Thomas Compton CV

LinkedIn Medium GitHub

Statement

Interdisciplinary NLP researcher with a unique foundation in social science, English, and computer science. Experienced in applying NLP methods to large-scale textual corpora, including topic modelling, OCR pipelines, and discourse analysis. Strong communicator with a proven track record in undergraduate teaching and outreach across Maths, English, and Social Policy, spanning primary to university level.

Experience

School for Business and Society
Graduate Teaching Assistant & Research Assistant
May 2023 – Aug 2025 | University of York

- Organised and interviewed 10 practitioners to understand their approaches to study EDI policies.
- Created chatbot to allow staff to understand EDI policy changes, using staff survey data.
- Delivered seminars to undergraduate students receiving positive feedback in evaluations.

Many Flavours of Wellingborough CIC
Founder & Director Sept 2023 – Sept 2024 | Wellingborough

- Founded, registered and managed Many Flavours of Wellingborough CIC, including developing safeguarding, data protection, EDI and other policies.
- Advertised and managed volunteers for a youth drama group with 11 regular students.
- Secured funding from Made With Many (£750), Wellingborough Town Council (£1000), Asda (£586), Cinema4all (£440).

Projects

OCR Accuracy Evaluation: GitHub

- Compare WER & CER of different OCR pipelines.
- Develop batch approach to OCR with Gemini to avoid rate limits.
- **Skills:** EasyOCR, PaddleOCR, Tesseract, Google Gemini, Jiwer, Deepseek, Qwen.

BERTopic Evaluation and Fine-Tuning: GitHub

- Developed metrics for evaluating multiple BERTopic runs and exploring trade-offs.
- Created batch method to select the best topic model from multiple runs.
- **Skills:** BERTopic, Sentence Transformers, scipy, matplotlib, UMAP.

Education

University of York, 2021 – 2025 (Expected)

- 1+3 ESRC Funded PhD in NLP approaches to large historical corpora.
- Thesis: *Historical (dis)Continuities in Community Unionism Between The Boot and Shoe Union and Unite Community.*

University of Birmingham, 2016 – 2020

- MSc Distinction, International Relations.
- BA 2:1, English and Creative Writing.

Skills

- **Programming:** Python (3+ years)
- **NLP & ML:** spaCy, NLTK, FAISS, BERTopic, LDA, Sentence Transformers, PyTorch, Hugging Face, scikit-learn
- **LLMs:** Google Gemini, OpenAI, Qwen, TrOCR
- **Data & Tools:** pandas, matplotlib, requests, AWS, Docker, GitHub, Gradio, SQL, PostgreSQL
- **Visualization:** UMAP, Pandas, scipy

Open Source Projects

Literature Review: GitHub

- Modular tools for literature reviews using RAG, FAISS, and sentence embeddings.
- **Skills:** RAG, FAISS, Sentence Transformers, SpaCy, sklearn.

PDF-to-Speech Reader: GitHub

- Converts PDFs into audio using text-to-speech.
- **Skills:** pypdf, pyttsx3, pydub, AWS.

Hansard OCR Dataset: GitHub

- Prepared historical UK parliamentary debates for OCR training.
- Finish pipeline to be used in upcoming projects with Northampton Lab, ST Mary's Church Rushden's Archive.
- **Skills:** TrOCR, torch, BeautifulSoup, transformers, SQL, postgres

Samuel Smiles DA: GitHub

- Educational repo for historical text analysis using embeddings and topic modelling.
- **Skills:** sentence-transformers, FAISS, BERTopic, spaCy, nltk.

ngram-2-topic: GitHub

- Combines n-gram frequency with sentence embeddings for interpretable topics.
- **Skills:** sentence-transformers, spaCy.

Publications

Articles

Compton, T. (2025). Holistic evaluations of topic models. https://arxiv.org/abs/2507.23364

Conferences

- BUIRA Conference (accepted), 2025, *Community Unionism Reconsidered*
- BUIRA Conference (accepted), 2024, *Northampton Boot & Shoe Then & Now*
- BSA Work, Employment and Society Conference (accepted), 2023, poster *Northampton Boot & Shoe Then & Now*
- PSA Conference 2023, *Humanitarian Intervention*

Thomas Compton Portfolio

thomasocompton@outlook.com | GitHub |
arXiv:2507.23364

Introduction

- Interdisciplinary NLP researcher applying computational methods to historical and policy texts. Focus on OCR, topic modelling, and RAG-based analysis.
- Where appropriate, the GitHub README files include Google Colab links, allowing the code to be run online.

Projects

9.1 Historical Union Debate

GitHub

- Uses RAG to simulate a dialogue between two databases. A FAISS index enables semantic search over generated debates.
- Skills: Google Gemini, RAG (Langchain), FAISS, Gradio
- Research Output: blog
- Live demo: Hugging Face Space

9.2 OCR Accuracy Evaluation for Image-Based Text Extraction

GitHub

- Compare WER & CER across OCR+LLM pipelines.
- Developed batch processing to bypass Gemini API rate limits.
- Skills: EasyOCR, PaddleOCR, Tesseract, Google Gemini, Jiwer, Deepseek, Qwen
- Research output GitHub: OCR-evaluation
- blog

Table 1: OCR + LLM Accuracy (Lower = Better)

Engine	**WER**	**CER**	**LLM**
Gemini 2.0 Flash	0.04	0.02	Yes
Qwen3-235B-A22B	0.06	0.03	Yes
Deepseek-V3-R1	0.29	0.26	Yes
ChatGPT-4o	0.58	0.45	Yes
Tesseract	0.69	0.43	No
PaddleOCR	0.79	0.76	No
EasyOCR	0.89	0.67	No

WER = Word Error Rate, CER = Character Error Rate. Evaluated using `jiwer` against ground-truth text.

9.3 BERTopic Evaluation and Fine-Tuning

Jan 2024 – Present | University of York | Project Lead

- Developed metrics for evaluating multiple BERTopic runs and exploring trade-offs (e.g., topic inequality vs. sentence coverage).
- Implemented randomized parameter search to generate diverse topic models.
- Research output: blog, article

Code snippet: Randomized BERTopic parameter search

```python
for i in range(60):
    print(f"\nIteration {i+1}")

    min_cluster_size = random.choice(range(5, 80, 3))
    min_topic_size = random.choice(range(5, 80, 3))

    hdbscan_model = HDBSCAN(
        min_cluster_size=min_cluster_size,
        metric='euclidean',
        cluster_selection_method='eom',
        prediction_data=True
    )
    umap_model = UMAP(
        n_neighbors=random.choice([5,10,15]),
        n_components=5,
        min_dist=0.0,
        metric='cosine',
        random_state=42
    )
    vectorizer_model = CountVectorizer(
        stop_words="english",
        min_df=2,
        ngram_range=(1, 2),
        token_pattern=r"(?u)\b\w{3,}\b"
    )

    topic_model = BERTopic(
        embedding_model=None,
        umap_model=umap_model,
        hdbscan_model=hdbscan_model,
        vectorizer_model=vectorizer_model,
        min_topic_size=min_topic_size,
        nr_topics=50,
        top_n_words=10,
        verbose=False
    )

    topics, probs = topic_model.fit_transform(all_sentences, embeddings)
    topic_info = topic_model.get_topic_info()
    error_size = topic_info['Count'].iloc[0]

    if error_size in error_sizes:
        continue

    if len(topic_info) > 19:
        topic_info_filtered = topic_info[topic_info["Topic"] != -1]
        top_df = topic_info_filtered.nlargest(20, "Count").copy()
        topic20_size = topic_info['Count'].iloc[19]
    else:
        top_df = pd.DataFrame()
        topic20_size = 0

    if not top_df.empty:
        topic_counts = top_df["Count"].values
        gini_score = gini(topic_counts)
        ngram_value = 0
        for topic_id in top_df['Topic'][:20]:
            for word, _ in topic_model.get_topic(topic_id):
                if word in term_freq_dict:
                    ngram_value += term_freq_dict[word]
    else:
        gini_score = 0
```

Graph: Trade-off between sentence coverage and topic inequality (Gini score).

9.4 Literature Review Tools

GitHub

- Modular tools for literature reviews using Python.
- Skills: RAG, FAISS, Sentence Transformers, SpaCy, NLTK, sklearn (cosine similarity, TfidfVectorizer)

9.5 PDF-to-Speech Reader

GitHub

- Converts PDFs to audio using text-to-speech.
- Skills: pypdf, pyttsx3, pydub, AWS

9.6 Hansard OCR Dataset

GitHub

- Curated dataset of historical UK parliamentary debates for OCR training.
- Skills: torch, TrOCR, BeautifulSoup, transformers

```
def parse_args():
    parser =
    ↪   argparse.ArgumentParser(description="Train
    ↪   a TrOCR OCR model on a preprocessed
    ↪   dataset")
    parser.add_argument("--dataset_path",
    ↪   type=str, required=True,
                        help="Path to the
                        ↪   preprocessed dataset")
    parser.add_argument("--output_dir", type=str,
    ↪   default="./trocr-ocr-model",
                        help="Directory to save
                        ↪   trained model and
                        ↪   logs")
    parser.add_argument("--batch_size", type=int,
    ↪   default=8,
                        help="Batch size per
                        ↪   device during
                        ↪   training")
    parser.add_argument("--num_train_epochs",
    ↪   type=int, default=5,
                        help="Number of training
                        ↪   epochs")
    parser.add_argument("--fp16",
    ↪   action="store_true",
                        help="Use FP16 precision
                        ↪   if available")
    parser.add_argument("--save_strategy",
    ↪   type=str, default="epoch",
                        choices=["no", "epoch",
                        ↪   "steps"],
                        help="When to save model
                        ↪   checkpoints")
    parser.add_argument("--model_name", type=str,
    ↪   default="microsoft/trocr-base-stage1",
                        help="Base model to
                        ↪   fine-tune")
    return parser.parse_args()
```

Design focus: User-friendly CLI for efficient OCR model training.

9.7 Samuel Smiles DA

GitHub

- Educational repo for collecting, embedding, clustering, and analyzing historical texts.
- Skills: sentence-transformers, faiss-cpu, bertopic, nltk, spacy, scikit-learn, matplotlib

9.8 ngram-2-topic

GitHub

- Combines n-gram frequency and sentence embeddings to generate interpretable topics.
- Skills: sentence-transformers, spaCy

Table evaluating the impact of changing cosine similarity threshold on Topic distributions (Gini), topic overlap (inverted Pairwise Uniqueness Values), and Topic Model Corpus Coverage (Percentage Appearance, which is the number of sentences covered by the model compared to the total sentences)

Table 2: Selected Topic Model Metrics

Gini	**PUV**	**% Appearance**	**Threshold**
0.64	0.18	72.75	0.40
0.62	0.18	70.19	0.41
0.60	0.18	67.37	0.42
0.59	0.19	64.37	0.43
0.56	0.18	61.12	0.44
0.55	0.21	57.79	0.45
0.53	0.21	54.39	0.46
0.52	0.21	50.93	0.47
0.52	0.21	47.39	0.48
0.51	0.21	43.79	0.49