PEOPLE'S DEMOCRATIC REPUBLIC OF ALGERIA

MINISTRY OF HIGHER EDUCATION AND SCIENTIFIC RESEARCH

FERHAT ABBES UNIVERSITY of SETIF 1



Master's Thesis

Submitted in Partial Fulfilment of the Requirements for the Academic Master's Degree

Field: Computer Science

Specialty: Image processing and data engineering

Title:

BERT APPROACHES FOR FAKE NEWS DETECTION

Presented by:	Under the supervision of:
LACHEHEUB Samy Ismail	Dr. Ahlem DRIF
SIDHOUM Dhia Eddine	

Academic Year 2025 – 2026

Abstract

The rapid growth of fake news across digital platforms poses serious risks to public discourse and democratic stability, particularly during geopolitical crises in which false reports spread virally before verification can occur. This thesis presents a hybrid deep learning framework integrating BERT with Large Language Models (LLMs) to enhance fake news detection across English domain.

The proposed methodology fine-tunes RoBERTa for binary classification on the WELFake dataset — comprising 72,134 labelled English news articles — and evaluates the resulting system on a curated subset of FakeNewsNet PolitiFact under zero-shot cross-domain conditions. LLMs contribute through zero-shot prompting to generate structured authenticity features, including manipulation intensity scores, semantic vagueness assessments, and numerical evidence indicators, which are fused with RoBERTa's contextual embeddings via a gated late-fusion mechanism.

Experimental results demonstrate that the hybrid architecture consistently outperforms the RoBERTa-only baseline across all classification metrics, achieving accuracy and weighted F1-score improvements of approximately two percentage points on the in-domain test set, with stronger relative improvements under cross-domain evaluation. An ablation study confirms that both the gated fusion architecture and the LLM inference-time scores contribute independently to the observed gains.

Keywords: *hybrid BERT–LLM framework, fake news detection, WELFake dataset, Arabic Fake News Dataset, zero-shot prompting, transformer models, misinformation detection, RoBERTa, gated fusion.*

Résumé

La prolifération rapide des fausses informations sur les plateformes numériques représente une menace sérieuse pour le débat public et la stabilité démocratique, particulièrement lors des crises géopolitiques où des informations erronées se propagent viralement avant toute vérification. Ce mémoire présente un cadre d'apprentissage profond hybride intégrant BERT et les grands modèles de langage (LLMs) afin d'améliorer la détection des fausses informations dans le domaine anglophone.

La méthodologie proposée affine RoBERTa pour une classification binaire sur les données WELFake — composé de 72 134 articles de presse anglais étiquetés — et évalue le système obtenu sur un sous-ensemble sélectionné de FakeNewsNet PolitiFact dans des conditions de transfert inter-domaines en zéro-shot. Les LLMs contribuent via un prompting zéro-shot pour générer des caractéristiques d'authenticité structurées, notamment des scores d'intensité de manipulation, des évaluations de vaguité sémantique et des indicateurs de preuves numériques, qui sont fusionnés avec les représentations contextuelles de RoBERTa via un mécanisme de fusion tardive à porte apprise.

Les résultats expérimentaux montrent que l'architecture hybride surpasse systématiquement le modèle de référence RoBERTa seul sur l'ensemble des métriques de classification, avec des gains d'environ deux points de pourcentage en précision et en F1-score pondéré sur l'ensemble de test en domaine, et des améliorations relatives plus marquées en évaluation inter-domaines. Une étude d'ablation confirme que l'architecture de fusion à porte et les scores LLM contribuent indépendamment aux gains observés.

Mots-clés : *cadre hybride BERT–LLM, détection de fausses informations, jeu de données WELFake, jeu de données arabe de fausses informations, prompting zéro-shot, modèles transformeurs, détection de désinformation, RoBERTa, fusion à porte.*

ملخص

يُشكّل الانتشار السريع للأخبار الكاذبة عبر المنصات الرقمية تهديدًا جسيمًا للنقاش العام والاستقرار الديمقراطي، لا سيما في أوقات الأزمات الجيوسياسية التي تنتشر فيها المعلومات المضللة بشكل متسارع قبل أن تتم عملية التحقق منها. تقدّم هذه المذكرة إطارًا هجينًا للتعلم العميق يدمج نموذج BERT مع النماذج اللغوية الكبيرة (LLMs) بهدف تعزيز قدرات الكشف عن الأخبار الكاذبة في النطاق الإنجليزي.

تعتمد المنهجية المقترحة على الضبط الدقيق لنموذج RoBERTa لأداء مهمة التصنيف الثنائي على مجموعة بيانات WELFake — التي تضم 72,134 مقالًا إخباريًا إنجليزيًا موسومًا — وتقييم النظام الناتج على مجموعة فرعية مختارة من FakeNewsNet PolitiFact في ظروف تقييم عبر النطاقات بأسلوب الصفر-shot. تُسهم النماذج اللغوية الكبيرة من خلال التوجيه الفوري بأسلوب الصفر-shot في توليد سمات موثوقية منظّمة، تشمل درجات شدة التلاعب، وتقييمات الغموض الدلالي، ومؤشرات الأدلة العددية، والتي تُدمج مع التمثيلات السياقية لـ RoBERTa عبر آلية دمج متأخر بوابي متعلَّم.

تُظهر النتائج التجريبية أن البنية الهجينة تتفوق باستمرار على النموذج المرجعي RoBERTa المنفرد عبر جميع مقاييس التصنيف، محققةً تحسينات في الدقة ودرجة F1 الموزونة بحوالي نقطتين مئويتين على مجموعة الاختبار داخل النطاق، مع تحسينات نسبية أكثر وضوحًا في التقييم عبر النطاقات. وتؤكد دراسة الاستئصال أن كلًا من بنية الدمج البوابي ودرجات النموذج اللغوي تُسهمان بشكل مستقل في المكاسب الملاحظة.

الكلمات المفتاحية: إطار هجين BERT–LLM، كشف الأخبار الكاذبة، WELFake، مجموعة بيانات الأخبار الكاذبة العربية، التوجيه الفوري بأسلوب الصفر-shot، نماذج المحوّل، كشف المعلومات المضللة، RoBERTa، الدمج البوابي.

Table of Contents

General Introduction

Research Objectives

Key Contributions

Document Organisation

Chapter 1

1.1 Introduction to Advanced NLP

1.2 Historical Development of NLP

1.2.1 Symbolic and Rule-Based Era (1950s–1980s)

1.2.2 Statistical NLP Era (1990s–2000s)

1.2.3 Neural NLP Era (2011–2017)

1.2.4 Transformer Era (2017–Present)

1.3 NLP Task Taxonomy

1.3.1 Lexical Level Processing

1.3.2 Syntactic Analysis

1.3.3 Semantic Processing

1.3.4 Discourse and Pragmatics

1.4 Text Representation Evolution

1.4.1 Static Word Embeddings

1.4.2 Contextual Embeddings

1.5 Sequence Modelling Architectures

1.5.1 Recurrent Neural Networks and Their Limitations

1.5.2 The Transformer Architecture (2017): Self-Attention and Positional Encoding

1.6 Pre-training Objectives: MLM and CLM

1.7 Multilingual NLP and Cross-Lingual Transfer

1.7.1 mBERT and Shared Subword Vocabularies

1.7.2 XLM-R and Massively Multilingual Pre-training

1.7.3 Arabic-Specific Challenges: Morphology, Dialects, and Script

1.8 Large language models (LLMs)

1.8.1 From Encoders to Decoders: Architectural Paradigm Shift

1.8.2 Defining Characteristics of LLMs: Scale, Emergence, and Instruction-Following

1.8.3 Key LLM Families and Comparative Overview

1.8.4 Prompting Paradigms

1.8.4.1 Zero-Shot Prompting

1.8.4.2 Few-Shot and In-Context Learning

1.8.4.3 Chain-of-Thought Reasoning

1.8.5 Limitations of LLMs

1.8.6 BERT vs. LLMs:

1.9 Conclusion

Chapter 2

2.1 Problem Definition

2.2 Formal Definitions and Information Disorder Taxonomy

2.3 Typology of Fake News

2.4 Real-World Consequences of Fake News

2.4.1 Political Manipulation

2.4.2 Public Health Harm

2.4.3 Economic Consequences

2.4.4 Violence and Conflict Escalation

2.5 Sources and Propagation of Fake News

2.6 Approaches to Fake News Detection

2.6.1 Rule-Based and Feature-Engineering Approaches

2.6.2 Classical Machine Learning (SVM, Naive Bayes, N-grams)

2.6.3 Deep Learning Approaches (CNN, LSTM, BiLSTM)

2.6.4 Transformer-Based Approaches

2.6.5 LLM-Augmented and Hybrid Approaches

2.7 Benchmark Datasets for Fake News Detection

2.7.1 English Datasets: ISOT and WELFake

2.7.2 Arabic Dataset: AFND

2.7.3 Zero-Shot Evaluation Set: FakeNewsNet PolitiFact

2.8 Cross-Lingual Fake News Detection

2.9 Critical Synthesis and Identified Research Gaps

Chapter 3

3.1 Problem Formulation

3.2 BERT Architecture: Pre-training and Classification

3.3 RoBERTa

3.4 DistilBERT and FakeBERT

3.5 DeBERTa, ELECTRA, ALBERT, XLNet

3.6 AraBERTv2 and CAMeLBERT

3.7 XLM-RoBERTa for Cross-Lingual Detection

3.8 Comparative Summary: BERT-Family Models

3.9 Limitations of Current BERT-Based Approaches

3.10 Proposed Framework Architecture

4.1 Introduction

4.2 Development Environment and Software Stack

4.3 Dataset Description

4.4 Text Preprocessing Pipeline

4.5 LLM Feature Extraction via Groq API

4.6 Hybrid Architecture Design

4.6.1 RoBERTa-Only (Baseline)

4.6.2 Hybrid RoBERTa + LLM

4.7 Training Configuration and Hyperparameters

4.8 Experimental Protocol and Evaluation Metrics

4.9 Training Dynamics: Loss and Accuracy Curves

4.10 Quantitative Results on WELFake Test Set

4.11 Zero-Shot Cross-Domain Generalization: FakeNewsNet PolitiFact

4.12 Ablation Study: Contribution of Each LLM Feature

4.13 Confusion Matrices and Error Analysis

4.14 Comparison with Prior Work

4.15 Discussion: Strengths, Limitations, and Failure Modes

4.16 Chapter Summary

General Conclusion

Summary of Findings

Principal Contributions

Limitations and Future Research Directions

References