A Bilingual Deep Learning Framework for Identifying Gender Impersonation on Social Media via Text, Image, and Metadata Fusion

By E. Mohammad Omar Mahairi

Syrian Virtual University – Department of Web Science
Supervised by: Dr. Basel Alkhateb (Director of the University Program)
June 2025

Abstract

With the proliferation of fake and impersonator profiles on social media, identifying malicious actors, especially those impersonating female identities, has become a critical challenge. This thesis proposes a bilingual deep learning framework that combines gender classification with multimodal signals to detect gender impersonators on Facebook. The system uses MARBERT for Arabic text and BERTweet for English text, fused with auxiliary scores from profile image analysis (reverse image search + EXIF metadata) and Instagram link detection as a social authenticity signal. The final decision layer combines all scores using a weighted or meta-classifier strategy. The framework achieves strong F1 scores across both languages, demonstrating the value of integrating diverse signals for impersonation detection.

Acknowledgments

I would like to express my deepest gratitude to my supervisor, **Dr. Basel Alkhateb**, for his guidance, support, and continuous encouragement throughout this work. I also thank my family, colleagues, and the Syrian Virtual University's Department of Web Science for their invaluable support during my thesis journey.

العنوان باللغة العربية

إطار عمل ثنائي اللغة قائم على التعلم العميق للكشف عن انتحال الهوية الجنسية على وسائل التواصل الاجتماعي باستخدام دمج النصوص والصور والبيانات الوصفية

1. Introduction

Gender impersonation, particularly the act of men impersonating women on social media, poses serious risks ranging from emotional manipulation to financial fraud. While prior efforts have focused on either text-based gender classification or image analysis, few have integrated these into a holistic impersonation detection framework. This thesis addresses this gap by proposing a multimodal detection system leveraging language-specific deep learning models, image analysis, metadata extraction, and social verification cues.

2. Related Work

2.1 Gender Classification

Previous research on gender detection has largely focused on stylometric and deep learning models trained on social media or blog corpora. The PAN 2017 Author Profiling dataset and the Blog Authorship Corpus are common benchmarks. Deep models like BERT, MARBERT, and BERTweet have outperformed traditional ML techniques by capturing nuanced linguistic features.

2.2 Impersonation Detection

Research on social media impersonation is limited. Some works utilize reverse image search or profile feature extraction (e.g., username similarity, account age), but lack robust multimodal integration or focus on gender-based deception.

2.3 Multimodal Detection

Multimodal approaches combine text, image, and metadata (EXIF, social links) for enhanced detection. Some systems use reverse image search to detect stock or stolen images and EXIF metadata to infer camera authenticity. Instagram link detection can serve as a proxy for account authenticity, especially in female-presenting profiles.

2.4 Instagram Link Detection in Facebook Profiles

Few studies analyze the presence of cross-platform social signals. We introduce a novel component that scrapes Facebook profile data to detect references to Instagram, using both HTML scraping and text-based regex matching. This signal improves the model’s ability to distinguish between real and impersonator accounts.

3. Methodology

3.1 Dataset Overview

- **Arabic PAN 2017**: 1,620 samples (tweets labeled by gender).
- **English PAN 2017**: 3,600 samples.
- **Blog Authorship Corpus**: 20,000 labeled blog posts (balanced by gender).

3.2 Preprocessing

All datasets underwent standard cleaning, deduplication, and balancing. For the Blog Corpus, we used a filtered subset of ~20k entries. Tokenization was done using model-specific tokenizers:

- **Arabic**: MARBERT tokenizer with truncation and padding.
- **English**: BERTweet tokenizer (`use_fast=False`, normalization enabled, max length 128).

3.3 Text Classification Models

- **Arabic**: Fine-tuned MARBERT on PAN 2017 Arabic data.
- **English**: Fine-tuned `vinai/bertweet-base` on PAN 2017 English + Blog Authorship Corpus.

3.4 Instagram Link Detection

A hybrid module combining:

- **Scraping**: Searches for `instagram.com` in Facebook profile pages using `BeautifulSoup`.
- **Regex Matching**: Detects handles like `@username`, mentions of "insta", or international variants (e.g., "انستغرام").

Returns a **social_score** (1.0 if detected, else 0.0).

3.5 Image & EXIF Analysis

- **Reverse Image Search** via SerpAPI to detect stock or reused images.
- **EXIF Metadata Scoring**: Extracts `Make`/`Model` fields using `PIL` to assess authenticity.

3.6 Decision Layer

Combines four scores:

- `text_score`, `image_score`, `exif_score`, `social_score`

Using a weighted fusion formula:

```
python
CopyEdit
```

```
final_score = 0.30 * text + 0.40 * image + 0.15 * exif + 0.15 * social
```

If `final_score ≥ 0.5`, classify as `real_female`, else `impersonator`.

Optional: Meta-classifier (Logistic Regression) trained on validation data for dynamic weighting.

3.7 QR Code Repository Link

🔗 Scanning this links to the GitHub repository containing all training scripts, evaluation logs, and sample data.

4. Experiments and Results

4.1 Arabic Model (MARBERT)

Epoch	Train Loss	Val Loss	Accuracy	F1	Confusion Matrix
1	0.6009	0.5240	0.7458	0.7188	[[101, 23], [38, 78]]
2	0.4560	0.7566	0.7208	0.6378	[[114, 10], [57, 59]]
3	0.2401	0.8895	0.7583	0.7642	[[88, 36], [22, 94]]
4	0.2403	1.4520	0.7333	0.6831	[[107, 17], [47, 69]]
5	0.1170	1.3558	0.7583	0.7456	[[97, 27], [31, 85]]

Best Model: Epoch 3
Final Evaluation:

- Accuracy: 76.66%
- F1: 0.7862
- Precision: 0.7357
- Recall: 0.8443

4.2 English Model (BERTweet)

Best at Epoch 3:

- Accuracy: 71.18%
- F1: 0.7202
- Precision: 0.7090
- Recall: 0.7316

Male Precision/Recall: 0.7149 / 0.6916
Female Precision/Recall: 0.7091 / 0.7316

5. Discussion

5.1 Interpretation of Results

Arabic MARBERT model outperformed English, likely due to the more focused PAN 2017 dataset vs. diverse Blog corpus.

5.2 Effectiveness of Decision Layer

The integration of `image_score`, `exif_score`, and `social_score` significantly improved the impersonation detection.

5.3 Practical Considerations

- Instagram link detection has limitations with private/logged-in profiles.
- Reverse image search requires external APIs or browser automation.

5.4 Limitations

- Datasets are platform-specific (e.g., Facebook/Twitter), which may limit generalization.
- EXIF data is often stripped from images.

5.5 Recommendations

- Augment training data using Masader or QADI (post-hydration).
- Explore multilingual joint training models like XLM-R for future expansion.

6. Conclusion and Future Work

This thesis presented a bilingual, multimodal framework for detecting gender impersonators on social media using a fusion of text classification, image analysis, EXIF metadata, and Instagram link detection. The system achieved strong performance in both Arabic and English, with the weighted decision layer significantly improving classification robustness.

Future Work:

- Merge MARBERT and BERTweet into a unified multilingual transformer.
- Apply the system to live profile monitoring tools.
- Use larger-scale multimodal datasets for fine-tuning.

References

1. Devlin et al., “BERT: Pre-training of Deep Bidirectional Transformers,” NAACL 2019.
2. Abdul-Mageed et al., “MARBERT: Large-scale Arabic Language Model,” EMNLP 2021.
3. Nguyen et al., “BERTweet: A pre-trained language model for English Tweets,” NAACL 2020.
4. Rangel et al., “Overview of the PAN 2017 Author Profiling Task,” CLEF 2017.
5. Argamon et al., “Gender, genre, and writing style in formal written texts,” Text, 2003.
6. Blog Authorship Corpus – http://u.cs.biu.ac.il/~koppel/BlogCorpus.htm
7. Masader: Arabic NLP Datasets Repository – https://masader.ai
8. SerpAPI Documentation – https://serpapi.com
9. Hugging Face Transformers – https://huggingface.co
10. BeautifulSoup: HTML Parsing – https://www.crummy.com/software/BeautifulSoup/

Appendices

A. Sample Entries – PAN 2017 (English)

text	label
“oif chief warrant officer …”	1
“urllink bush urges tax …”	0

B. Sample Entries – Blog Corpus

text	label
“wgf gn bq xfojb …”	1

C. Sample Entries – Arabic PAN

text	label
“قامت على طاولتها …”	1

D. Decision Layer Code

Python:

```
def classify_profile(text_score, image_score, exif_score, social_score,
weights=None, threshold=0.5):
    if weights is None:
        weights = {"text": 0.30, "image": 0.40, "exif": 0.15, "social": 0.15}
    final_score = (
        weights["text"] * text_score +
        weights["image"] * image_score +
    weights["exif"] * exif_score +
    weights["social"] * social_score
)
decision = "real_female" if final_score >= threshold else "impersonator"
return final_score, decision
```