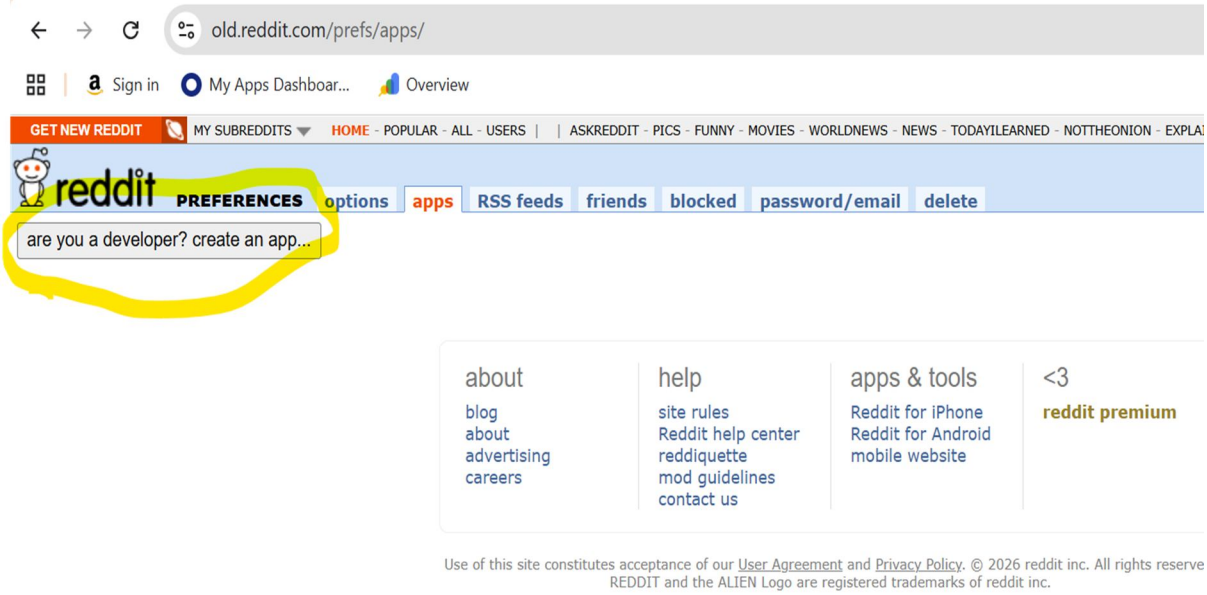


Figure 1: Screenshot of Reddit developer app creation page and app credentials. Access here: https://old.reddit.com/prefs/apps/.

The next step is to fill out this form shown below:

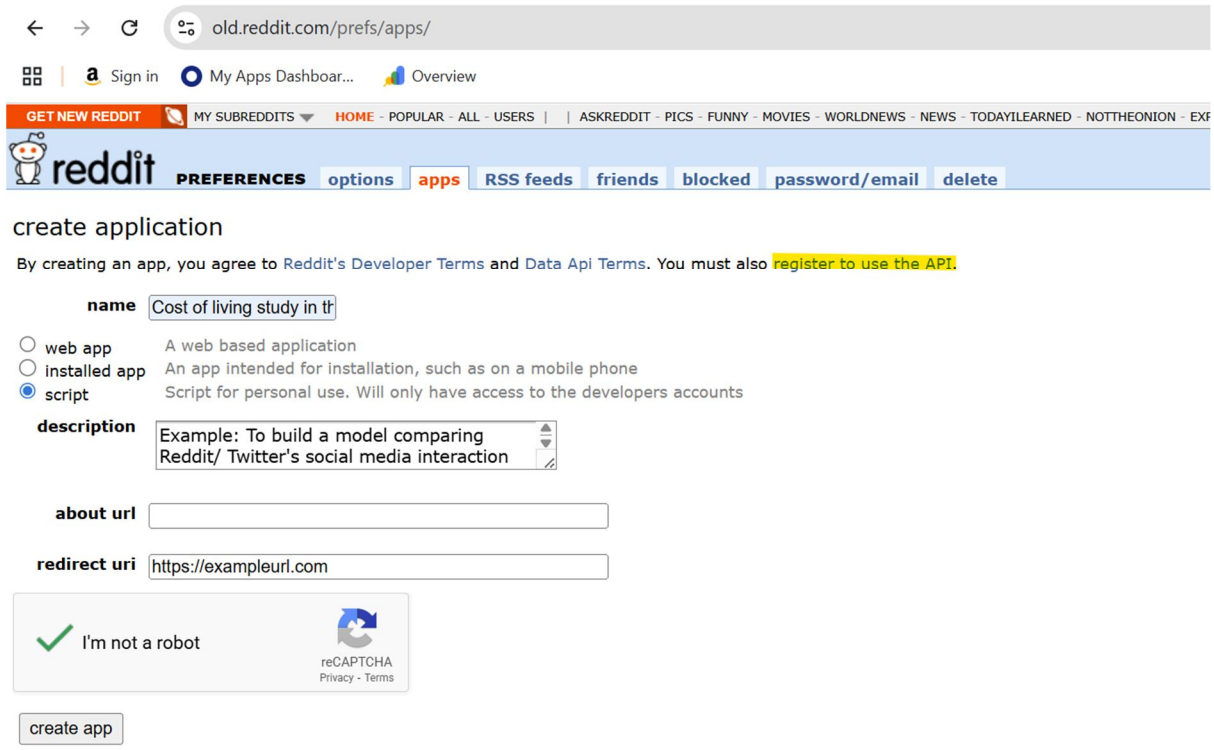


Figure 2: Fill out the key fields and select the format you would like to use.

The important aspect to note in figure 2 is the yellow marked text where a user needs to register for the API separately. As shown above, the key fields typically include name of the app, format you want to use to code, description, about url and redirect url. It is advisable to fill out this information in as detailed as possible to illustrate this.

b) Authenticate using an API library

Researchers commonly use a wrapper library (e.g., PRAW in Python) to handle authentication and requests. The chapter does not require readers to be advanced programmers; the point is to show what authentication consists of and how access is obtained. For handbook purposes, it is sufficient to explain the flow: credentials > token > query > returned data.

Below is the example of working code as an alternative method to those who are more familiar with coding and technical methods. This working code includes notes for guidance too so the beginners can also navigate through this.

```
"""
Reddit API Authentication + Example Data Pull (for handbook screenshot)
Topic: UK cost of living (illustrative)
NOTE: Replace placeholders with your own credentials before running.
For screenshots, KEEP placeholders or mask credentials (recommended).
"""

import praw
from datetime import datetime, timezone
import pandas as pd
# -------------------------
# 1) AUTHENTICATION CONFIG
# -------------------------
# Create a Reddit script app here:
# https://www.reddit.com/prefs/apps  (or https://old.reddit.com/prefs/apps)
```

```
reddit = praw.Reddit(
    client_id="YOUR_CLIENT_ID",            # <-- replace
    client_secret="YOUR_CLIENT_SECRET",    # <-- replace
    username="YOUR_REDDIT_USERNAME",       # <-- optional; include if needed
    password="YOUR_REDDIT_PASSWORD",       # <-- optional; include if needed
    user_agent="UKCostOfLivingStudy/0.1 (by u_YOUR_REDDIT_USERNAME)"
)

# Quick auth test (will raise an error if auth fails)
print("Authenticated as:", reddit.user.me())


# -------------------------
# 2) STUDY PARAMETERS
# -------------------------
SUBREDDITS = ["UKPersonalFinance", "AskUK", "unitedkingdom", "HousingUK"]

QUERIES = [
    '"cost of living"',
    "inflation",
    "rent",
    "mortgage",
    '"energy bills"',
    "groceries",
    '"food prices"',
    "wages",
    "salary",
    '"housing costs"',
    '"council tax"'
]

START_TS = datetime(2025, 1, 1, tzinfo=timezone.utc).timestamp()
END_TS = datetime(2025, 12, 31, 23, 59, 59, tzinfo=timezone.utc).timestamp()

MIN_WORDS = 10          # filter out very short texts
MAX_PER_QUERY = 300     # number of results per query per subreddit (adjust if needed)


# -------------------------
# 3) COLLECT POSTS
# -------------------------
rows = []

for sub in SUBREDDITS:
    subreddit = reddit.subreddit(sub)

    for q in QUERIES:
        # Note: Reddit search is not perfect; you may need multiple queries.
        for submission in subreddit.search(q, sort="new", limit=MAX_PER_QUERY):
```

```
        created = submission.created_utc
        if created < START_TS or created > END_TS:
            continue

        title = submission.title or ""
        body = submission.selftext or ""
        text = (title + "\n" + body).strip()

        # Skip removed/deleted/empty/short
        if not text or text in ["[deleted]", "[removed]"]:
            continue
        if len(text.split()) < MIN_WORDS:
            continue

        rows.append({
            "platform": "reddit",
            "year": 2025,
            "subreddit": sub,
            "text": text,
            "created_utc": int(created),
            "post_id": submission.id
        })

print("Raw collected rows:", len(rows))


# -------------------------
# 4) DEDUPE + SAMPLE TO 2500
# -------------------------
df = pd.DataFrame(rows)

# Deduplicate by post id, then by text (extra safety)
df = df.drop_duplicates(subset=["post_id"])
df = df.drop_duplicates(subset=["text"])

print("After dedupe:", len(df))

# If you collected more than 2500, sample down
TARGET_N = 2500
if len(df) >= TARGET_N:
    df = df.sample(n=TARGET_N, random_state=42)
else:
    print(f"Warning: only {len(df)} texts collected. "
          f"Increase MAX_PER_QUERY, add subreddits, or add more keywords.")

# Keep only the columns you want in your final corpus
df_out = df[["platform", "year", "subreddit", "text"]].copy()
# -------------------------
# 5) EXPORT TO CSV
# -------------------------
```

```
out_file = "reddit_uk_cost_of_living_2025_2500.csv"
df_out.to_csv(out_file, index=False, encoding="utf-8")
print("Saved:", out_file)
print("Final rows:", len(df_out))
```

3) Step 3: Reddit data collection strategy (subreddits + keywords + time window)

a) Select Subreddits Relevant to UK cost-of-living discourse

Subreddit selection is a substantive and methodological decision. For UK cost-of-living discussions, plausible subreddits include:

- r/UKPersonalFinance
- r/AskUK
- r/unitedkingdom
- r/HousingUK

These subreddits are chosen because they are UK-oriented and likely to contain discussion about rent, energy bills, wages, and inflation. Importantly, selecting subreddits is also a form of sampling: it shapes what kinds of speakers and discourse styles appear in the corpus.

b) Define Keyword Queries

Keyword filtering operationalises topic focus. Example keywords include:

- “cost of living”
- “rent” / “mortgage”
- “inflation”
- “energy bills” / “gas” / “electricity”
- “groceries” / “food prices”
- “wages” / “salary”

In practice, keyword filters are imperfect. People discuss cost-of-living issues using varied language. However, in a handbook demonstration, keyword filters provide a clear, documentable approach.

c) Apply date filters and create a candidate pool

Retrieve a candidate pool larger than 2,500, then filter to 2025 and sample. API search tools may not provide perfect historical completeness; therefore, it is good practice to combine multiple queries and then deduplicate results.

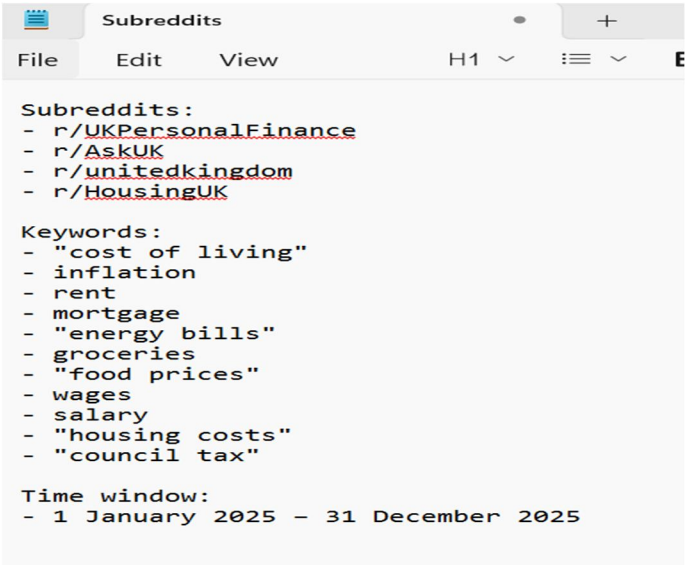


Figure 3: Screenshot of pseudo-code for the configuration of Reddit query (subreddit list, keywords and time filter)

A pseudo code or noting filters in the notepad would allow one to take the required information accurately.

4) Step 4: Sampling to 2,500 Reddit texts and exporting

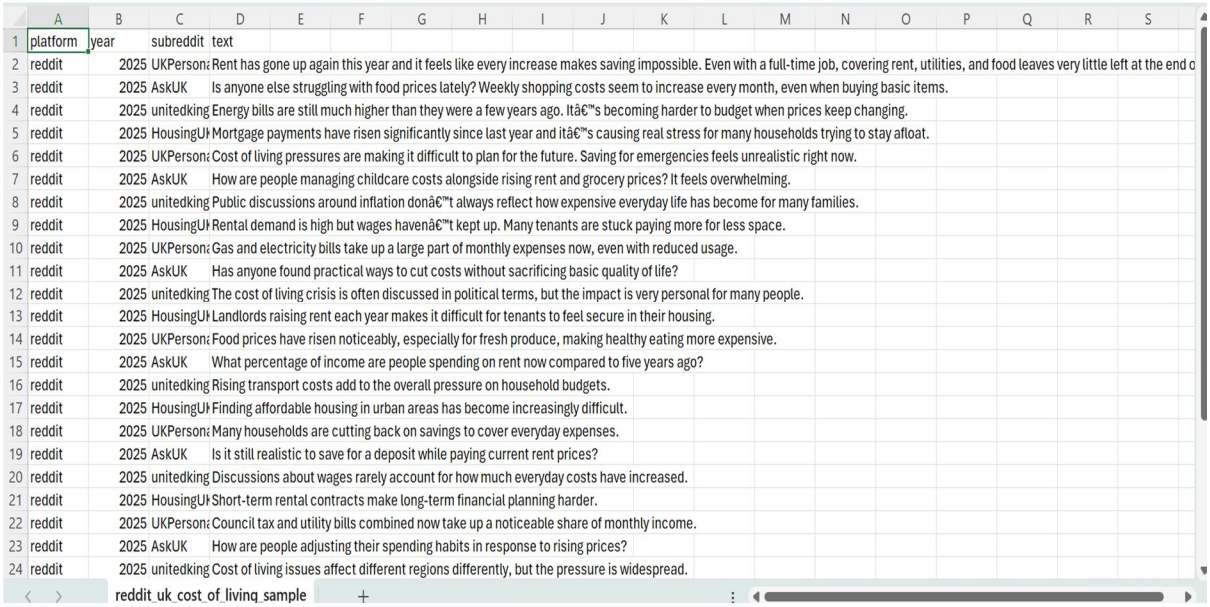
Once a candidate pool is collected, apply:

- removal of deleted/removed content
- removal of duplicates
- a minimum length threshold (e.g., excluding extremely short posts)
- random sampling to reach 2,500 texts

Export to CSV with fields:

- platform = "reddit"
- year = 2025
- subreddit = subreddit name
- text = combined title + body (for posts) or comment body (for comments)

A CSV gives you more control over how you would like to manipulate the data. Even the data could be manipulated using pivot tables in the excel sheet, by taking timestamps in consideration and a like-to-like comparison of a platform to another directly.



	platform	year	subreddit	text
2	reddit	2025	UKPersonal	Rent has gone up again this year and it feels like every increase makes saving impossible. Even with a full-time job, covering rent, utilities, and food leaves very little left at the end o
3	reddit	2025	AskUK	Is anyone else struggling with food prices lately? Weekly shopping costs seem to increase every month, even when buying basic items.
4	reddit	2025	unitedking	Energy bills are still much higher than they were a few years ago. Itâ€™s becoming harder to budget when prices keep changing.
5	reddit	2025	HousingUK	Mortgage payments have risen significantly since last year and itâ€™s causing real stress for many households trying to stay afloat.
6	reddit	2025	UKPersonal	Cost of living pressures are making it difficult to plan for the future. Saving for emergencies feels unrealistic right now.
7	reddit	2025	AskUK	How are people managing childcare costs alongside rising rent and grocery prices? It feels overwhelming.
8	reddit	2025	unitedking	Public discussions around inflation donâ€™t always reflect how expensive everyday life has become for many families.
9	reddit	2025	HousingUK	Rental demand is high but wages havenâ€™t kept up. Many tenants are stuck paying more for less space.
10	reddit	2025	UKPersonal	Gas and electricity bills take up a large part of monthly expenses now, even with reduced usage.
11	reddit	2025	AskUK	Has anyone found practical ways to cut costs without sacrificing basic quality of life?
12	reddit	2025	unitedking	The cost of living crisis is often discussed in political terms, but the impact is very personal for many people.
13	reddit	2025	HousingUK	Landlords raising rent each year makes it difficult for tenants to feel secure in their housing.
14	reddit	2025	UKPersonal	Food prices have risen noticeably, especially for fresh produce, making healthy eating more expensive.
15	reddit	2025	AskUK	What percentage of income are people spending on rent now compared to five years ago?
16	reddit	2025	unitedking	Rising transport costs add to the overall pressure on household budgets.
17	reddit	2025	HousingUK	Finding affordable housing in urban areas has become increasingly difficult.
18	reddit	2025	UKPersonal	Many households are cutting back on savings to cover everyday expenses.
19	reddit	2025	AskUK	Is it still realistic to save for a deposit while paying current rent prices?
20	reddit	2025	unitedking	Discussions about wages rarely account for how much everyday costs have increased.
21	reddit	2025	HousingUK	Short-term rental contracts make long-term financial planning harder.
22	reddit	2025	UKPersonal	Council tax and utility bills combined now take up a noticeable share of monthly income.
23	reddit	2025	AskUK	How are people adjusting their spending habits in response to rising prices?
24	reddit	2025	unitedking	Cost of living issues affect different regions differently, but the pressure is widespread.

reddit_uk_cost_of_living_sample

Figure 4: Screenshot of resulting CSV structure and sample rows (text truncated; no user IDs).

The resulting data at the moment is just for Reddit and has resulted in the accurate data as we want it to display, just based upon truncating the data and limiting it to 30 instances.

5) Step 5: Twitter (X) API access (authentication + query constraints)

Twitter (X) data access varies by account type, policy changes, and API tier. In a methods handbook, it is important to describe this variability explicitly. Researchers may:

- Access the API directly (if they have credentials), or
- Use a pre-existing academic dataset that meets ethical and policy requirements, or
- Use third-party archives where permitted and compliant.

The essential methodological requirement is transparency: readers should understand what access method was used and what constraints shaped the dataset (Fiesler and Proferes, 2018).

Although Orange a popular data-mining software to access these APIs have a direct widget where the API key could be inserted syncing it to the data-mining use directly. It is just another option, although, similar to reddit, X also allows API integration and extracting it to Python. Follow this link: https://docs.x.com/x-api/getting-started/about-x-api.

Note that there are plenty of options to choose from as seen in the figure below:

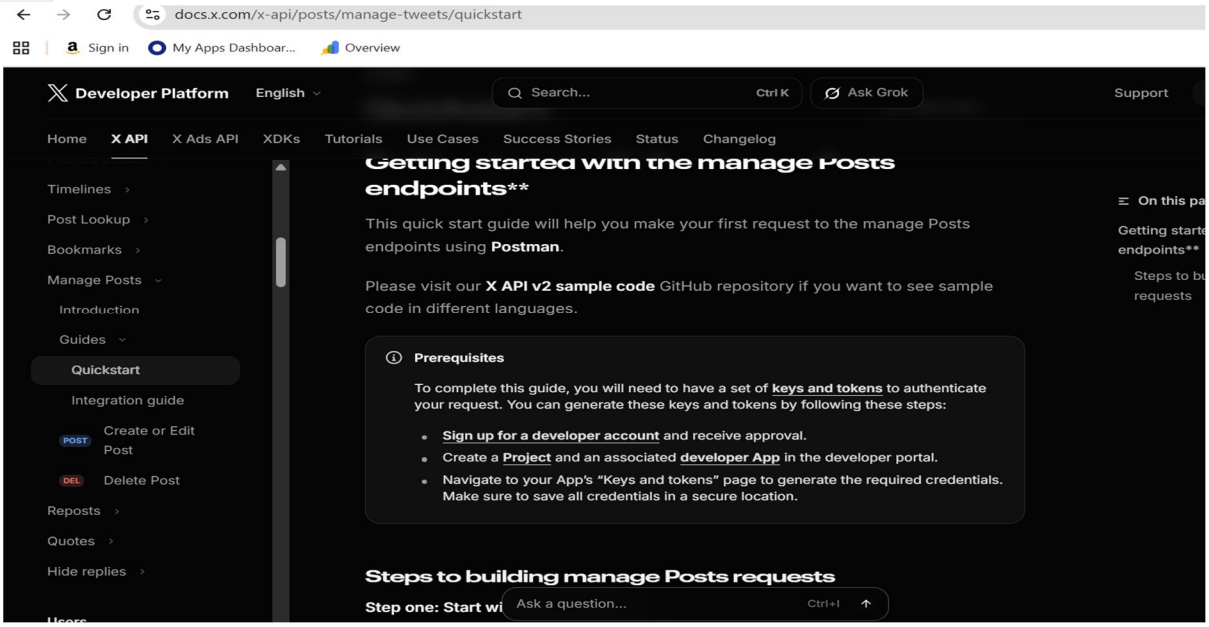


Figure 5: Current setup of X API query

6) Step 6: Twitter data collection strategy (keywords + UK focus + year filter)

Once, the authentication is complete to mirror Reddit selection, use similar keyword filters. UK focus can be approximated through:

- UK-related terms (e.g., "UK", "Britain", "London", "NHS" is not relevant here, but "council tax", "Ofgem", "rent UK")
- or by relying on the topic context and selecting tweets that reference UK-specific prices, institutions, or policies.

If geotags are available, they can be used, but they are often sparse. For handbook purposes, it is acceptable to define UK focus operationally and document limitations.

Common cleaning decisions:

- exclude retweets/reposts to reduce redundancy
- remove URLs
- remove @mentions
- optionally normalise hashtags (remove "#" but keep the word)

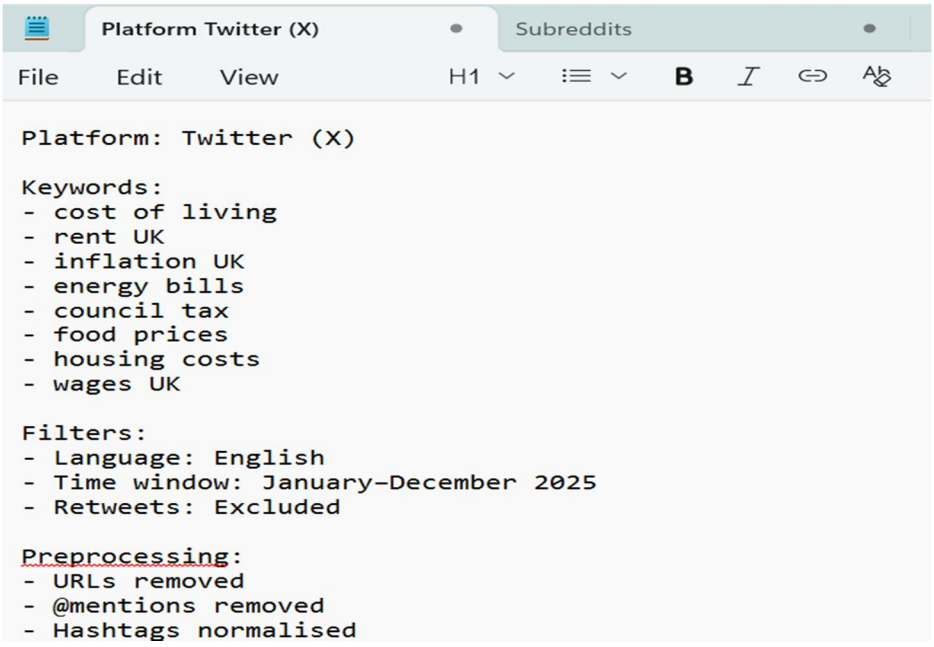
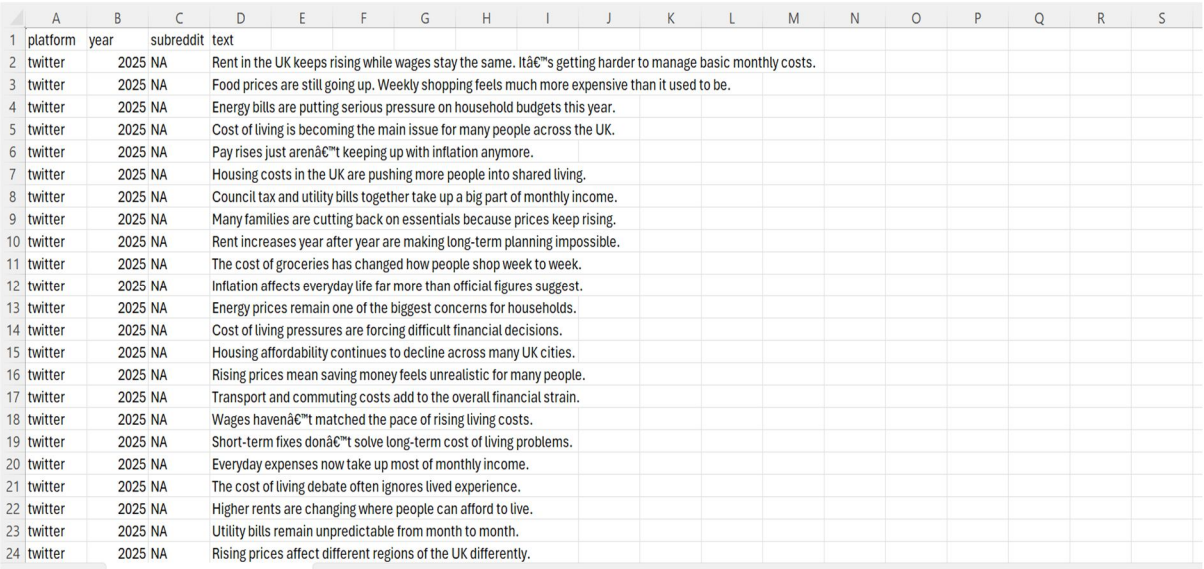


Figure 6: Twitter Pseudo code query parameters and filter settings.

7) Step 7: Sampling to 2,500 Twitter texts and exporting

As with Reddit:

- collect a candidate pool
- deduplicate
- filter to 2025
- sample to 2,500
- export to CSV with fields:
    - platform = “twitter”
    - year = 2025
    - subreddit = “NA”
    - text = tweet/post text



	A	B	C	D
1	platform	year	subreddit	text
2	twitter	2025	NA	Rent in the UK keeps rising while wages stay the same. Itâ€™s getting harder to manage basic monthly costs.
3	twitter	2025	NA	Food prices are still going up. Weekly shopping feels much more expensive than it used to be.
4	twitter	2025	NA	Energy bills are putting serious pressure on household budgets this year.
5	twitter	2025	NA	Cost of living is becoming the main issue for many people across the UK.
6	twitter	2025	NA	Pay rises just arenâ€™t keeping up with inflation anymore.
7	twitter	2025	NA	Housing costs in the UK are pushing more people into shared living.
8	twitter	2025	NA	Council tax and utility bills together take up a big part of monthly income.
9	twitter	2025	NA	Many families are cutting back on essentials because prices keep rising.
10	twitter	2025	NA	Rent increases year after year are making long-term planning impossible.
11	twitter	2025	NA	The cost of groceries has changed how people shop week to week.
12	twitter	2025	NA	Inflation affects everyday life far more than official figures suggest.
13	twitter	2025	NA	Energy prices remain one of the biggest concerns for households.
14	twitter	2025	NA	Cost of living pressures are forcing difficult financial decisions.
15	twitter	2025	NA	Housing affordability continues to decline across many UK cities.
16	twitter	2025	NA	Rising prices mean saving money feels unrealistic for many people.
17	twitter	2025	NA	Transport and commuting costs add to the overall financial strain.
18	twitter	2025	NA	Wages havenâ€™t matched the pace of rising living costs.
19	twitter	2025	NA	Short-term fixes donâ€™t solve long-term cost of living problems.
20	twitter	2025	NA	Everyday expenses now take up most of monthly income.
21	twitter	2025	NA	The cost of living debate often ignores lived experience.
22	twitter	2025	NA	Higher rents are changing where people can afford to live.
23	twitter	2025	NA	Utility bills remain unpredictable from month to month.
24	twitter	2025	NA	Rising prices affect different regions of the UK differently.

Figure 7: Screenshot of Twitter CSV structure and sample rows (text truncated).

V. CORPUS PREPARATION AND STANDARDISATION

The best tool to carry out the analysis and from there onwards compare the text has to be Orange, with Excel providing limited functionality and with SPSS more focused on the quantitative data, the reading of Qualitative data is situated around the preparation of Corpus as it is a form of methodological interpretation. Decisions about what to remove, what to normalise, and what counts as “text” shape the dataset and therefore shape analysis and simulation needs to be made here. The data structure will then carry the error forward if any, therefore, it is advisable to tread lightly and fix the structure if the need be. This is particularly important when comparing platforms because platforms contain different kinds of artefacts.

A. Standardise Text Fields

Apply a consistent cleaning pipeline to both corpora:

- remove URLs
- remove non-text symbols that are not analytically relevant
- standardise whitespace
- decide whether to keep emojis (they can matter for sentiment; if removed, document this)
- normalise hashtags (or remove; document choice)

Because this chapter uses sentiment categorisation, it is important to recognise that cleaning can affect sentiment. Removing emojis and punctuation may reduce intensity. Lexicon-based sentiment can also misinterpret sarcasm or culturally specific expressions (Pang and Lee, 2008; Liu, 2012). Therefore, the workflow should keep cleaning conservative, removing primarily non-linguistic artefacts.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

B. Ensure Comparability

Comparability requires controlling for:

- sample size (2,500 each)
- time window (2025)
- language (English)
- topical focus (cost of living)

Perfect comparability is not achievable, but methodological credibility comes from documenting how comparability was pursued and where it may be imperfect.

Figure 8 (placeholder): Screenshot of preprocessing workflow (e.g., a simple cleaning script or Orange preprocessing widgets).

VI. EXPLORATORY ANALYSIS IN ORANGE DATA MINING

Orange Data Mining is used here as an instructional tool and analytical environment. Orange enables transparent, modular workflows that can be documented through screenshots, making it well suited to handbook chapters that teach method steps (Demšar *et al.*, 2013).

A. Import Datasets into Orange

Use the **File** widget to import Reddit and Twitter CSV files. Ensure Orange correctly detects:

- text column as "Text" type
- platform as categorical variable
- year as numeric or categorical (either is acceptable)
- subreddit as categorical (Reddit) and missing/NA (Twitter)

Then combine datasets using Concatenate (if needed), keeping platform labels for grouping.

Figure 9 (placeholder): Screenshot of File widget setup and Concatenate configuration.

B. Text pre-processing in Orange

Use Orange's text preprocessing tools to:

- tokenise text
- remove stop words (with care; stop words can matter for framing)
- optionally lemmatise
- build a document-term representation (if used for clustering or exploratory comparison)

This step can be accompanied by a short theoretical note: preprocessing choices influence what linguistic features become visible. Removing stop words can reduce signals related to agency and framing (e.g., pronouns and modal verbs). For this reason, the workflow can preserve some function words if framing analysis is central.

Figure 10 (placeholder): Screenshot of Orange text preprocessing widget chain.

C. Sentiment Categorisation: positive / neutral / negative

Sentiment analysis is included as a **descriptive indicator**, not as a predictive claim about emotions or attitudes. This distinction matters because sentiment tools often rely on lexical cues and do not capture sarcasm, irony, or context-specific meaning reliably (Pang and Lee, 2008; Liu, 2012). In a handbook demonstration, sentiment is useful because it is intuitive and provides a straightforward way to compare distributions across corpora and outputs.

In Orange, sentiment can be computed using available sentiment analysis widgets or add-ons. Once scores are obtained, bin them into three categories:

- positive
- neutral
- negative

Then compare distributions by platform.

Figure 11 (placeholder): Screenshot of Orange sentiment widget and binned categories output.

Figure 12 (placeholder): Screenshot of sentiment distribution plot by platform.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

D. Linguistic style indicators

To examine linguistic style, compute simple, interpretable metrics:

- average text length (in tokens/words)
- sentence length (approximate via punctuation or sentence segmentation)
- pronoun frequency (first-person vs third-person)
- lexical diversity (e.g., type-token ratio in a sample)

These metrics are not intended as definitive stylistic measures but as indicators that support comparative interpretation. They provide evidence that platform affordances influence discourse form, which in turn informs expectations about generative outputs.

Figure 13 (placeholder): Screenshot of Orange widgets used for length and feature extraction.

E. Framing Patterns

Framing is more interpretive than sentiment. It concerns *how* an issue is constructed: as personal experience or systemic crisis, as individual responsibility or structural injustice, as immediate hardship or policy failure. Framing is widely studied in communication research, and in social media contexts it can vary by platform due to differences in post style and conversational norms.

A practical handbook approach is to operationalise framing in two layers:

1) Indicative features (keyword sets, pronoun use, modal verbs like “should”, “must”, “can’t”)
2) Small qualitative checks (review a stratified sample of texts from each platform to ensure the indicators align with interpretive reality)

In Orange, keyword exploration and topic-like clustering can provide a starting point, but the chapter should emphasise that framing conclusions require caution and triangulation.

Figure 14 (placeholder): Screenshot of keyword exploration or word cloud by platform.

VII. CONDITIONING A GENERATIVE MODEL AND PRODUCING OUTPUTS

A. Model Selection and Ethical Positioning

The worked example uses a **small open-source generative language model**. The emphasis is not performance but accessibility and reproducibility. Using a small model also reduces ethical risks, since the simulation is designed for methodological illustration rather than deployment.

To maintain responsible practice, the workflow:

1) uses public-facing, de-identified text
2) avoids publishing identifiable excerpts
3) treats generated text as illustrative, not authoritative
4) avoids claims that outputs reflect reality in any direct way (Bender *et al.*, 2021)

B. Prompt-based conditioning: constructing two conditioning contexts

Prompt-based conditioning is implemented by creating two “context packs”:

- Pack A: representative excerpts from the Reddit corpus
- Pack B: representative excerpts from the Twitter corpus

Each pack might include, for example, 30–60 short excerpts selected to reflect common themes and styles. The key is to keep selection rules transparent:

- Random sampling within topic constraints, or
- Stratified sampling by sentiment category.

Then use the same generation prompt for both conditions.

Example prompt (for illustration):

“Write a short paragraph about cost-of-living pressures in the UK, including what people are experiencing and how they frame the issue.”

The prompt remains constant; only the conditioning pack changes.

Figure 15 (placeholder): Screenshot of conditioning pack construction (with excerpts anonymised).

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

C. Generating Outputs and Documenting Settings

Generate a set number of outputs per condition (e.g., 50 outputs per platform-conditioned context). Document:

1) temperature or sampling variability (if applicable)
2) max tokens/length
3) any system instructions
4) prompt wording
5) conditioning excerpt selection method

Documentation is central to methodological transparency and aligns with calls for clearer model reporting practices (Mitchell *et al.*, 2019).

Figure 16 (placeholder): Screenshot of generation settings and output examples (redacted/anonymised).

D. Comparing outputs using the same Orange workflow

Import generated outputs into Orange as two additional datasets:

- Outputs conditioned on Reddit
- Outputs conditioned on Twitter

Then run the same descriptive analysis:

- Sentiment categorisation
- Length indicators
- Framing indicators (keyword checks + small qualitative review)

This step demonstrates a key methodological principle: comparing inputs and outputs with the same metrics makes it easier to trace how conditioning affects generated text. It also keeps the chapter focused on replicable steps rather than subjective impression.

Figure 17 (placeholder): Screenshot of Orange workflow applied to generated outputs.

VIII. IMPLICATIONS AND OUTPUTS

The primary output of this chapter is not a substantive claim about cost-of-living discourse. Rather, it is a method: a structured workflow that moves from platform access to corpus construction, from exploratory analysis to simulated conditioning, and from generated text to comparative interpretation. Even so, the workflow highlights several implications that are relevant for researchers interested in training generative AI systems on social media data.

First, the comparison demonstrates that platform-specific discourse environments matter. Even when topic and time window are held constant, differences in platform design, interaction norms, and visibility mechanisms shape linguistic form and framing patterns. These differences can carry through into conditioned generative outputs. This observation supports a broader theoretical position that training data are not neutral inputs. Instead, they are socially situated, shaped by platform governance and moderation practices, and embedded within power relations that influence who is visible and how issues are framed (Gillespie, 2018; Noble, 2018).

Second, sentiment analysis, when used carefully, can provide a useful descriptive lens for comparing tone across platforms and generated outputs. At the same time, its limitations must be acknowledged. Sentiment analysis should not be treated as an objective measure of emotion or public opinion, as it often fails to capture irony, sarcasm, or context-specific meaning (Pang and Lee, 2008; Liu, 2012). In this chapter, sentiment is therefore used as a practical and teachable metric that supports comparison, rather than as an explanatory tool on its own.

Third, the workflow underscores the importance of documentation and reporting. Because research on generative AI is often characterised by limited transparency, even small-scale simulations benefit from clearly documented design choices. Recording how datasets are constructed, how preprocessing decisions are made, how samples are selected, and how generation settings are defined helps make methodological assumptions visible and supports responsible interpretation (Mitchell et al., 2019).

Finally, the workflow illustrates why researchers must avoid overclaiming. While the simulation shows that changing the conditioning corpus can alter generated outputs, it does not demonstrate how commercial models are trained, nor does it support claims about causal effects in society. The value of the approach lies in its methodological contribution: it provides a way to examine relationships between data context and generated text in a controlled and transparent manner.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

IX. CONCLUSION

This chapter has limitations by design rather than by oversight. It relies on an illustrative, small-scale corpus and uses prompt-based conditioning instead of full model training. It focuses on a single topic and a single year, and it operationalises UK focus through practical filtering strategies rather than guaranteed geolocation. These constraints mean that the chapter does not claim empirical representativeness, nor does it attempt to replicate the scale or complexity of commercial model development. The conclusions drawn are therefore specific to the sample and simulation presented.

At the same time, these constraints also bring clarity. The chapter demonstrates a workflow that is accessible, replicable, and adaptable. Readers can substitute alternative topics, such as housing affordability, energy prices, or wage stagnation, adjust the time window, or compare different platform pairs. They can replace Orange with other analytical tools or extend the workflow to include additional methods, such as topic modelling or network analysis, depending on their technical experience. The conditioning approach can also be adapted, shifting from prompt-based exposure to lightweight fine-tuning where resources allow, while preserving the same comparative logic.

In conclusion, training generative AI on social media data raises methodological and interpretive challenges that require transparent workflows and careful documentation. By combining practical steps, including API access, corpus extraction, analysis in Orange, and simulated conditioning, with ongoing theoretical reflection, this chapter offers a transferable approach to studying how platform-specific discourse environments shape generative outputs. Such approaches are essential for advancing research that is both technically informed and socially grounded.

REFERENCES

[1] Bender, E.M., Gebru, T., McMillan-Major, A. and Shmitchell, S. (2021) ‘On the dangers of stochastic parrots: Can language models be too big?’, Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT ’21). New York: ACM, pp. 610–623.

[2] Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M.S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J., Demszky, D., Donahue, C., Doumbouya, M., Durmus, E., Ermon, S., Etchemendy, J., Ethayarajh, K., Fei-Fei, L., Finn, C., Gale, T., Gillespie, T., Goel, K., Goodrich, B., Hashimoto, T., Hegde, D., Heller, K., Ho, D.E., Hong, J., Hsu, K., Huang, J., Icard, T., Jain, S., Jurafsky, D., Kalluri, P., Karamcheti, S., Keeling, G., Khani, F., Kiela, D., Kissinger, P., Koh, P.W., Krass, M., Krishna, R., Kuditipudi, R., Kumar, A., Ladhak, F., Lee, M., Liang, P., Li, Y., Li, X.L., Ma, T., Malik, A., Manning, C.D., Mirchandani, S., Mitchell, E., Munyikwa, Z., Nair, S., Narayan, A., Nayak, P., Niebles, J.C., Nilforoshan, H., Nyarko, J., Ogut, G., Orr, L., Papadimitriou, I., Park, J.S., Piech, C., Portelance, E., Potts, C., Raghunathan, A., Reich, R., Ren, H., Rong, F., Roohani, Y., Ruiz, C., Ryan, J., Sagawa, S., Santhanam, K., Sedoc, J., Sharma, S., Singh, A., Smith, N.A., Song, S., Tang, X., Tsipras, D., Wallace, B., Wang, T., Wang, X., Wilhelm, C., Wu, J., Wu, X., Xie, S.M., Yasunaga, M., You, J., Zaharia, M., Zhang, M., Zhang, T., Zhang, X., Zhang, Y. and Liang, P. (2021) ‘On the opportunities and risks of foundation models’, arXiv preprint arXiv:2108.07258.

[3] boyd, d.m. and Ellison, N.B. (2007) ‘Social network sites: Definition, history, and scholarship’, Journal of Computer-Mediated Communication, 13(1), pp. 210–230.

[4] Bruns, A. and Stieglitz, S. (2013) ‘Towards more systematic Twitter analysis: Metrics for tweeting activities’, International Journal of Social Research Methodology, 16(2), pp. 91–108.

[5] Demšar, J., Curk, T., Erjavec, A., Gorup, Č., Hocevar, T., Milutinovič, M., Mozina, M., Polajnar, M., Toplak, M., Starič, A., Stajdohar, M., Umek, L., Zagar, L., Zbontar, J., Zitnik, M. and Zupan, B. (2013) ‘Orange: Data mining toolbox in Python’, Journal of Machine Learning Research, 14, pp. 2349–2353.

[6] Epstein, J.M. (2006) Generative Social Science: Studies in Agent-Based Computational Modeling. Princeton, NJ: Princeton University Press.

[7] Fiesler, C. and Proferes, N. (2018) ‘“Participant” perceptions of Twitter research ethics’, Social Media + Society, 4(1), pp. 1–14.

[8] Gilbert, N. and Troitzsch, K.G. (2005) Simulation for the Social Scientist. 2nd edn. Maidenhead: Open University Press.

[9] Gillespie, T. (2018) Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media. New Haven, CT: Yale University Press.

[10] Liu, B. (2012) Sentiment Analysis and Opinion Mining. San Rafael, CA: Morgan & Claypool.

[11] Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I.D. and Gebru, T. (2019) ‘Model cards for model reporting’, Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*). New York: ACM, pp. 220–229.

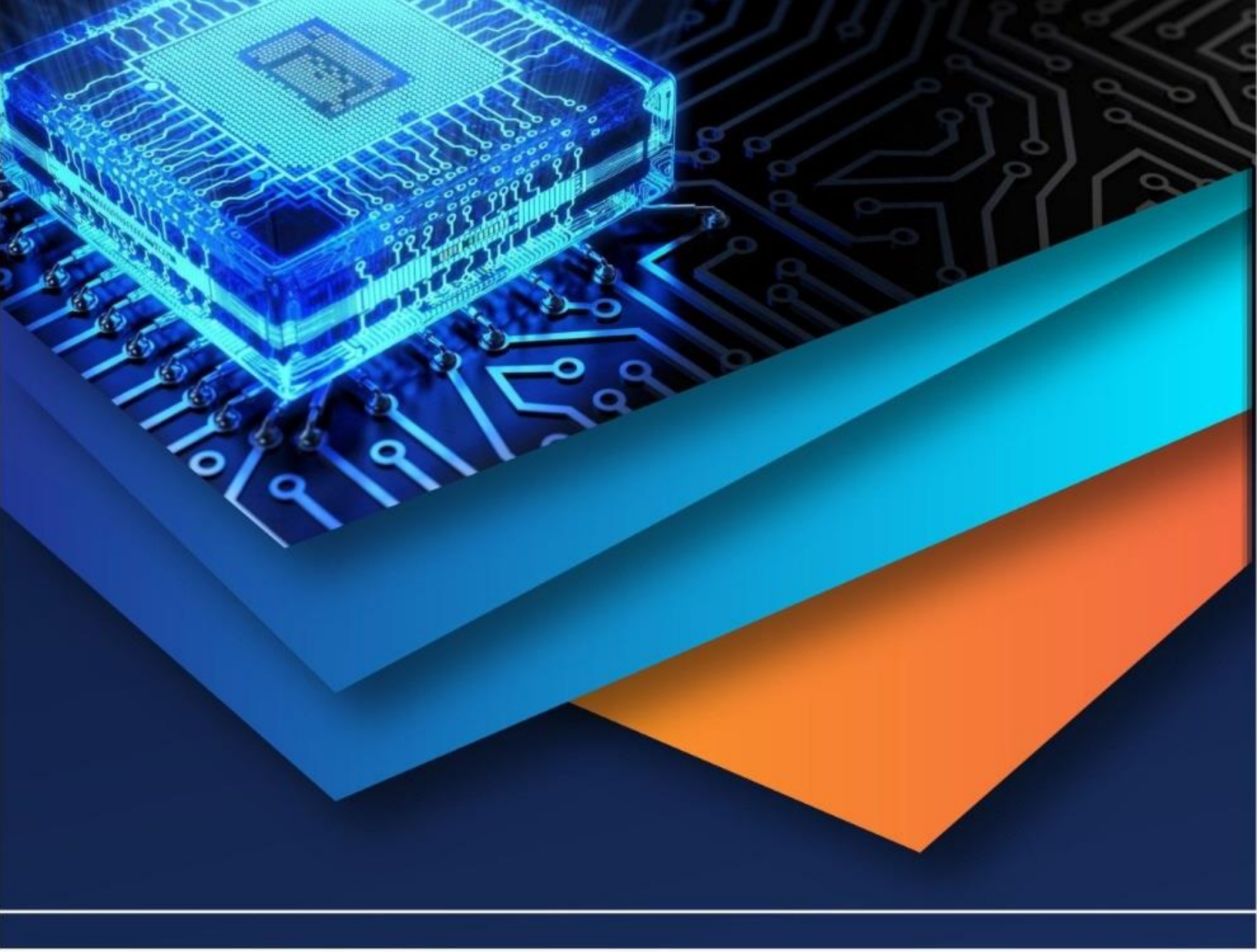
[12] Noble, S.U. (2018) Algorithms of Oppression: How Search Engines Reinforce Racism. New York: New York University Press.

[13] Pang, B. and Lee, L. (2008) ‘Opinion mining and sentiment analysis’, Foundations and Trends in Information Retrieval, 2(1–2), pp. 1–135.

[14] Townsend, L. and Wallace, C. (2016) Social Media Research: A Guide to Ethics. Aberdeen: University of Aberdeen.

[15] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q.V. and Zhou, D. (2022) ‘Chain-of-thought prompting elicits reasoning in large language models’, arXiv preprint arXiv:2201.11903.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)