People's Democratic Republic of Algeria

Ministry of Higher Education and Scientific Research

University of Ferhat Abbas Setif -1-

Department of Computer Science

In order to obtain **the Doctor Degree** in Computer science (LMD)

Field of Study: **Computer Science**

Senator Ferial

Structured Emotion Analysis from Arabic Text

Defended on: [**Month, Year**] *In front of the jury composed of*

Dr. Chabane KHENTOUT	Setif 1 University of Ferhat ABBAS	President
Pr. Abdelaziz LAKHFIF	Setif 1 University of Ferhat ABBAS	Supervisor
Dr. Chahrazed MEDIANI	Setif 1 University of Ferhat ABBAS	Co-Supervisor
Dr. Mohamed SAIDI	Setif 1 University of Ferhat ABBAS	Examiner
Pr. Samir AKHROUF	M'Sila University	Examiner
Pr. Farid NOUIOUA	B.B.A University	Examiner
Pr. Allaoua REFOUFI	Retired professor	Invited

Statement of Originality

I hereby declare that this thesis is my own original work and that all sources used have been acknowledged.

All ideas, data, and content taken from other sources have been appropriately cited and referenced.

This thesis represents the results of my independent research conducted under the supervision of **Pr. Lakhfif Abdelaziz**, in fulfillment of the requirements for the degree of **Doctor in Computer Science**.

Signed: **Date: November 2025**

Senator Ferial

Declaration

Note on Publications Included in This Thesis: At the time of submission, three chapters of this thesis are heavily based on papers submitted for publication or published in conferences and journals:

- **Chapter 7:** Senator, F., Boutouta, H., Lakhfif, A. and Mediani, C., 2023, March. Seman- tic Role Labeling of Arabic Emotional Text in Tweets. In 2023 International Conference on Advances in Electronics, Control and Communication Systems (ICAECCS) (pp. 1-6). IEEE.
- **Chapter 8:** Senator, F., Boutouta, H., Lakhfif, A., Mediani, C. (2025). The impact of semantic roles in emotion detection using the AraBERT model. In Proceedings of the Mediterranean Conference on Computer Science and Artificial Intelligence (MCCSAI 2025) (accepted, to appear). Springer.
- **Chapter 9:** Senator, F., Lakhfif, A., Zenbout, I., Boutouta, H., and Mediani, C. (2025). Leveraging ChatGPT for Enhancing Arabic NLP: Application for Semantic Role Labeling and Cross-lingual Annotation Projection. IEEE Access.

Note on Publications Not Included in This Thesis: As well as the above papers, the following works have been published during the period of research for this thesis; they have helped my deeper understanding of the work towards this thesis; however, these publications do not fit into the narrative of this thesis and have not been included in the text.

- H. Boutouta, A. Lakhfif, F. Senator, and C. Mediani, "Enhancement of Implicit Emotion Recognition in Arabic Text: Annotated Dataset and Baseline Models," *IEEE Access*, vol. 13, pp. 165096–165116, 2025, DOI: 10.1109/ACCESS.2025.3611337.

- H. Boutouta, A. Lakhfif, F. Senator, and C. Mediani, “A Transformer-based Hybrid Model for Implicit Emotion Recognition in Arabic Text”, *Eng. Technol. Appl. Sci. Res.*, vol. 15, no. 3, pp. 23834–23839, Jun. 2025, DOI: 10.48084/etasr.10261

- H. Boutouta, A. Lakhfif, F. Senator, and C. Mediani, 2025. From Context to Emotion: Leveraging LLMs for Recognizing Implicit Emotions. In: *Proceedings of the 8th International Conference on Natural Language and Speech Processing (ICNLSP-2025)*, pages 399–409, Odense, Denmark. Association for Computational Linguistics. URL: https://aclanthology.org/2025.icnlsp-1.39/

- H. Boutouta, A. Lakhfif, F. Senator, and C. Mediani, (2025, November). Do LLMs Grasp Emotion or Just Keywords? A Benchmark for Implicit Emotion Recognition. In: Proceedings of *The First IEEE/International Conference on Artificial Intelligence and Cyber Physical Systems (AICPS'25)* , Mascara, Algeria. (accepted, to appear) IEEE.

- H. Boutouta, F. Senator, Z. Lakab, C. Mediani, and A. Lakhfif, 2025. Multi-Label Emotion Recognition in Low-Resource Dialects: A Case Study on Algerian Arabic with Large Language Models. In: *ACLing 2025: 7th International Conference on AI in Computational Linguistics*, Dubai, UAE. (accepted, to appear)

Acknowledgement

In the name of Allah, the Most Gracious, the Most Merciful. All praise is due to Allah, whose blessings, guidance, and mercy have accompanied me throughout this journey and made this work possible.

I would like to express my deepest and sincerest gratitude to my supervisor, Pr. **Lakhfif AbdelAziz**, for their unwavering support, exceptional mentorship, and constant encouragement. Your guidance has shaped not only this thesis but also my way of thinking and working. Thank you for your patience, your generosity with your time, and your invaluable scientific insight.

My heartfelt thanks also go to my co-supervisor, **Dr.Mediani Chahrazed**, for their constructive feedback, continuous assistance, and commitment throughout the development of this research.

I am sincerely grateful to the members of my thesis committee, for accepting to evaluate my work and for their insightful comments, suggestions, and valuable time.

I would also like to express my profound appreciation to all my teachers, from primary school to university. Each of them has contributed, in their own way, to shaping my academic journey and nurturing my passion for learning.

I extend my warm thanks to my colleagues, friends, and research group members for their support, discussions, and encouragement, which made this journey more enriching and enjoyable.

A special mention goes to my colleagues, friends and sisters " **zenboute Imene** and **Boutouta Hanane** ", for their endless love, and belief in me. Your support has been my greatest source of strength.

Finally, I would like to thank anyone who contributed, directly or indirectly, to the realization of this thesis. Your help and presence have been invaluable.

Dedication

To my beloved father "**AL MOUHOUB**", whose strength, wisdom, and unwavering support have guided me throughout my life. Your encouragement has always pushed me to strive for excellence, and your belief in me has been a constant source of inspiration.

To my beloved mother"**Naima**", whose endless love, prayers, and sacrifices have shaped every step of my journey. Your kindness, patience, and devotion are gifts I can never repay. I owe so much of who I am to you.

To my dear husband "**Sami**", my companion and greatest supporter. Thank you for your patience, understanding, and constant presence through every challenge. Your love has given me courage, and your support has helped me reach this milestone.

To my wonderful sisters "**Rania, Chiraz, and Lidya**", for your unconditional love, your encouragement, and your ability to lift me up whenever I needed it. Your presence has always been a source of comfort and strength.

To my cherished daughter "**Assinete**" and my son "**Assil**", the lights of my life, to my sweetheart "**Watine**". Your smiles, innocence, and pure love have inspired me every single day and reminded me of the purpose behind all my efforts.

To my mother-in-law, father-in-law, and sisters-in-law, thank you for your kindness, support, and heartfelt prayers. Your encouragement has meant more to me than words can express.

To my dear friends "**Hanane, Imene, Souha, Doudi, Lamiss, Asma, and Imene**, thank you for your motivation, positivity, and companionship. Your support made the hardest moments easier and the good moments brighter.

I dedicate this work to all of you, with heartfelt gratitude and profound love.

Abstract

In the field of Natural Language Processing (NLP), emotion analysis aims to map textual content with a predefined set of human emotions, typically including joy, anger, fear, surprise, disgust, and sadness. Current state-of-the-art research mainly focuses on identifying emotions in text using categories inspired by psychological theories, such as Ekman's (1992) basic emotions. Despite the importance of emotion detection, most analyses are shallow and insufficient for tasks that require a deeper understanding of emotional meaning in context. Such applications necessitate addressing key questions, including identifying the cause that triggered the emotion (Cause), determining who experienced it (Experiencer), and more generally addressing structural questions such as who did what (Cue), to whom (Target), why (Cause), and how (Manner). This doctoral thesis aims to propose original and effective solutions to address the lack of resources and models dedicated to the structural analysis of emotions in Arabic text. To achieve this, we introduce a novel approach for analyzing the argument structure of emotions in Arabic, leveraging recent advances in Transformer-based architectures and, in particular, the capabilities of large language models (LLMs) for Arabic.

The main contributions of this thesis are multifold. The first contribution consists of the construction and annotation of the first Arabic corpus dedicated to structured emotion analysis, named 'AraERL'. The thesis also provides an in-depth examination of the impact of each semantic argument on the performance of emotion identification. In addition, it explores the use of ChatGPT for annotating Arabic texts with semantic roles and emotions through an interlingual annotation projection approach. The work further evaluates ChatGPT's ability to accurately translate English semantic and emotional annotation into Arabic. Finally, it offers a comprehensive comparison of the performance of open large language models for these tasks.

Keywords— Arabic, NLP, SRL, Cross-Lingual Annotation Projection, LLMs, ChatGPT, Emotion Analysis, Structural emotions

Résumé

Dans le domaine du traitement automatique des langues (TAL), l'analyse des émotions vise à associer un contenu textuel à un ensemble prédéfini d'émotions humaines, incluant généralement la joie, la colère, la peur, la surprise, le dégoût et la tristesse. Les recherches récentes se concentrent principalement sur l'identification des émotions dans les textes en s'appuyant sur des catégories inspirées par les théories psychologiques, telles que les émotions de base proposées par Ekman (1992). Malgré l'importance de la détection des émotions, la majorité des analyses restent superficielles et insuffisantes pour des tâches nécessitant une compréhension plus approfondie du sens émotionnel en contexte. De telles applications exigent de répondre à des questions clés, notamment l'identification de la cause ayant déclenché l'émotion (Cause), la détermination de la personne qui l'a ressentie (Expérient), et, plus généralement, la prise en compte d'informations structurelles telles que qui a fait quoi (Indice), à qui (Cible), pourquoi (Cause) et comment (Manière).

Cette thèse doctorale vise à proposer des solutions originales et efficaces pour pallier le manque de ressources et de modèles dédiés à l'analyse structurelle des émotions dans les textes arabes. Pour ce faire, nous introduisons une nouvelle approche d'analyse de la structure argumentaire des émotions en arabe, en tirant parti des avancées récentes des modèles fondés sur les Transformers et, en particulier, des capacités des grands modèles de langue (LLMs) pour l'arabe.

Les principales contributions de cette thèse sont multiples. La première contribution consiste en la construction et l'annotation du premier corpus arabe dédié à l'analyse structurée des émotions, nommé "AraERL". La thèse propose également une étude approfondie de l'impact de chaque argument sémantique sur la performance de l'identification des émotions. Elle explore ensuite l'utilisation de ChatGPT pour annoter des textes arabes avec des rôles sémantiques et des émotions à travers une approche de projection interlinguale. Ce travail évalue également la capacité de ChatGPT à projeter avec précision en arabe les annotations sémantiques et émotionnelles issues de l'anglais. Enfin, il offre une comparaison complète des performances des modèles ouverts de grande taille (open-LLMs) pour ces différentes tâches.

Mots-clés: Arabe, TALN, SRL, Projection interlinguale d'annotations, LLMs, ChatGPT, Analyse des émotions, Émotions structurelles

الملخص

في مجال معالجة اللغات الطبيعية، يهدف تحليل الانفعالات إلى ربط المحتوى النصي بمجموعة محددة مسبقًا من الانفعالات البشرية، والتي تشمل عادةً الفرح والغضب والخوف والدهشة والاشمئزاز والحزن. تركز الأبحاث الحديثة بشكل أساسي على تحديد الانفعالات في النصوص باستخدام تصنيفات مستوحاة من النظريات النفسية، مثل الانفعالات الأساسية التي اقترحها إيكمان (1992). ورغم أهمية الكشف عن الانفعالات، فإن معظم التحليلات المتوفرة تبقى سطحية وغير كافية للمهام التي تتطلب فهمًا أعمق للمعنى الانفعالي في السياق. وتتطلّب مثل هذه التطبيقات الإجابة عن أسئلة أساسية، من بينها تحديد السبب الذي أدى إلى حدوث الانفعال (السبب)، وتحديد الشخص الذي اختبره (المُجرِّب)، وبشكل أوسع معالجة المعلومات البنيوية مثل من قام بماذا (الدليل)، ولمن (المستهدف)، ولماذا (السبب)، وكيف (الطريقة).

تهدف هذه الأطروحة إلى اقتراح حلول أصيلة وفعّالة لسدّ النقص في الموارد والنماذج المخصّصة للتحليل البنيوي للانفعالات في النصوص العربية. ولتحقيق ذلك، نقدّم منهجًا جديدًا لتحليل البنية المجاجية للانفعالات في اللغة العربية، بالاعتماد على التقدّم المحرز في النماذج القائمة على Transformers، وخاصة قدرات النماذج اللغوية الضخمة LLMs الخاصة بالعربية.

تتعدّد الإسهامات الرئيسية لهذه الأطروحة. فهي تبدأ بتقديم بناء ووسم AraERL، وهو أول مدونة عربية مخصّصة للتحليل البنيوي للانفعالات. كما تقدّم الأطروحة دراسة معمّقة لتأثير كل حجة دلالية على أداء تحديد الانفعالات. وتستكشف كذلك استخدام ChatGPT في وسم النصوص العربية بالأدوار الدلالية والانفعالات من خلال منهجية الإسقاط بين اللغات. ويقيّم هذا العمل أيضًا قدرة ChatGPT على نقل الوسوم الدلالية والانفعالية من الإنجليزية إلى العربية بدقة. وأخيرًا، تقدّم الأطروحة مقارنة شاملة لأداء النماذج اللغوية المفتوحة الضخمة في هذه المهام.

الكلمات المفتاحية: اللغة العربية، معالجة اللغة الطبيعية، الأدوار الدلالية، الإسقاط التعليقي بين اللغات، النماذج اللغوية الكبيرة، ChatGPT، تحليل المشاعر، المشاعر البنيوية

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

AI Artificial Intelligence.

AraBERT Arabic Bidirectional Encoder Representations from Transformers.

AraERL Arabic Emotion Role Labeling.

BAMA Buckwalter Arabic Morphological Analyzer.

BERT Bidirectional Encoder Representations from Transformers.

BiLSTM Bidirectional Long Short-Term Memory.

BIO Begin-Inside-Outside Tagging Scheme.

BLEU Bilingual Evaluation Understudy.

CBR Case-Based Reasoning.

ChatGPT Chat Generative Pre-trained Transformer.

CL Cross-Lingual.

CLAP Cross-Lingual Annotation Projection.

CNN Convolutional Neural Network.

CoNLL Conference on Natural Language Learning.

CRF Conditional Random Field.

DepecheMood DepecheMood Emotion Lexicon.

DL Deep Learning.

ECE Emotion Cause Extraction.

EMOBANK Emotion-Bank.

ERL Emotion Role Labeling.

FFN Feedforward Neural Network.

GCN Graph Neural Network.

GNE GoodNewsEveryone.

GNN Graph Neural Network.

IEMOCAP Interactive Emotional Dyadic Motion Capture.

IOB Inside–Outside–Beginning.

ISEAR International Survey on Emotion Antecedents and Reactions.

LLM Large Language Models.

LSTM Long Short-Term Memory.

mBART Multilingual Bidirectional and Auto-Regressive Transformers.

mBERT Multilingual BERT.

MELD Multimodal Emotion Lines Dataset.

ML Machine Learning.

MSA Modern Standard Arabic.

NER Named Entity Recognition.

NLP Natural Language Processing.

PATB Penn Arabic Treebank.

POS Part-of-Speech.

PropBank Proposition Bank.

RECOLA Remote Collaborative and Affective Interactions Dataset.

REMAN Relational Emotion Annotation for Fiction.

RNN Recurrent Neural Network.

RoBERTa Robustly Optimized BERT Pretraining Approach.

SRL Semantic Role Labeling.

SRL4E Semantic Role Labeling for Emotion.

SVM Support Vector Machine.

SVO Subject–Verb–Object.

TEC Twitter Emotion Corpus.

VSO Verb–Subject–Object.

XLM-RoBERTa Cross-lingual Language Model - Robustly Optimized BERT Approach.

CHAPTER 1

GENERAL INTRODUCTION

Emotional expressions are fundamental to human experience and social interaction. One of the major challenges in Natural Language Processing (NLP) is understanding how humans express emotions and interpreting the underlying meaning of the linguistic elements involved.

Despite notable advances in recent years, traditional approaches to emotion detection primarily rely on lexical and syntactic features. These methods often overlook deeper semantic and syntactic relations (such as identifying who experiences an emotion and why) which are essential for achieving a complete and accurate understanding of emotional expressions.

Semantic Role Labeling (SRL) addresses this limitation by assigning semantic roles to sentence constituents, providing rich and structured information crucial for advanced NLP tasks such as machine translation, question answering, text summarization, sentiment analysis, and others. The availability of corpora annotated with semantic roles has proven to be a determining factor in improving meaning-based NLP models.

On the other hand, emotion analysis in Natural Language Processing seeks to characterize emotional states, determine their causes and intensities, and examine the contextual factors that influence them. It typically maps textual content to predefined emotion categories such as anger, fear, joy, disgust, sadness, and surprise. This field plays an essential role in enhancing applications related to social media analysis, narrative understanding, news interpretation, and human–computer interaction. However, although existing research has primarily focused on associating textual expressions with psychological models of emotion, less attention has been devoted to identifying the linguistic elements that indicate the experiencer, the stimulus, or modifiers that shape the emotional meaning.

Recently, the integration of SRL into emotion analysis has gained increasing interest, particularly in English [1] and Chinese [2]. However, to the best of our knowledge, no equivalent work has been conducted for the Arabic language. As a low-resource language, Arabic lacks annotated corpora and linguistic tools required for machine learning–based emotion analysis and SRL. Constructing alternative methods to generate annotated data is therefore a crucial step toward developing robust models. To enhance the ability of automatic systems to capture the semantic components of emotions, large and well-annotated datasets are essential. However, building such datasets is challenging due to the complexity of manually identifying emotional structures and semantic roles.

Several challenges arise when integrating SRL into emotion analysis. The first is the absence of corpora annotated with semantic roles and emotion categories for low-resource languages such as Arabic. Existing datasets are usually designed for sentiment analysis and tend to focus on polarity (positive, negative, neutral), without capturing deeper semantic relationships. A second challenge is the limited generalization of existing SRL cross-lingual models, which are primarily trained in English and thus struggle with the Arabic language, due to its unique syntactic and morphological complexity. Moreover, automatic SRL models require large annotated datasets and high computational resources, making their development and application more difficult. Despite these constraints, recent advances in pre-trained Transformer-based models (e.g., mBERT, XLM-RoBERTa) offer promising opportunities for cross-lingual SRL and emotion analysis, especially when combined with fine-tuning and transfer learning.

Motivation and Research Objectives

This thesis aims to enhance Arabic NLP by developing resources enriched with emotion elements and semantic roles, enabling the improvement of computational models and applications for the Arabic language. It also explores the relationship between SRL and emotion analysis, demonstrating how semantic roles support emotion identification and classification. In this context, we built a corpus of Arabic texts manually annotated from various domains—such as news, novels, and social networks, using a semi-automatic platform that combines automatic annotation with human validation to ensure both quality and efficiency.

Inspired by advancements in Large Language Models (LLM), particularly Transformer-based and service-based models such as ChatGPT, this work further investigates how these technologies can strengthen SRL, refine emotion analysis, and support cross-lingual generalization in low-resource settings.

The main objectives of this research are:

- To build a large Arabic dataset of emotional expression, combining semantic role labels with emotion categories, collected a thousands of expressive and meaningful sentences from diverse domains including: tweets, news headlines, religious text, academic books, Arabic novels, and children's book stories.

- To investigate the impact of SRL on improving emotion detection performance in arabic.

- To leverage AI-based chat-bots and LLMs, particularly ChatGPT, to enhance SRL and emotion analysis for the Arabic language.

- To explore cross-lingual transfer learning techniques for SRL-based emotion analysis especially by projecting labels from English to Arabic.

Evaluating open-LLMs (e.g., mBERT, mBART) for automatic SRL and cross-lingual annotation projection.

Proposal and contribution

The main contributions of this thesis are as follows:

1. Construction of the first Arabic dataset annotated with both semantic roles and emotion categories, establishing a foundational resource for structural emotion analysis in Arabic.

2. Exploring the influence of semantic roles (specifically cue, experiencer, target, and cause), on the effectiveness of emotion identification in Arabic language.

3. Leveraging LLMs and AI-based chatbots, especially ChatGPT, for several NLP tasks, including:

    - The assessment of ChatGPT's performance in several tasks, namely, emotion category identification, intensity prediction, and semantic role labeling with cue, experiencer, cause, and target arguments.
    - The investigation of ChatGPT's capability in projecting semantic role labels from English datasets to Arabic.
    - The construction of two benchmark Arabic data sets: one manually translated and annotated, the other generated using ChatGPT.
    - The evaluatinon of open-LLMs (e.g., mBERT, mBART) for automatic SRL and Cross-Lingual Annotation Projection.

Thesis structure

This thesis is divided into two main parts. The first part covers background concepts and literature review, while the second presents the contributions. The outline of this thesis is as follows:

Part I: Background and Literature Review

- **Chapter 2**: Overview of Arabic language processing, including linguistic characteristics, challenges, techniques, and available NLP resources.

- **Chapter 3**: Definition, taxonomy, datasets, annotation methods, and learning techniques for Semantic Role Labeling.

- **Chapter 4**: introduces key definitions of emotion and emotional expression, outlines major psychological models of emotion. The chapter also reviews existing emotion analysis datasets.

- **Chapter 5**: Explores the principles underlying Emotion Role Labeling, including semantic frames, core emotion roles, annotation examples, and existing ERL datasets.

- **Chapter 6**: A selected overview of relevant work in the research areas under study.

Part II: Contributions

- **Chapter 7**: Dataset construction and annotation of Arabic texts with semantic roles, emotion categories, and emotional structures.

- **Chapter 8**: Demonstrate the affect of SRL on emotion detection using transformer based models.

- **Chapter 9** : Present the third contribution using deverse LLMs to improve SRL.

The Thesis is closed by a general conclusion that summarizes our overall preception of the aquired knowledge throughout the doctorate cycle, as well as, it highlights the perspectives, and future works.

Publications

[1] Senator, F., Lakhfif, A., Zenbout, I., Boutouta, H., and Mediani, C. (2025). Leveraging ChatGPT for Enhancing Arabic NLP: Application for Semantic Role Labeling and Cross-lingual Annotation Projection. IEEE Access.

[2] Senator, F., Boutouta, H., Lakhfif, A. and Mediani, C., 2023, March. Semantic Role Labeling of Arabic Emotional Text in Tweets. In 2023 International Conference on Advances in Electronics, Control and Communication Systems (ICAECCS) (pp. 1-6). IEEE.

[3] Senator, F., Boutouta, H., Lakhfif, A., & Mediani, C. (2025). The impact of semantic roles in emotion detection using the AraBERT model. In Proceedings of the Mediterranean Conference on Computer Science and Artificial Intelligence (MCCSAI'2025) (accepted, to appear). Springer.

Part I

Background and Literature Review

CHAPTER 2
ARABIC LANGUAGE PROCESSING

Arabic has become one of the most widely used languages in the world, both in writing and speaking. It is a Semitic language, closely related to Hebrew, Aramaic, and the Semitic languages of Ethiopia. Arabic has been spoken for over fifteen centuries and serves as the official or national language in twenty-two (22) countries, collectively known as the Arab League. It is considered one of the most important languages globally, especially among the fastest-growing languages on the internet. Additionally, Arabic is used by many Muslims around the world and is spoken by approximately 400 million people today [3].

With the exception of some early poetry, the Holy Quran—the sacred book of Muslims worldwide—is the oldest known written work in Arabic.

In this chapter, we explore the unique characteristics of the Arabic language and their implications for Natural Language Processing (NLP). We begin by discussing the fundamental linguistic features of Arabic, including its script and alphabet, rich and complex morphology, flexible syntax, and the significant variation introduced by regional dialects. These characteristics pose distinct challenges for NLP tasks. Furthermore, the chapter highlights key NLP resources that have been developed to support Arabic computational linguistics, including the Penn Arabic Treebank (PATB), the Quranic Arabic Corpus, theBuckwalter Arabic Morphological Analyzer (BAMA), and Arabic WordNet. These resources play a crucial role in enabling more accurate and effective analysis of Arabic text.

2.1 Major Challenges in Arabic Language Processing

Arabic presents several challenges for NLP due to its complex linguistic structure and variability. The language is morphologically rich and highly inflectional, where words are derived from roots through different patterns; for instance, from the root [كتب] (k-t-b, "to write") we obtain [كتب] ("he wrote"), [كاتب] ("writer"), and [مكتوب] ("written"). Moreover, Arabic is typically written without short vowels (diacritics), leading to ambiguity in interpretation—e.g., [علم] can mean ᶜalam ("flag"), ᶜilm ("knowledge"), or ᶜalima ("he knew"). Another major issue is diglossia, as Modern Standard Arabic (MSA) coexists with numerous regional dialects that differ lexically, morphologically, and syntactically, making cross-dialectal NLP difficult. Orthographic variation (such as inconsistencies in the use of alif— [أ / ا] —or ya— [ى / اي]) further complicates tokenization and normalization. The limited availability of large, annotated Arabic corpora compared to English adds to the difficulty, as do ambiguous words like [فرح] which may mean "joy," "he rejoiced," or be a personal name. Together, these features make Arabic a linguistically rich but computationally challenging language to process effectively [4–6]

2.2 Characteristics of the Arabic Language

Arabic has some specific characteristics, such as:

- **Script and Alphabet:** unlike English and many other languages that are written from left to right, Arabic is written and read from right to left. In terms of Alphabet, Arabic uses a special script of 28 letters. The script is cursive; that is most letters are connected within words. Also, Arabic does not have capital and lowercase letters; all letters are in a similar form regardless of their position in the sentence.

- **Morphology:** Arabic words are typically derived from (03) triliteral (three-consonant) roots, using patterns of vowels and affixes (prefixes, suffixes, infixes) added to derive meanings. Such a rich morphology, let Arabic the power of generating numerous word forms from a single root. Example: The root "k-t-b" can produce words like "kitāb" (book), "maktab" (office), "kātib" (writer). Also, Inflectional and derivational Morphology provide Arabic with many ways to form new words or change word meanings (e.g., verbs are conjugated for person, gender, and tense). In Arabic, nouns, adjectives, and verbs are defined as either masculine or feminine. Depending on the gender, the word's form frequently changes. Nonetheless, the conjugation of Arabic verbs is based on the person, gender, number, tense, and mood. This impacts pronouns, adjectives, and verbs alike.

- **Complex Verb System:** In Arabic, verbs can convey a variety of moods and tenses, including indicative, subjunctive, jussive, and future. In comparison to many Indo-European languages, the verb system is more intricate. The structure of verbs is based on a root-pattern system, in which the root serves as the fundamental unit of meaning and different patterns—often involving vowel changes—are used to convey grammatical relationships.

- **Syntax:** Arabic language typically follows a Verb–Subject–Object (VSO) word order, However in modern contexts, particularly in everyday conversation, () order is also frequently used. In addition, Arabic features complex syntactic structures, utilizing case markers and prepositions to show the grammatical relationships between words.

- **Dialects and Variants:** Arabic can be divided on Modern Standard Arabic (MSA), colloquial dialects and diglossia. MSA is a formal version employed in writing and in formal speech and discussion, etc. MSA is not commonly spoken in daily conversation. While Arabic has numerous spoken dialects (e.g., Egyptian, Levantine, Gulf, Maghrebi) that vary significantly in vocabulary, pronunciation, and grammar from MSA. There is a significant difference between written Arabic

(MSA) and spoken Arabic, creating a situation known as ”diglossia.”

2.3 NLP resources for Arabic

Arabic language has attracted a lot of researches, last decades. Various resources for NLP for Arabic were published to tackle issues raised within Arabic NLP applications. Developing effective NLP systems for Arabic involves using NLP computational resources for Arabic. There are various annotated corpora for several tasks such as Part-of-Speech (POS) tagging, syntactic parsing, NER, etc.

- **Penn Arabic Treebank (PATB):** is one of the most well-known resources for Arabic syntactic annotation, by providing a treebank for Arabic, with annotations for syntactic structure, POS tagging, and more. PATB is used in syntactic parsing and POS tagging [7].
- **The Quranic Arabic Corpus:** An annotated corpus of the Quran with morphological, syntactic, and grammatical information. Quranic Arabic Corpus is useful for linguistic analysis of Quranic Arabic [8].
- **Buckwalter Arabic Morphological Analyzer (BAMA):** a rule-based computational tool developed by Tim Buckwalter to analyze Arabic words and provide their possible morphological analyses. [9]
- **Arabic WordNet:** a lexical database for MSA, modeled after the original Princeton English WordNet. It organizes Arabic words into sets of synonyms (called synsets) and encodes various semantic relationships among them. [10]

2.4 Conclusion

In summary, this chapter underscores the significance of the Arabic language, both in its linguistic richness and its growing importance in the digital world. By examining its linguistic features and the challenges they present to NLP, we gain a deeper understanding of why specialized tools and resources are essential for effective Arabic language processing. The reviewed resources not only reflect the progress made in Arabic NLP but also lay the foundation for future advancements aimed at addressing the complexity and diversity of Arabic in various computational applications.

CHAPTER 3

SEMANTIC ROLE LABELING

Natural Language Processing (NLP) enable computers to understand, interpret, generate, and manipulate human language. NLP encompasses a wide range of tasks, including but not limited to: Part-of-Speech (POS) Tagging [11], Named Entity Recognition (NER) [12], Sentiment and Emotion Analysis [13], Machine Translation, Text Summarization, Question Answering [14], Semantic Role Labeling (SRL) [15], Language Modeling and Text Generation.

In this chapter, we provide a comprehensive overview of SRL. We begin by defining SRL and highlighting its importance in understanding the meaning of text. The chapter then reviews prominent SRL datasets available for both English and Arabic. We discuss the various annotation methodologies used to construct these resources. Furthermore, we explore different approaches to SRL, and we also address different learning strategies employed in SRL tasks. Finally, we outline the standard evaluation metrics used to assess SRL systems.

3.1 Overview and Definition

Semantic Role Labeling is a fundamental task in NLP that aims to identify the semantic relationships between sentence constituents. Deeper language comprehension requires understanding these relationships by identifying the roles of each word or sentence. SRL is useful in numerous downstream NLP applications, including information extraction [16], automatic document categorization [17], question answering [18, 19], and Twitter information extraction [20].

Semantic roles represent one of the oldest and most fundamental concepts in NLP. Their origins can be traced back to the Paninian Karaka theory [21], which dates as far back as 1966 and remains a cornerstone in the study of predicate-argument structures.

Jurafsky and Gildea [15], in one of the seminal works on SRL, define it as: *"The task of identifying each argument of a predicate with a semantic role, such as Agent, Patient, Instrument, etc., to represent the relationship between the predicate and its arguments."*

Similarly, the Berkeley FrameNet Project [22] defined semantic roles based on semantic frames as follows: *"The act of identifying and assigning labels to words or phrases in a sentence, according to their semantic functions within a specific frame. These frames depict various categories of events, activities, or circumstances, and the corresponding roles (frame elements) explain the participants and characteristics engaged in these frames."*

SRL is an NLP task that aims at identifying the predicate-argument structure of a sentence by assigning roles to each input sentence component according to their semantic relationships with a predicate. These roles define who did what to whom, when, where, why, and how, providing a structured representation of sentence meaning. SRL typically relies on linguistic frameworks such as Proposition Bank (PropBank) and FrameNet.

Diverse terms are used to denote the annotation of semantic roles we cite: Semantic Role Labeling, Frame Semantic Parsing, Thematic Role Labeling, PropBank Role Labeling, and Argument Structure Parsing.

Figure 3.1: Semantic role representation

The aim of SRL is to identify and classify the arguments of each target verb into semantic roles. For example, for the sentence "Last year, the luxury car manufacturer sold 2000 cars in U.K." SRL yields the following outputs: [*Time* Last year] [*Agent* the luxury car manufacturer] [*Verb* sold] [*Theme* 2000 cars] [*Location* in U.K.]. Here, Agent represents the seller, Theme represents the thing sold, Time represents the timing of the action, Location is where the action happens and Verb represents the action.

3.1.1 Labeling steps

This task is usually accomplished in four steps:

- **Predicate identification:** Determine which words in the sentence are predicates, usually verbs.
- **Arguments identification:** Determine the phrases or words in the sentence that function as arguments to the predicate.
- **Roles assignment:** Assign semantic roles based on PropBank, FrameNet, or Thematic Roles
- **Verification:** Expert review and validate labels.

3.2 Semantic role labeling projects and resources

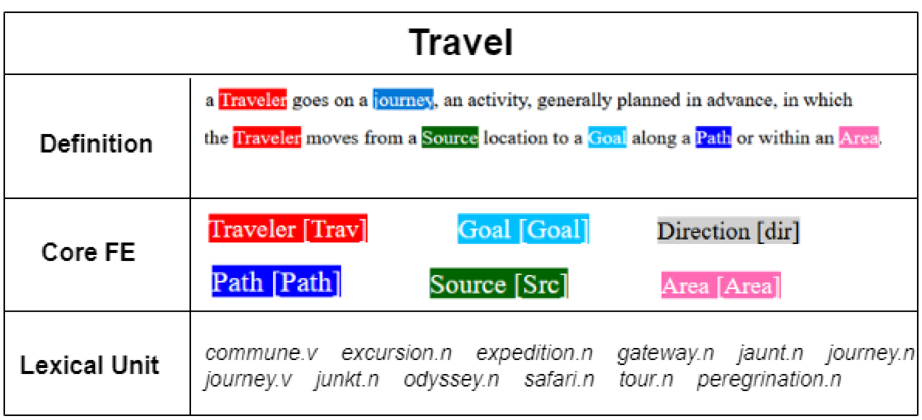
High-quality datasets that offer annotated instances of predicate-argument structures are crucial for the development and assessment of SRL systems. Several publicly accessible SRL datasets are frequently used as training data for SRL tasks, including

PropBank, FrameNet, and CoNLL Shared Task datasets, and have significantly contributed to the advancement of research in this domain. These resources empower researchers to construct, train, and evaluate SRL models across various languages and domains.

3.2.1 English Resources

3.2.1.1 The Berkeley FrameNet project

An English lexical database, created by Baker et al. [22], based on the theory of frame semantics, which associates word meanings with conceptual structures known as frames. The project explains how lexical units evoke specific semantic frames, where lexical units are words or phrases associated with these frames. FrameNet offers a lexicon that includes more than 13,000 word senses, accompanied by more than 200,000 manually annotated sentences, along with more than 1,200 semantic frames, which constitute an interesting training dataset for the identification of semantic roles [23].
The concept was initially developed for English and has then affected in other languages like Spanish [24], German, Russian [25], and Japanese [26], improving cross-linguistic semantic analysis. Commonly applied in several NLP tasks such as SRL, information extraction, machine translation, and sentiment analysis.



Travel	
Definition	a Traveler goes on a Journey, an activity, generally planned in advance, in which the Traveler moves from a Source location to a Goal along a Path or within an Area.
Core FE	Traveler [Trav] Goal [Goal] Direction [dir] Path [Path] Source [Src] Area [Area]
Lexical Unit	commune.v excursion.n expedition.n gateway.n jaunt.n journey.n journey.v junkt.n odyssey.n safari.n tour.n peregrination.n

Figure 3.2: Example of frame from Framenet: the Travel frame and its elements

The Table 3.2 describes the Travel frame, which represents scenarios where an entity moves from one location to another. It includes key frame elements (roles) that define different aspects of the travel process.

- **Traveler:** the person or entity making the journey.
- **Source:** the starting point of the travel.
- **Direction:** the endpoint or target location of the journey.
- **Path:** the route taken during the travel.
- **Area:** the area in which the traveling takes place.

3.2.1.2 Propbank

Palmer et al. [27] created the linguistic resource PropBank, which helps to make the relationships between verbs and their arguments by giving semantic role labels to the verbs' arguments. Each verb in PropBank is annotated with roles like Arg0 (Agent), Arg1 (Patient), and other roles like Arg2, Arg3, etc. Usually, each number represents a distinct semantic role. For instance, the agent is often denoted by Arg0 and the patient by Arg1 [28]. PropBank was first created for English, but has since been expanded to Chinese [29], Arabic [30], Turkish [31], and Finnish [32]. Its versatility enables it to increase performance in tasks like SRL, information extraction, and machine translation, as well as to improve language comprehension in specialized disciplines.

Figure 3.3: Sentence example (annotated in PropBank format)

3.2.1.3 CoNLL Shared Task

Conference on Natural Language Learning (CoNLL), a premier event within the Association for Computational Linguistics (ACL). The challenge offers annotated datasets for training and testing, and each year it focuses on a different topic of NLP. Initially conducted in English in 2005, SRL tasks were then expanded to six languages: Spanish, Catalan, Chinese, Czech, German, and Japanese [33]. These shared tasks have significantly contributed to advancements in semantic understanding, syntactic analysis, and cross-lingual processing. The CoNLL Shared Tasks on SRL were pivotal in advancing research in the field and took place in the following years: CoNLL -2004, CoNLL -2005 [33], CoNLL -2008 [34], CoNLL -2009 [35].

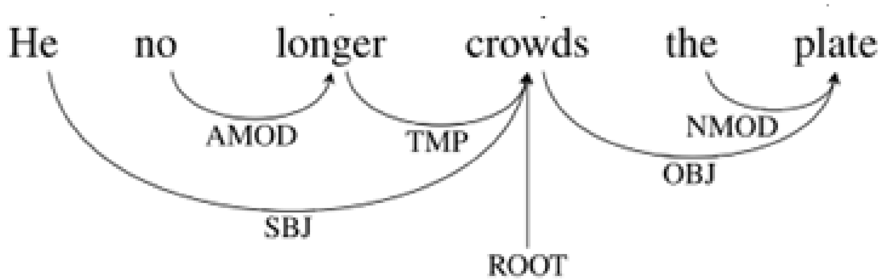


Figure 3.4: A sentence from the CoNLL-2009 English dataset annotated with semantic roles. [36]

3.2.1.4 VerbNet

Hierarchical lexical resource for English, which organizes verbs into several classes based on Levin's verb classes, taking into account their syntactic and semantic characteristics. Each verb class in VerbNet offers information about syntactic structures and their associated semantic roles.

VerbNet is especially useful for SRL and invaluable for applications like machine translation, data extraction, and event detection. There is a mapping from VerbNet to other semantic resources like FrameNet and PropBank. [37]

3.2.2 Arabic resources

3.2.2.1 The Arabic PropBank

A linguistic resource, was developed by the University of Colorado, to provide semantic role annotations for Arabic texts, and both Chinese and English PropBanks are available. However, the process of developing an Arabic PropBank is considerably different due to the morphological and syntactic representation of Arabic [30]. The example below is taken from [30]

- Sentence:

[وكانت مجموعة اسلامية مسلحة اقامت حاجزا وهميا على الطريق الدولي]

- Annotation:

[وكانت مجموعة اسلامية مسلحة] Arg0 [أقامت] Predicate [حاجزا وهميا] Arg1 [على الطريق الدولي] ArgM-LOC

3.2.2.2 Arabic VerbNet

An extensive verb vocabulary that categorizes Arabic verbs into hierarchical classes. It aims to illustrate the syntactic and semantic relationships of each verb in the lexicon, which contains 334 classes, 7,672 verbs, and 1,393 frames.[38]

3.2.2.3 The Arabic FrameNet

Inspired from original FrameNet project for English, Lakhfif & Laskri [39] proposed a resource for Arabic following the principal of frame semantics tenet. Their work involved building a semantic parser capable of extracting deep semantic structures at lexical, morphological, syntactic, and semantic levels to translate Arabic into Sign Language. To support this, they constructed an Arabic FrameNet by developing a rule-based method that leverages mappings between Arabic WordNet and FrameNet. As a result, they automatically generated over 8,000 Arabic lexical units and more than 600 semantic frames.

3.3 Taxonomy of SRL task

Figure 3.5 represents the taxonomy of SRL tasks and serves as an overview of the chapter. It encompasses: (1) publicly available datasets and labeling strategies aimed at expanding training data, particularly for low-resource scenarios; (2) methods and learning approaches designed to enhance the performance of SRL models; and (3) evaluation metrics used to assess SRL models.

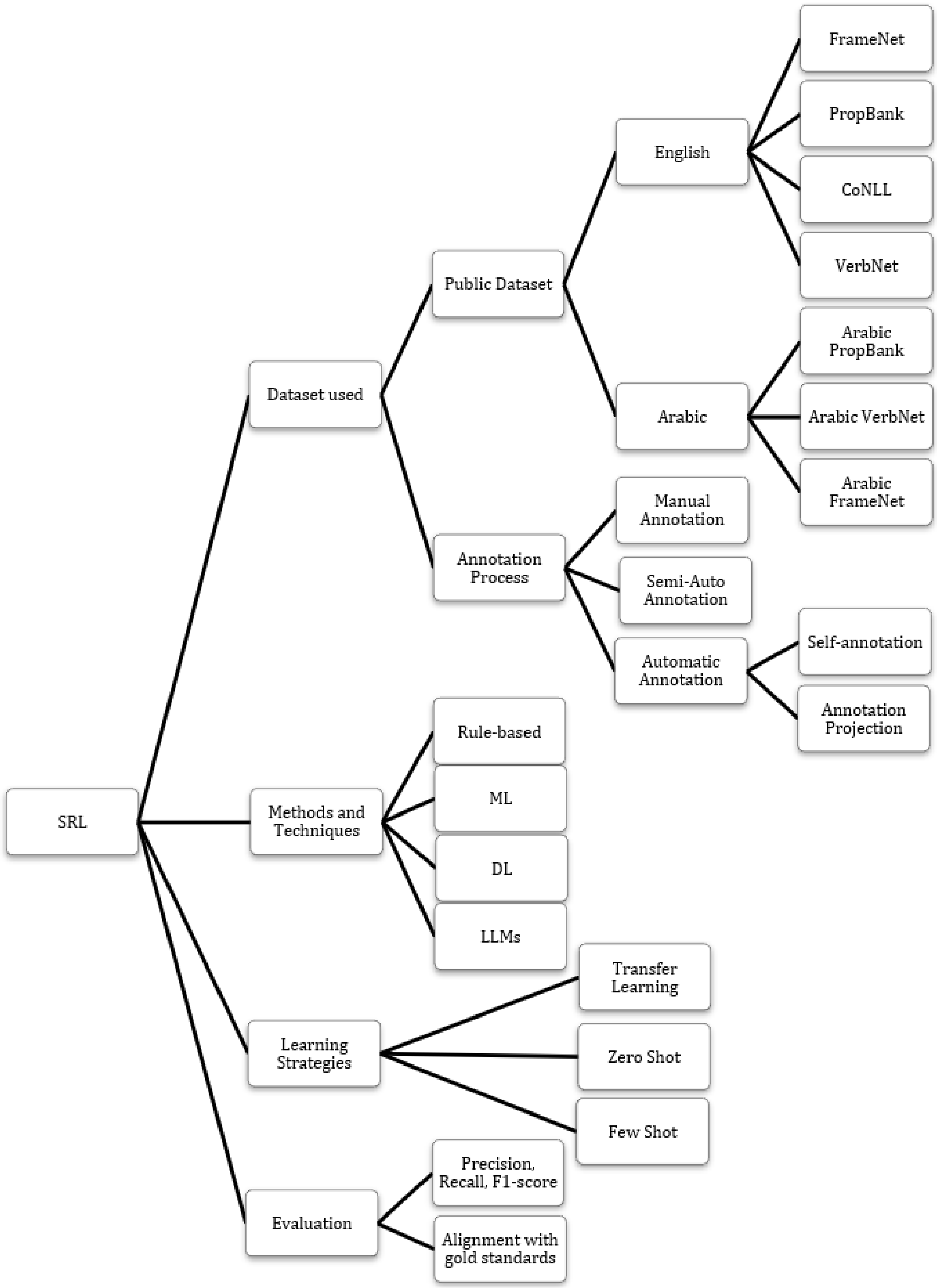


Figure 3.5: Taxonomy of SRL

3.3.1 SRL Annotation process

Various methods have been developed over time to accomplish labeling and assigning semantic roles, reflecting developments in computational linguistics, language processing and the increasing availability of annotated corpora. According to the needs of the task and the resources available, SRL annotation process can be divided into three categories: manual, automatic, and semi-automatic.

3.3.1.1 Manual SRL Annotation

Refers to the process of assigning and annotating text by a human (in general, a human is often a domain expert) to identify roles and their relationships with predicates in a sentence. In this method, the annotator reads the text attentively and understands the meaning based on the intelligence and linguistic expertise to accurately assign relevant labels [40].

Manual annotation was used by Zhang et al. [41]; they extract narratives from English news texts. Two annotators; an expert in the dynamics of regeneration and a doctoral student, independently annotated news texts, and inter-annotator agreement was measured. Campagnano et al. [42] aligned emotion categories and semantic roles manually to fit the unified Semantic Role Labeling for Emotion (SRL4E) framework, this made existing datasets more consistent and better. In [43], the annotation was conducted using Amazon's Mechanical Turk service to crowdsource the annotation of electoral tweets, in order to encompass the experiencer, stimulus, and the cue. Manual annotation was crucial for understanding the emotional content of tweets due to the informal context of social media. Bostan et al. [1] build the GoodNewsEveryone (GNE) corpus, which consists of news headlines annotated with emotions, semantic roles, and reader perception. 310 Annotators follow precise guidelines and are trained to ensure consistency. In the final phase, inter-annotator agreement is measured to validate the quality of the annotations.

3.3.1.2 Semi-automatic SRL Annotation

A hybrid methodology that annotates text with semantic roles by combining automated techniques, including rule-based systems or pre-trained models or an annotation tool, with human participation. The initial tagging is done by automated techniques, and the output is then refined by human reviewers to ensure accuracy and consistency [44]. Cai et al. [45] use an automatic SRL tool to identify arguments, followed by human intervention to correct errors in trigger identification and argument labeling. Senator et al. [46] create an emotional Arabic corpus annotated with semantic roles using a semi-automatic tool and a team of annotators.

3.3.1.3 Automatic SRL Annotation

The process of annotating data without the need for human interaction, using techniques such as rule-based systems, algorithms, or Machine Learning models. To improve training data, SRL usually requires the use of pre-trained models or syntactic parsers to identify the roles of sentence components [47]. Can be achieved through strategies like self-augmentation and annotation projection.

3.3.1.3.1 Self-augmentation Is a data enhancement technique where a model or dataset is used to generate additional training examples, leveraging its own outputs or variations of its data. Data augmentation was developed to address the problem of limited data and to increase data diversity, especially in Deep Learning-based tasks like SRL, which require accurate labeled data [48]. Liu et al. [49] introduce a self-augmentation approach to enhance event argument extraction and to enrich the training set.

3.3.1.3.2 Annotation projection Leverages the existing well-annotated resource in the source language to overcome the rarity of resources in other languages based on the translational and structural equivalences present in the aligned data [50]. This technique can help reduce the effort in terms of time and both material resources and cost required to generate annotations or models for a new target language [45].

A tool for SRL tasks in Sinhala was proposed by Gunasekara et al. [51] utilizing an annotation projection approach, initially, AllenNLP tools are used to annotate the source sentence (in English) with semantic roles. The target sentence in the Sinhala language will then be projected with the semantic role labels.

3.3.2 SRL Methods and techniques

The semantic role classification and labeling process has seen significant progress in its methodologies over time, passing through distinct stages of development. In this work, we divide SRL methods into three main categories: Traditional, Machine Learning, and Deep Learning. At the begining, SRL systems depended on syntactic dependency structures and conventional rule-based methodologies. With the emergence of Machine Learning, SRL systems shifted to utilizing manually labeled corpora and feature-based statistical models, achieving improved accuracy and broader applicability. But in the last ten years, DL introducing powerful methods such as transformer-based models like Bidirectional Encoder Representations from Transformers (BERT) and Robustly Optimized BERT Pretraining Approach (RoBERTa), Graph Neural Network (GCN), and Recurrent Neural Network (RNN).

3.3.2.1 Rule based

Rule-based methods are a traditional approach that aims to solve computational linguistics problems using manually created rules. These methods have been extensively used for various NLP applications, SRL, NER, and syntactic parsing. When applied to both structured and unstructured text in various domains, such as formal language in news articles or non-formal text from social media, rule-based algorithms are particularly effective. Rule-based methods are used in SRL because it can produce precise SRL results, especially in domains or datasets with limited training data. [52]

3.3.2.2 Machine Learning

A branch of AI aims to enable computers to learn from data and enhance their performance on tasks without explicit programming.

Machine Learning involves algorithms that analyze and process data, identify patterns, and make predictions. There are three main types in ML: When the algorithm learns from labeled data, where each input has its corresponding output, it is referred to as supervised learning. In contrast, unsupervised learning refers to an algorithm that explores data without labeled outputs. Reinforcement learning means that an agent learns to make decisions by interacting with an environment. [53]

Machine Learning approaches have considerably enhanced the evolution of SRL systems through the automation of this process. In Machine Learning-based SRL, models are trained and evaluated using statistical techniques. Statistical methods are used to find patterns and correlations in the manually annotated datasets. [54]

3.3.2.3 Deep Learning

Deep Learning is a subset of ML that utilizes artificial neural networks with multiple layers to model and learn complex patterns in large amounts of data. It is particularly effective for tasks such as NLP, image recognition, and speech processing. DL models, such as Convolutional Neural Network (CNN) and transformers, rely on large datasets and high computational power to achieve state-of-the-art performance in various domains. [55]

Deep Learning has significantly improved SRL by replacing traditional feature engineering with automated feature extraction using neural networks.

Fine-tuning DL models on SRL datasets, such as CoNLL -2005, CoNLL -2012, or Arabic corpora, has achieved state-of-the-art results [56–58].

3.3.2.4 Large Language Models

Large Language Models (LLM) are advanced Artificial Intelligence (AI) models trained on massive amounts of text data based on deep neural network architectures, most notably the Transformer. Transformer architectures have transformed language models by introducing attention mechanisms, greater parallelization, and an improved ability to capture long-range dependencies. These advancements have made it possible to create more powerful and flexible models that have achieved state-of-the-art performance on a wide range of NLP tasks, from translation to text generation. The pre-training and fine-tuning approach, popularized by Transformers, has now become a standard in the field. [59]

- **Open-LLM:** Open Large Language Models (LLM) are publicly available models that allow access to their architecture, weights, and training data (partially or fully). These models can be fine-tuned, modified, or deployed by researchers and developers for various applications. Examples include BERT, LLaMA, Falcon, BLOOM, and Mistral, which enable greater transparency and customizability. [60]

- **Close-LLM:** Close LLM are proprietary models that restrict access to their architecture, weights, or training data. Users typically interact with these models via APIs. Examples include GPT-4 (OpenAI), Claude (Anthropic), and Gemini (Google DeepMind). These models limit transparency and customization.[60]

3.3.3 Learning Strategies for SRL

The main challenge in using Machine Learning and DL approaches is to collecting enough of training data (expensive to collect and label, or inaccessible) for each category. This limitation reduce the model's ability to represent some classes. Addressing this difficulty has motivated the exploration of novel learning strategies adapted to situations with limited labeled data. We classify learning scenarios for SRL tasks into three principal paradigms: transfer learning, zero-shot learning, and few-shot learning.

3.3.3.1 Transfer Learning / Cross-lingual Learning

Transfer learning entails utilizing knowledge from a source domain, language, or tasks which has abundant training data and transfering it to a target domain, language or tasks. It generally depends on pre-trained models, cross-lingual embeddings, or fine-tuning methodologies.[61]. Subburathinam et al.[62] breaks down the challenge of identifying the deep semantic structure of entities and their interactions, as well as events, from text. The authors propose a novel cross-lingual structure transfer learning approach that combines the distributed characteristics, for instance contextualized embeddings, with the language-independent symbolic features such as the dependency paths and the Part-of-Speech (POS) tags. The results obtained while applying the suggested approach on Arabic, Chinese and English demonstrate that the suggested approach performs competitively.

3.3.3.2 Zero-Shot Learning

Zero-shot learning refers to the ability of a model to identify and label semantic roles in sentences from previously unseen languages (without being explicitly trained on annotated examples for those specific cases). Zero-shot models rely on knowledge transfer from pre-trained language models or high-resource datasets, enabling them to generalize to new tasks with no labeled examples. [63]

In practice, zero-shot SRL is often achieved using multilingual pre-trained models like Multilingual BERT (mBERT) or Cross-lingual Language Model - Robustly Optimized BERT Approach (XLM-RoBERTa), which learn deep syntactic and semantic representations across multiple languages. These models can perform SRL in low-resource languages (like dialectal Arabic) by leveraging the training knowledge obtained from high-resource languages (like English). [64]

Bombieri et al. [65] created a novel SRL-based approach for extracting procedural knowledge from surgical texts, evaluated using multiple language models under zero-shot settings.

3.3.3.3 Few-Shot Learning

In SRL, few-shot learning refers to the model's ability to learn to identify semantic roles from a very limited number of labeled examples. few-shot SRL adapts quickly to new tasks through limited fine-tuning or prompt-based learning. Advanced methodologies such as meta-learning enhance models' ability to acquire knowledge with less data. [66]

3.3.4 Evaluation Metrics

Evaluation metrics are essential tools for measuring the performance of Machine Learning models in NLP tasks such as SRL and Sentiment Analysis. The most widely used metrics include Precision, Recall, and F1-score, along with Alignment with the Gold Standard.

3.3.4.1 Precision, Recall, and F1-score

The concepts of precision, recall, and F1 score are used to assess how well SRL tasks are performed. These terms include: (1) true positive, which refers to data that the system correctly labels; (2) false positive, which refers to data that the system incorrectly labels; and (3) false negative, which denotes instances that the system missed or failed to detect. The weighted mean of accuracy and recall, or F1, is useful in cases when the distribution of classes is unbalanced. The precision, recall, and F1 score computations are found in Equations below.

$$\textbf{Precision} = \frac{Truepositives}{Truepositives + Falsepositives}$$

$$\textbf{Recall} = \frac{Truepositives}{Truepositives + Falsenegatives}$$

$$\textbf{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

3.3.4.2 Alignement with Gold Standard

Alignment with the gold standard is the degree to which the model's output closely resembles the manually annotated reference data (the gold standard). It ensures the correctness and consistency of the predicted labels compared to expert human annotations. A high level of alignment means that the model accurately simulates human judgment. Research by Jindal et.al [67] introduces PriMeSRL, a stricter evaluation metric that provides a more accurate assessment of SRL model performance by aligning outputs with gold standard annotations.

3.4 Conclusion

This chapter provides a comprehensive overview of the key principles pertaining to the fields of SRL presenting its foundational concepts, linguistic relevance, and practical applications within NLP. It provides an overview of key resources in both English and Arabic, discusses annotation methodologies, and reviews a range of computational approaches from traditional models to cutting-edge large language . By outlining current learning paradigms and evaluation practices, this chapter offers the necessary background of SRL which is central to the focus of this thesis.

CHAPTER 4 EMOTION ANALYSIS

In this chapter, we explore the foundational aspects of emotion analysis within Natural Language Processing (NLP). We begin by defining emotion analysis and examining its significance in understanding human affective states in text. Then we discuss explicit emotion expressions, including a discussion of emotion categories based on established psychological theories. Furthermore, we review several representative emotion-annotated corpora.

4.1 introduction

Traditional sentiment analysis can determine the polarity of a statement, however, it lacks the granularity needed to understand what specific emotion is being expressed.

Our actions and words invariably mirror our feelings, if not always explicitly [68, 69]. To understand the underlying behavior of humans, we need to analyze these emotions through some emotional data, sometimes referred to as affect data. This data may include text, audio, facial expressions, etc. The use of emotional data for the analysis of emotions is an interdisciplinary domain known as Affective Computing.

Despite the progress, challenges remain, especially in dealing with low-resource languages, ambiguous expressions, and the subjective nature of emotions. Therefore, emotion analysis continues to be an active area of research, aiming to bridge the gap between human emotional expression and machine understanding.

4.2 Definition

According to the Oxford Advanced Learner's Dictionary, the noun emotion is defined as: *"A strong feeling such as love, fear, or anger; the part of a person's character that consists of feelings."* [70]. While, the Cambridge Dictionary defines emotion as *"a strong feeling such as love or anger, or strong feelings in general"* .[71]. The American Psychological Association defines emotion as *"a complex pattern of changes, including physiological arousal, feelings, cognitive processes, and behavioral reactions, made in response to a situation perceived to be personally significant"*[72].

The complexity and the subjectivity of emotions make them difficult to interpret and analyze, especially in large-scale or automated settings.

Emotion analysis is an effective study of human emotions that tries to identify the proper emotion from context, and analyze it according to some predefined emotion class models, it aims to bridge the gap between human emotional expression and machine understanding by enabling computational systems to detect, classify, and interpret emotional cues in text [73]. Emotion analysis typically based on psychological theories of Ekman (1992)[74], Plutchik [75] (2001), Russell (1980) [76] or Scherer (2005) [77]. The task has evolved from being only a research issue to sweeping a number of applications, such as social media mining, intelligent agents, clinical diagnosis of mental diseases, and dialog systems (chatbots, teaching systems) [78].

4.3 Emotions Expression

Explicit expressions directly state the emotion using clear emotion-bearing words. These are typically adjectives, verbs, or nouns that clearly denote a specific emotional state [79]. For example:

- "I am angry."
- "She felt joy after receiving the news."

- ”His sadness was overwhelming.”

In such cases, the emotion is linguistically marked and easily identifiable because the language includes direct references to emotions such as *angry, joy, or sadness.* These are often easier to annotate and detect using lexicon-based or supervised approaches.
From a psychological standpoint, Ekman’s model [74] of six basic emotions (anger, disgust, fear, joy, sadness, and surprise) and Plutchik’s wheel of emotions (see Figure 4.1) [75], are the most widely used.

Since the 2010s, automatic emotion detection in text has gained significant interest, particularly with the rise of social media [80]. Emotion recognition systems are applied in diverse domains, such as public opinion monitoring [81], consumer insights [82], financial forecasting [83], and election prediction [84]. Furthermore, they play a key role in developing empathetic virtual assistants and chatbots [85, 86].

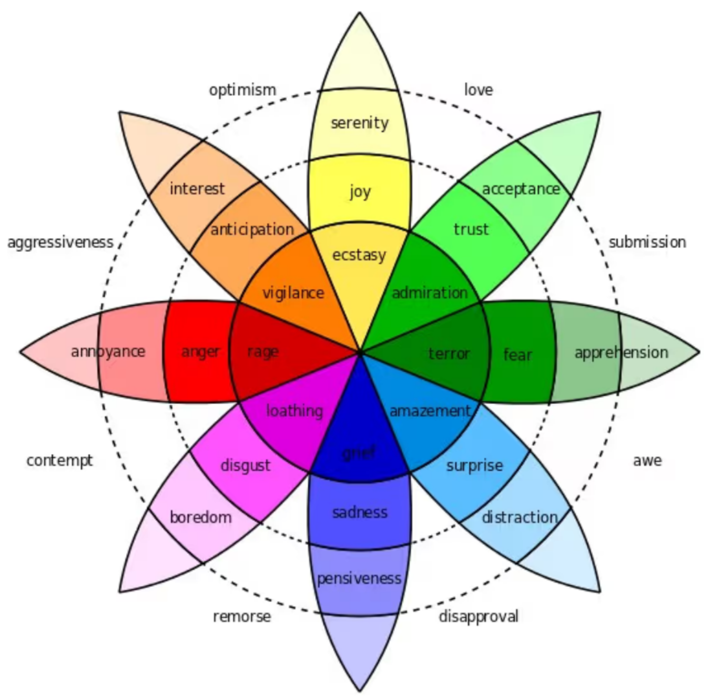


Figure 4.1: Plutchik’s Wheel of Emotions [87]

The definitions of each emotions category provided in the tables below 4.1, and 4.2 are adapted from established frameworks in affective computing. In particular, they are drawn from Robert Plutchik’s theory of emotions [75], Paul Ekman’s basic emotions model [74] and the APA Dictionary of Psychology and standard glossaries in emotion research [88, 89]. These definitions have been rephrased and condensed to suit the context of the thesis.

Table 4.1: Basic emotions and their definitions

Emotion	**Definition**
Anger	A strong feeling of displeasure, hostility, or antagonism toward someone or something perceived as harmful or unjust.
Anticipation	A state of looking forward to a future event, often with excitement, hope, or anxiety.
Disgust	A feeling of revulsion or disapproval aroused by something unpleasant or offensive.
Fear	An emotional response to a perceived threat or danger.
Joy	A feeling of great pleasure, happiness, or delight.
Sadness	A state of emotional pain associated with loss, disappointment, or sorrow.
Surprise	A sudden feeling of astonishment or amazement caused by something unexpected.
Trust	A sense of security or confidence in someone or something.

Table 4.2: Derived emotions based on Plutchik's model

Derived Emotion	**Dyad Components**	**Definition**
Optimism	Anticipation + Joy	Hopefulness and confidence about the future or successful outcome.
Love	Joy + Trust	Deep affection, attachment, and care for someone or something.
Submission	Trust + Fear	Yielding to authority or another person due to respect or fear.
Awe	Fear + Surprise	A feeling of reverential respect mixed with fear or wonder.
Disappointment	Surprise + Sadness	Sadness caused by unmet expectations or outcomes.
Remorse	Sadness + Disgust	Deep regret or guilt for a wrong committed.
Contempt	Disgust + Anger	A mix of anger and disgust directed at someone regarded as beneath consideration.
Aggressiveness	Anger + Anticipation	A readiness or tendency to attack or confront.

Table 4.3: Comprehensive list of emotion classification corpora: languages, modalities, emotion types, and dataset resources.

Ref	Corpus	Language(s)	Modality	Emotion Types	Size
[90]	GoEmotions	English	Text	28 emotions	58,000
[77]	ISEAR	Multilingual	Text	7 basic emotions	7,000
[91]	XED	multilingual	Text	11 emotions	7500
[92]	DailyDialog	English	Text	7 emotions	13,000
[93]	EMOBANK	English	Text	VAD + categories	10,000
[89]	TEC	English	Text	6 emotions	21,000
[94]	EmpatheticDialogues	English	Text	25 emotions	25,000
[95]	EmoEvent	Multilingual	Text	6 emotions + neutral	Varies
[96]	MELD	English	Multimodal	7 emotions	13,000
[97]	IEMOCAP	English	Speech/Video	6 emotions	12 hours
[98]	RECOLA	French	Multimodal	VAD + emotion	5 hours
[99]	DepecheMood	English/Italian	Text	Emotion scores	25,000

Table 4.3 presents an overview of major emotion classification corpora that have been widely used in NLP and affective computing research. The datasets are organized by citation, name, language coverage, modality (text, speech, or multimodal), the type or number of annotated emotions, and dataset size.

This compilation includes corpora that represent a range of emotion annotation frameworks. It encompasses monolingual and multilingual resources and captures data from various modalities. Notable corpora such as GoEmotions, DailyDialog, and Multimodal Emotion Lines Dataset (MELD) have contributed significantly to benchmarking emotion classification models. Others, like Interactive Emotional Dyadic Motion Capture (IEMOCAP) and Remote Collaborative and Affective Interactions Dataset (RECOLA) provide multimodal emotional data, useful for emotion recognition beyond text. The XED dataset stands out for being extended to over 35 languages through annotation projection.

In terms of methodology, research has evolved from dictionary-based methods and classical Machine Learning to Deep Learning and transformer-based models such as BERT [90]. Advanced strategies like multi-task learning [100], zero-shot learning [101], and few-shot learning [102] are also gaining traction.

4.4 Conclusion

This chapter has established a foundational understanding of emotion analysis in NLP by highlighting its importance in interpreting affective content in textual data. We have discussed explicit emotion expressions and we have explain the classification of emotions in established psychological theories.

CHAPTER 5

EMOTION ROLE LABELING

In this chapter, we explore the semantic frame structures commonly associated with emotion, detailing how emotions are framed through specific linguistic patterns, and justify the use of FrameNet as a semantic resource for modeling emotion events. We then discuss existing Emotion Role Labeling (ERL) datasets and their coverage across languages and domains. Particular attention is given to the task of emotion stimulus detection. Finally, we highlight the specific issues encountered when applying emotional Semantic Role Labeling in low-resource languages, where the lack of annotated data and linguistic tools poses significant limitations for model development and evaluation.

5.1 Introduction

Emotion Role Labeling (ERL) is the task of automatically identifying and assigning semantic roles related to emotions in text.

Unlike traditional emotion classification tasks that assign a single emotion label to a sentence or document (e.g., joy, anger, fear), ERL attempts to answer who is feeling the emotion, what causes or triggers the emotion, and, in some cases, what contextual or topical information frames the emotional response. [103]

5.2 Semantic frames associated with emotions

In SRL field, the choice of underlying semantic resources plays a pivotal role in the effectiveness of the resulting representations. While PropBank and FrameNet are among the most widely used resources for SRL, our study adopts FrameNet as the foundational framework for integrating SRL with emotion analysis. This decision is motivated by the conceptual semantic depth that FrameNet offers.

Unlike PropBank, FrameNet adopts a frame-semantic approach that captures ”frames” associated with events, experiences, and emotions. This makes it ideal for modeling the relationships between entities involved in emotional situations, such as the experiencer, the cause, and the stimulus.

Moreover, FrameNet explicitly includes emotion related frames (e.g., Experiencer_focus, Emotion_directed, Emotions_of_mental_activity, etc.) that align with the objectives of emotion detection and analysis. Also, FrameNet enriches the representation of emotional meaning in text. Its frame-based organization makes it a more appropriate resource than PropBank or other alternatives when the task requires deep semantic interpretation aligned with emotional roles. Until now, FrameNet is the only semantic resource that provides a structured background for emotion scenarios.

Table 5.1: Emotion-Related Frames in FrameNet

Frame Name	**Description**	**Example Lexical Units**
Experiencer focused emotion	Emotion described primarily from the experiencer's perspective, without emphasizing an external cause.	feel, rejoice, suffer, enjoy
Emotion directed	Emotion is explicitly directed at a particular target or object.	admire, envy, resent
Stimulate emotion	A stimulus (often an agent or event) deliberately evokes an emotional response.	thrill, amuse, inspire
Cause emotion	An external event or situation causes emotion, regardless of intention.	devastate, depress, cheer
Emotion heat	Emotion metaphorically framed in terms of heat or physical intensity.	boil, simmer, seethe, stew
Emotions by stimulus	Emotions elicited by various stimuli, focusing on the trigger.	discover, annoy, please
Emotions of mental activity	Emotions arising from cognitive processes like realization or reflection.	surprise, realize, regret
Desiring	Emotional states related to longing, wanting, or craving.	want, desire, crave, long for

Detecting the who feel what, and towards whom is essentially a semantic role-labeling problem. The core and non-core roles in FrameNet is shown in Table 5.2. The definitions are primarily derived from FrameNet's Emotions frame and the four roles investigated in this thesis are shown in bold as commonly adopted in emotion analysis studies [1], [43].

Table 5.2: Core and Non-Core Semantic Roles in Emotion Representation

Role	**Definition**
	Core Roles
Event	Describes the situation or occurrence in which the emotional state is experienced.
Experiencer	The individual or entity who feels or undergoes the emotional reaction.
Expressor	Refers to the physical or verbal cues (e.g., gestures, facial expressions) through which the emotion is manifested.
Target	entity or aspect that is the focus of an emotion expressed in a text.
Stimulus	The cause or trigger (person, event, or condition) that provokes the emotional response.
Topic	A broader subject area or context in which the emotion is situated or directed.
	Non-Core Roles
Circumstances	The contextual background or conditions under which the emotional response takes place.
Degree	Indicates the intensity or strength of the emotional experience compared to a typical baseline.
Empathy Target	The person or group with whom the Experiencer emotionally aligns or empathizes.
Manner	Describes the specific way the emotion is felt, beyond what is captured by other roles.
Parameter	The domain or dimension in which the emotional reaction is experienced (e.g., time, social space).
Reason	Explains the rationale or justification behind the emotional reaction triggered by the stimulus.

5.3 Detailed Emotion Role Labeling example

- The sentence:

[غضب المتظاهرون من الحكومة بسبب سياساتها الاقتصادية]

- The translation:

Protesters were angry at the government over its economic policies.

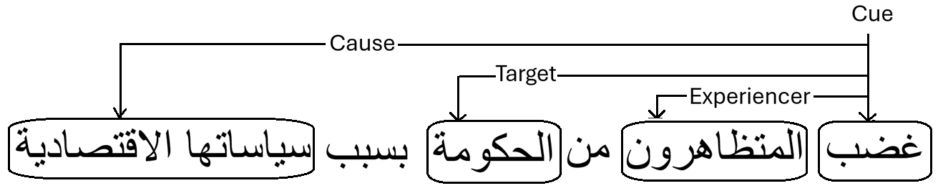


Figure 5.1: Detailed ERL example

The sentence illustrates a fully structured instance of ERL in Arabic, capturing four key semantic roles:

- Cue [غضب]: the verbal expression that signals the presence of an emotional state.
- Experiencer [المتظاهرون]: the entity experiencing the emotion.
- Target [الحكومة]: the entity that the emotion is directed toward.
- Cause [سياساتها الاقتصادية]: the reason or stimulus that triggered the emotion.

5.4 Existing Emotion Role Labeling dataset's

Table 5.3 summarizes the existing datasets for ERL, covering a range of languages, emotion classes, annotated emotion roles, and relevant notes:

Dataset	**Language**	**Emo. classes**	**Annotated Roles**	**Notes**
GNE [1]	English	7 emotion classes	Cue, Experiencer, Cause, Target	Comprehensive ERL in headlines
Elections [43]	English	7 emotion classes	Cue, Experiencer, Cause, Target	Early dataset for Emotion Role Labeling in social media
REMAN [104]	English	8 emotion classes	Cue, Experiencer, Cause, Target	Captures perceived emotion of the reader
Emo-Stimulus [103]	English	4 emotion classes	cue, Stimulus	Focused on emotion stimulus in text.
GERSTI [105]	German	10 emotion classes	cue, Experiencer, Cause	First stimulus-annotated corpus in German.

Table 5.3: Overview of existing datasets for Emotion Role Labeling (ERL).

5.5 Emotion stimulus detection

Causality refers to a semantic relationship in which one event (the effect) is viewed as the result of another (the cause). Traditionally, the identification of causal relationships relied on manually crafted [103, 105].

There is especial interested in the stimulus due to its compelling applications in understanding the reasons for emotional occurrences, such as in consumer behavior analysis and mental health treatment. It highly beneficial for systems that respond to inquiries such as "how [x] feels about [y]" or "why [x] feels [y]". It also has practical significance for text summary, since emotion expressions and stimuli are often the most information in an expressive sentence.[103]

Emotion stimulus detection has been explored primarily in resource rich languages, with strong attention given to Mandarin Chinese and, to a lesser extent, English. In Mandarin, this research area has seen considerable development with corpora such as the Weibo Emotion Cause Corpus serving as benchmarks [106]. Early work by Lee et al. [107], and subsequent extensions by Gui et al. [106] [108] and Cheng et al. [109]established foundational methods for extracting emotional triggers. In English, fewer annotated datasets exist, though efforts such as those by Neviarouskaya and Aono [110], Mohammad et al. [43], Kim and Klinger [104] and Bostan et al. [1] have contributed corpora and evaluations targeting emotion cause or stimulus detection. Other languages like Italian and Japanese have received few attention, Russo et al. [111] constructed an Italian emotion corpus from news texts, while Yada et al. [112] annotated Japanese texts from news and Q&A websites.

5.6 Emotional SRL challenges for low resource language

Emotion Role Labeling in low-resource languages faces several significant challenges due to the scarcity of linguistic resources, annotated data, and language-specific tools [113]. One of the primary issues is the lack of labeled datasets that capture the full structure of emotion, including roles such as cue, experiencer, target, and cause [1, 114]. Without such corpora, training supervised models becomes nearly impossible. Additionally, emotion lexicons (essential for identifying emotional cues) are often incomplete or entirely absent in these languages, limiting the system's ability to recognize or classify emotional expressions [113]. Cultural and linguistic differences further complicate the task, as emotions are expressed differently across societies, often using metaphors, or indirect language that may not translate well from high-resource languages like English [115]. Moreover, pretrained language models tailored for low-resource languages are either underdeveloped or unavailable, leading to suboptimal performance on ERL tasks, even when multilingual models like mBERT or XLM-RoBERTa are used, they often fail to capture the emotion expression and role structure accurately in less-resourced contexts [116]. Another difficulty arises from the complexity of detecting multiple roles in a sentence, which requires accurate syntactic and semantic parsing resources that are also limited in these languages [113]. To address these challenges, researchers employ strategies mentioned above in section 3.3 such as cross-lingual transfer learning, machine translation with projection of annotations, data augmentation, and few-shot or zero-shot learning.

5.7 Conclusion

In summary, this chapter provides a comprehensive foundation for understanding Emotion Role Labeling by examining how emotions are represented within the framework of semantic frames [22], particularly those derived from FrameNet. By surveying existing ERL datasets and focusing on the challenges of emotion stimulus detection. Furthermore, we underscore the difficulties faced in low-resource language settings.

CHAPTER 6

LITERATURE REVIEW

This chapter delves into the research area of Semantic Role Labeling (SRL) for emotion classification task portraying the latest scientific findings in this field, we also conducted a comprehensive assessment of the literature, identifying, evaluating, and analyzing the most important studies conducted in the many contexts of SRL, including the integration of SRL and emotion identification. The remainder of this chapter is organized into three main sections. The first section provides an overview of SRL, where we categorize existing studies based on the learning methods employed, ranging from the most recent to earlier approaches. We also distinguish between models built using closed and open LLM. The second section discusses existing studies of SRL in Arabic. The final section focuses on research that integrates SRL with emotion identification, highlighting efforts that address the detection of emotional roles. Special attention is given to studies on emotion cause extraction, as this subtask has gained significant interest due to its relevance in fine-grained emotion understanding.

6.1 Semantic Role Labeling

This section provides a comprehensive review of previous research on SRL, beginning with traditional feature-based approaches, followed by classical ML methods. The discussion then moves to neural SRL models, covering both end-to-end architectures and transformer-based techniques, and concludes with recent advancements that utilize LLM. Special attention is given to studies focused on Arabic SRL.

6.1.1 Rule-based SRL

The history of SRL traces back to early linguistic theories of thematic roles in the period between 1960 and 1970 by many scientists, headed by Charles Fillmore, which introduced the *Case Gramma*r framework in his paper ”The Case for Case” [117], which suggested that verbs determine the syntactic and semantic structure of sentences by governing specific roles, or ”cases.”

In the late 1970s, he expanded his work with the concept of frame semantics [22], this theory posits that words elicit conceptual structures referred to as ”frames”.

The need for large-scale annotated data led to the development of resources like FrameNet and PropBank. The FrameNet project in 1998, was a significant achievement in the NLP field; it provided an extensive annotated database with frame elements linked to verbs.

At the beginning of the 21st century, Martha Palmer et al. [27] created *PropBank*, moving away from particular frames and instead concentrating on directly naming the predicate’s arguments. The authors have established a set of more generalized roles (e.g., ARG0, ARG1) for each verb.

Rule-based SRL has several advantages, first in its interpretability, the manual designed rules ensure transparency in decision-making. furthermore, it provides high linguistic accuracy, as well-formed rules effectively capture precise syntactic and semantic patterns. This approach also performs well in specialized domains, such as legal or medical texts. However, rule-based SRL has notable limitations, including a lack of generalization, as handcrafted rules are not easily transferable across languages. Moreover, maintaining such systems is costly. In addition, Scalability is another challenge, as expanding rule-based SRL to cover diverse linguistic phenomena requires a lot of manual effort, making it less suitable for large-scale applications.

6.1.2 Classical Machine Learning-Based SRL

The reliance of SRL on hand-crafted rules and linguistic theories required human labor to specify the relationships between predicates (verbs) and their arguments (roles like Agent, Patient). These approaches were frequently rigid and faced challenges in expanding their capacity. Adopting ML was the solution.

Traditional methods used linear classifiers, such as Support Vector Machine (SVM) or Maximum Entropy models with feature templates, to extract a variety of features from syntactic parses.The seminal work in SRL introduced by Gildea and Jurafsky (2002) [118], used feature-based models to develop the first automatic SRL in sentences. Based on FrameNet, they extracted a variety of syntactic and lexical features and trained a statistical models. To enhance traditional SRL (which involves full syntactic parsing) accuracy, Pradhan et al [119] combined multiple syntactic views, proposed a chunking-based method, and trained a ML model. This method reduces the intricacy of conventional parsing-based SRL and enhances the system's robustness to syntactic parsing errors. They found that the primary source of wrong prediction was the syntactic parser.

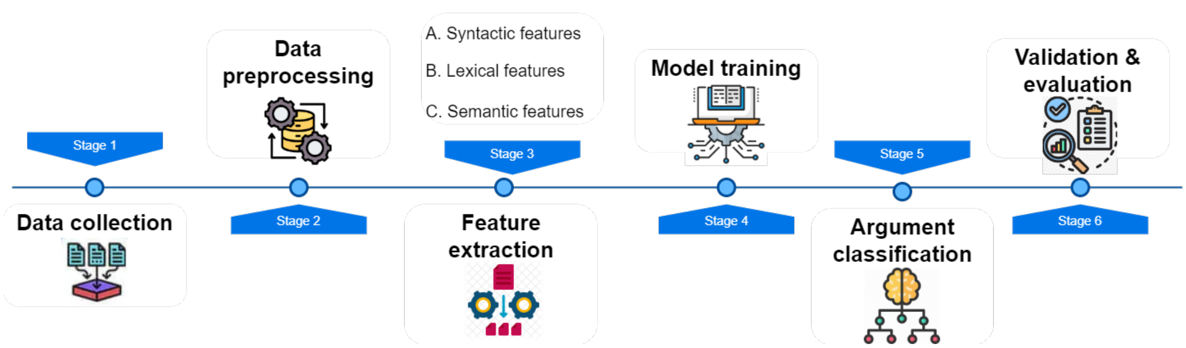


Figure 6.1: Statistical and Machine Learning-Based SRL pipeline.

6.1.3 Neural Network-Based SRL

Based on previous studies, other researchers chose another way to address the SRL problem. For instance, in [120], they gave up the combination of ML algorithms and hand-crafted features and adopted unified deep neural networks to build a multi-task learning system.

The authors in [121] present a three-step pipeline for SRL that uses a dependency parser to detect predicate arguments in various languages. The system was assessed in seven languages and obtained an average labeled semantic F1 score of 80.80.

Yuan [122] proposed incorporating valence structures into deep neural networks for SRL, particularly improving performance in Chinese by capturing syntactic and semantic relations more effectively. The model achieved an F1 score of 93.69% for verbal predicates and 79.23% for nominal predicates on Chinese datasets.

Fei et al. [123] introduced a transition-based neural model that jointly performs predicate identification and argument labeling. The model achieved state-of-the-art results on CoNLL -2009 and Universal Proposition Bank datasets. Munir et al. [124] developed an unsupervised neural SRL model that decomposes the task into argument identification and clustering. The model employs a Bidirectional Long Short-Term Memory (BiLSTM) with an adversarial layer and achieved competitive performance on the CoNLL -2009 English dataset without relying on annotated data. Devianti and Miyao [125], in 2024 explored the use of GCN for zero-shot cross-lingual SRL. Their study found that incorporating universal dependency trees enhances model transferability across 23 languages.

Table 6.1: Summary of Selected Neural Network-Based SRL Studies

Ref	Problematic	Objective	Dataset Used	Learning Algorithm	Key Finding	Challenges
[120]	Lack of unified neural models for NLP tasks	build a multi-task learning system using deep neural networks	PropBank, WSJ	CNNs and CRF	Showed that DL can handle various NLP tasks with minimal feature engineering	Requires large amounts of data
[121]	accurately determining and identifying the semantic roles of words in sentences across many languages	Build a consistent multilingual SRL framework	CoNLL -2009 Shared Task Dataset	linear classifiers	Demonstrated multilingual SRL is feasible with a shared schema Joint models outperform pipeline systems in multilingual settings	Variation in annotation schemes; dependence on high-quality syntactic parsers
[122]	Lack of syntactic structure in Chinese SRL systems	Introduce valence structure into neural SRL to improve accuracy	Chinese PropBank (CPB)	BiLSTM	Valence integration boosts performance on Chinese SRL	Difficulty of combining valence structures with end-to-end learning
[123]	Error propagation	Build an end-to-end model that jointly detects predicates and labels arguments	CoNLL -2009, PropBank	LSTM	Achieved better performance via joint modeling	long-distance arguments
[124]	Reliance on large labeled corpora for training SRL	Develop unsupervised neural SRL using clustering techniques	CoNLL -2009	BiLSTM	Effective SRL results without annotated data	Lower performance
[125]	SRL models struggle with cross-lingual generalization	Use GNN to incorporate syntax for better transfer	Treebanks	GCN	Enhanced cross-lingual SRL accuracy through syntactic structure	Aligning syntax/semantics across languages; domain variation

6.1.3.1 End-to-End SRL

Recent advancements in end-to-end deep models for SRL without syntactic input [56], [126], [57]. As pioneering work, Zhou and Xu [56] address the problem of using traditional SRL techniques that depend on supervised learning using syntactic features, which necessitate significant engineering efforts and risk prediction mistakes. The authors propose a deep Bidirectional Long Short-Term Memory (BiLSTM) model that processes raw text without syntactic input, evaluated on CoNLL -2005 and CoNLL -2012. The model uses pre-trained word embeddings and a Conditional Random Field (CRF) layer. Experiments prove that augmenting the depth of LSTM layers enhances performance; the first shared task achieved an F1 score of 81.07 and the second 81.27. He et al. [57] examine enhancements in SRL by applying a deep highway bidirectional LSTM (BiLSTM) model with constrained decoding. The model eliminates the necessity for syntactic parsing and attains a 10% relative error reduction compared to prior state-of-the-art systems on the CoNLL 2005 and 2012 datasets. The successes of end-to-end models demonstrate the potential of LSTMs to manage the underlying syntactic structure of sentences.

Despite these successes, neural network models have limitations, especially when applied to complex tasks like SRL. First, recurrent models capture sequential information within data and combine each word with its preceding hidden state recursively. RNN are applicable for sequential prediction; however, it is susceptible to the vanishing gradient problem. That's why it has difficulties in capturing long-range dependencies because they have to compress all the relevant information into a fixed-size hidden state. This compression can lead to information loss, especially with lengthy sequences. In addition to the problem of memory compression, the model necessitates larger memory ability to retain information for long sentences [127], [128].

Traditional neural models for SRL depend on extensive annotated datasets, making the application of these methods challenging in low-resource languages where labeled data is limited or expensive to obtain. Additionally, integrating unlabeled data into training is difficult, which makes it still underutilized. To tackle these challenges, [129] propose a semi-supervised SRL method that uses unlabeled data and applies syntactic constraints to enhance model outputs. The authors use the CoNLL -2012 dataset and integrate a syntactic-inconsistency loss function with cross-entropy loss to jointly train on both labeled and unlabeled data, using ELMo embeddings and a BiLSTM. The results show significant improvements, demonstrating the effectiveness of this method in enhancing SRL performance.

Table 6.2: Summary of some studies in neural Semantic Role Labeling

Ref	Problematic	Objective	Dataset Used	Learning Algorithm	Key Findings	Challenges
[56]	Traditional SRL depends heavily on syntactic parsers	Develop an end-to-end SRL system without syntax input	CoNLL -2005 CoNLL -2012	Deep BiLSTM	Showed that deep BiLSTM can learn SRL effectively without syntactic features	Handling long-range dependencies
[57]	Limitations in prior SRL models using heavy syntactic features	Investigate DL models with minimal linguistic features	CoNLL -2005 and CoNLL -2012	Deep BiLSTMs with highway connections	Achieved state-of-the-art results using only word and predicate embeddings	Limited reliance on large annotated corpora
[127]	Need for better machine reading systems	Integrate LSTM for machine reading in SRL	Penn Treebank	Long Short-Term Memory (LSTM)	Demonstrated that LSTM-based models perform well on reading comprehension and SRL	Generalization to complex semantic structures
[129]	Lack of annotated data limits performance of deep SRL models	Propose semi-supervised learning to reduce dependency on labeled data	CoNLL - 2005, CoNLL -2012	Deep BiLSTM	Semi-supervised approach improves SRL performance, especially in low-resource scenarios	

6.1.3.2 Transformer based SRL

Recently, a new model was introduced, called Transformer, a Deep Learning architecture designed for sequence-to-sequence tasks. It replaces recurrent layers with a self-attention mechanism, allowing the model to weigh the importance of different words in a sentence simultaneously rather than sequentially [130].

Vaswani et al. [130] offer a method for concurrently capturing dependencies across whole sequences. Self-attention layers enable each word in a sentence to pay attention to every other word, despite their distance. The vanishing gradient issue is resolved by this non-sequential processing. Compared to recurrent networks, self-attention mechanisms are more effective and scalable for huge datasets because they permit parallel processing [131].Transformers mark a significant improvement to many tasks, including machine reading [127], self-attentive sentence embedding [132], abstractive summarization [133], language understanding [134], and neural machine translation [130].

SRL has experienced considerable progress due to the transformer-based models. These advancements have enabled models to more effectively capture long-range dependencies and hierarchical linguistic structures, which is very essential for accurate SRL. in order to enhance the accuracy and performance of SRL, Tan et al. [58] propose a model that overcomes the commonly known problems of RNNs, such as the sequential restrictions in order to capture long-range dependency. Key components include positional encoding to maintain sequence order, Begin-Inside-Outside Tagging Scheme (BIO) tagging for easier role classification, and multi-head attention to record a variety of interactions. The model demonstrates the efficacy of self-attention in SRL and difficult NLP tasks by achieving state-of-the-art F1 scores tested on benchmark datasets.

Another innovative approach in this field, Strubell et al. [128], incorporates syntactic information within a self-attention framework. LISA employs multi-task learning for tasks such as dependency parsing, POS tagging, predicate detection, and SRL, in order to learn syntactic dependencies directly within the SRL model (allowing the model to utilize syntax efficiently without requiring considerable pre-processing), yielding accurate SRL performance.

In [131], the self-attention mechanism was guided by dependency trees. In order to capture pertinent linguistic structures within sentences, the authors incorporate syntactic information directly into the model, hence boosting SRL accuracy. This methodology has been especially efficacious for languages with complex syntactic structures, such as Chinese.

Another notable model [135] introduces a novel deep neural model that minimizes the necessity for recurrent updates by selectively linking attentive representations within the model. This approach efficiently mitigates the computational difficulties of recurrent models by implementing selective connections that encapsulate hierarchical dependencies, decreasing training duration by more than 60% while preserving good accuracy on benchmark datasets (e.g., CoNLL 2005 and CoNLL 2012). This methodology has established a standard for SRL models by integrating efficiency with robust hierarchical representation learning.

Table 6.3: Summary of studies on self-attention-based models for SRL

Ref	Problematic	Objective	Dataset Used	Learning Algorithm	Key Finding	Challenges
[130]	Sequential architectures are limited in capturing long-range dependencies	Introduce Transformer model based on self-attention	WMT 2014 English-German, English-French (MT tasks)	Transformer architecture (multi-head self-attention)	Self-attention significantly improves translation quality and training efficiency	Requires large amounts of data and computational resources
[58]	Traditional SRL models rely heavily on syntactic features	Apply self-attention to SRL without explicit syntax	CoNLL - 2005, CoNLL -2012	RNN, CNN, FFN	Achieves strong results without relying on syntactic parsers	Handling implicit syntactic structures without external parsers
[128]	Lack of syntactic guidance in self-attention models for SRL	Integrate linguistic features into self-attention layers	CoNLL - 2005, CoNLL -2012	Linguistically-informed Self-Attention (LISA)	Integrating syntax improves SRL accuracy over plain attention models	Dependency on quality of predicted syntax
[131]	Conventional SRL models are predominantly dependent on syntactic features	Use syntax trees to enhance attention scores	CoNLL - 2005, CoNLL -2012	Syntax-Enhanced Self-Attention	Syntax-aware attention improves SRL performance, especially on long-range dependencies	Complexity in integrating tree structures efficiently
[135]	Not all self-attention connections are equally important	Propose Selectively Connected Self-Attention to focus on critical parts of input	CoNLL - 2005, CoNLL -2012	Selectively Connected Self-Attention mechanism	Selective attention helps focusing in relevant input tokens for better role labeling	Balancing sparsity in selective connections

6.1.4 Large Language Models

With the emergence of LLM, SRL has seen significant advancements, particularly with closed-LLM like OpenAI's GPT, Google's Gemini, and Meta's LLaMA. These models, trained on massive corpora, offer powerful contextual representations and robust generalization across languages. Unlike open-source models, closed LLM benefit from extensive pretraining and access to vast computational resources, making them highly effective for SRL tasks.

Chat Generative Pre-trained Transformer (ChatGPT) has gained significant attention due to its applicability in various fields, particularly those related to NLP tasks [136], [137], such as question answering [138], [139], text summarization [140], machine translation [141], sentiment analysis [142], [143], and information extraction [144]. However, annotation tasks using ChatGPT remain anew and poorly explored in comparison to other NLP applications.

Cheng et al. [145] investigate the capability of LLM, such as ChatGPT, for SRL task using prompts without fine-tuning (few-shot learning), in addition to comparing the efficiency of ChatGPT with supervised methods. The SRL challenge has been presented as a natural language question-answering problem, ChatGPT attains comparable outcomes on CoNLL -2005 and CoNLL -2012.

Sun et al. [146] highlight ChatGPT's potential by examining a variety of NLP tasks, including SRL. In order to evaluate how effectively ChatGPT can adapt to this task without any training, they present techniques like "self-verification" to enhance SRL outputs. It also shows how techniques like paraphrasing and multi-prompt reasoning enhance ChatGPT's functionality.

These studies demonstrate that while ChatGPT provides promising results for SRL, it depends significantly on prompt design and task formulation to address its limitations in understanding nuanced semantics. Boutouta et al. [147] investigates implicit emotion recognition by leveraging large language models to infer emotions from contextual and semantic cues rather than explicit affective expressions. The results demonstrate that LLMs effectively capture complex contextual signals, outperforming traditional emotion detection approaches.

6.2 Semantic role labeling for Arabic

The earliest attempt to develop an SRL system for Arabic was introduced by Diab et al. [148]. Their approach adapted successful methodologies from English SRL, employing a supervised learning framework based on (SVM). The system was designed to perform both argument boundary identification and role classification tasks.

The proposed method of Diab et al. [149] was trained and evaluated using the dataset provided by SemEval 2007 Task 18 on Arabic Semantic Labeling. This dataset is built upon Arabic PropBank and Treebank resources and includes annotations for 95 verbs extracted from the Treebank. In this initial work, the system achieved promising performance, reporting an argument detection accuracy of 94.06% and an argument classification accuracy of 81.43%.

Following this, Diab et al. [150] leveraged the rich morphological characteristics of the Arabic language. Their system employed SVMs in combination with kernel-based methods, and was developed using the Arabic PropBank corpus. This enhanced methodology led to a significant performance improvement, achieving an argument classification accuracy of 82.17%.

In 2016, Lakhfif et al. [39] [151] [152], [153], [154] proposed a computational approach to Arabic text analysis and machine translation based on frame semantics. Their work involved building a semantic parser capable of extracting deep semantic structures at lexical, morphological, syntactic, and semantic levels to translate Arabic into Algerian Sign Language. To support this, they constructed an Arabic FrameNet by developing a rule-based method that leverages mappings between Arabic WordNet and FrameNet. As a result, they automatically generated over 8,000 Arabic lexical units and more than 600 semantic frames.

Meguehout et al. [155] proposed a SRL technique based on Case-Based Reasoning (CBR). Their method relies on Arabic PropBank as a source for semantic label and was evaluated on a corpus comprising 2,332 attributes and 5,291 arguments. The results highlighted the model's effectiveness in assigning semantic roles by identifying similarities with previously annotated instances, and they emphasized the importance of specific features in enhancing annotation quality.

Senator et al. [46] presented a corpus of Arabic emotional text annotated with cue, experiencer, cause and target. The proposed corpus of 3000 Arabic tweets was annotated using semi-automatic tool initially developed by Lakhfif et al. [151].

Table 6.4: Summary of SRL Studies on Arabic Language

Ref	Problematic	Objective	Dataset Used	Learning Algorithm	Key Findings	Challenges
[148]	Lack of resources for Arabic SRL	Build a basic SRL system for MSA	Annotated MSA corpus	Rule-based + supervised classifiers	First SRL prototype for Arabic	Limited annotated data
[149]	Evaluation of Arabic SRL systems	Provide benchmark for Arabic SRL	SemEval-2007 Task	SVM	Created first shared task for Arabic SRL	Sparse features and ambiguity in Arabic
[150]	Improve SRL for Arabic with ML	Apply kernel methods to Arabic SRL	SemEval-2007	Kernel-based SVMs	Syntactic features improve SRL performance	Lack of robust Arabic parsers
[151]	SRL for MT to sign language	Introduce Arabic FrameNet	Arabic text corpus	Rule-based	Frame semantics effective for translating Arabic into Sign Language	Arabic morphology
[155]	Enhance Arabic SRL using CBR	Case-based role labeling system	Arabic PropBank-style data	Case-Based Reasoning	CBR adaptable to new examples	Retrieval of relevant past cases
[46]	Lack of Emotion-SRL resources	Extend SRL to emotion-labeled tweets	Manually labeled Arabic tweets	–	Promising results in emotional SRL	Noise and informal language in tweets

6.3 Emotion Role Labeling

Laura Bostan in [1] addresses the challenge of the lack of detailed datasets linking emotions to their semantic roles in news headlines. To this end, the GoodNewsEveryone (GNE) was created, a novel dataset which provides fine-grained annotations of emotional arguments. The corpus consists of 5000 English news headlines annotated with emotion categories (joy, sadness, anger..), semantic roles (experiencer, target, cue..) , and reader perception. Crowd workers were used to annotate the dataset following high-quality guidelines. They formulate the task as sequence labeling of each role with a bidirectional long short-term memory networks with a Conditional Random Field (CRF) layer (BiLSTM-CRF) that uses ELMo embeddings as input to represent each word in context, and an Inside–Outside–Beginning (IOB) format as output. The dataset focuses on individual news headlines, which are often short and accurate, this makes it more difficult to understand the semantic roles and identify emotions since the subtleties may depend on details not found in the headline.

The main issue in SRL4E [42] is that the existing datasets and methods are fragmented,

with diverse annotation schemes, domains, and languages, making it difficult to compare and generalize models across different contexts. The authors seek to address these gaps by providing a unified evaluation framework (SRL4E) that harmonizes 6 datasets, Creating a basic neural baseline, made up of a stack of BiLSTM layers and a word representation module based on BERT. They find 56.5 F1-score.
The problematic addressed by Oberlander et al. in [156] revolves on comprehending how semantic roles might enhance NLP's emotion detection tasks. They investigate the relationship between ML models' capacity to deduce emotions from text and semantic roles. In order to correctly identify emotions, the authors examine which semantic roles (Experiencers, Stimuli, and Targets) are most important. The five datasets used include both emotion annotations and SRL information. They train BiLSTM with and without explicit semantic role, the results show that experiencer is the most critical role for emotion detection, followed by stimulus, while target is less significant.

By introducing semantic roles related to emotions in literary works, Kim et al. [104] tackles the problem of recognizing emotions in texts. The issue is that traditional emotion analysis frequently ignores the crucial semantic roles that provide deeper context. To fill this gap, the authors introduce Relational Emotion Annotation for Fiction (REMAN) by assembling a collection of fictional text and annotate the sentences manually with emotion category and semantic roles. They identified four roles: emotion (such as joy, sadness, or rage), cause, experiencer, and target. Inter-annotator agreement was measured to ensure the reliability of the annotations. The authors automated the extraction of these roles using Conditional Random Fields (CRFs) and BiLSTM-CRF as a baseline. The findings demonstrated that although the baseline models can identify some roles with an acceptable accuracy, there are still major obstacles to overcome, like handling complex sentence structures and implicit emotional cues.
In their study, Laura Oberländer and Roman Klinger [157] investigate the effectiveness of two approaches—token sequence labeling and clause classification—in detecting emotion stimuli in English texts. Although token sequence labeling and clause classification have been widely used in Mandarin and English, respectively, there hasn't been a direct

comparison of these approaches in English. The researchers developed an integrated framework to evaluate both approaches on four English datasets from diverse domains. Their findings indicate that token sequence labeling outperforms clause classification in three out of four datasets. They suggests that token sequence labeling is generally more effective for emotion stimulus detection in English.

Dang et al. [105] focus on emotion stimulus detection in German news headlines, introducing the GERSTI corpus with 2006 annotated headlines. It explores emotion categories, experiencers, and stimuli at the token level. A heuristic method is used to project annotations from English to German. The study evaluates CRF and XLM-RoBERTa, finding that in-corpus training outperforms cross-lingual models. XLM-RoBERTa achieved the best results.

In [43] the authors analyze tweets to detect emotions, experiencers, and stimuli, they use tweets from the 2012 US presidential elections as a dataset. Several classifiers used in this study, including: SVM, CRF, NB, BiLSTM ... ect. SVM classifier achieves 58.30% F-score in emotion classification. Identifying emotion stimuli is challenging due to informal language and limited context in tweets.

6.3.1 Emotion-cause extraction

Emotion Cause Extraction (ECE), the process of identifying the potential causes behind certain emotions in text, has attracted a lot of attention in recent years due to its wide applications.

While traditional emotion detection focuses on recognizing the type of emotion (e.g., happiness, anger, sadness), ECE goes a step further by determining what event, entity, or situation caused the emotion. Emotion stimulus detection received substantial attention in several languages. In Mandarin Lee et al. [107] constructed the first corpus for emotion cause detection, by extracting 6,058 entries of sentences, and created linguistic rules to extract emotion stimuli. Another study by Li and Xu [158] which built a dataset based on the Weibo website, identifying seven types of emotions, containing 16,485 posts and 1,305 posts with causes. Gui et al. [106] collected data from the same source. While, Cheng et al. [109] constructed a dataset for Emotion Cause Extraction

Table 6.5: Comparative Summary of Emotion Semantic Role Labeling Studies

Ref	Problematic	Main Focus	Dataset Used	Learning Algorithms	Key Findings	Limitations
[1]	Lack of emotion corpora that include both SRL and perception layers	Introduce a novel corpus supporting emotion analysis and SRL	GNE	BiLSTM-CRF	News headlines convey complex emotion semantics; F1-score: 0.27	Ambiguity in emotion perception, complexity in identifying semantic roles
[42]	Heterogeneous annotation formats across datasets	Create a unified evaluation framework for SRL and emotion analysis	GNE, REMAN , Blogs, Elections, NTCIR, EmoTweet	BERT-based, BiLSTM	Harmonized evaluation improves reproducibility; F1 (emo.class): 0.64; F1 (SRL): 0.6	Standardizing evaluation across diverse datasets remains complex
[104]	Comparison between token-based and clause-based stimulus detection	Compares token-level sequence labeling and clause classification for emotion stimulus detection	Emotion-Stimulus, ElectoralTweets, GNE, Emotion Cause	CRF, BiLSTM	Token-based models outperform clause-based in precision	Ambiguity in stimulus spans and difficulty in modeling long-distance dependencies
[105]	Need for SRL schema in low-resource settings (German)	Detect emotion stimuli in German headlines; evaluate in-corpus vs. cross-lingual learning	GERSTI corpus	CRF, XLM-RoBERTa	In-corpus training outperforms cross-lingual models	Ambiguity in emotional triggers, implicit expressions
[156]	Unclear role of semantic roles in aiding emotion inference	Investigates which roles are most informative for emotion classification	REMAN , ElectoralTweets, Emotion Cause, GNE, Emotion Stimulus	BiLSTM	Stimuli and targets play key roles in emotion prediction	Identifying clause boundaries and linking token/role-level data
[104]	Need for rich SRL annotation in literary texts	Introduces REMAN corpus to identify triggers and semantic roles in fiction	REMAN	CRF, BiLSTM-CRF	REMAN supports SRL modeling for emotion-rich literary texts	Genre-specific language may limit generalizability
[43]	Lack of Arabic emotion SRL resources in informal settings	Apply SRL for analyzing emotions in Arabic tweets	ElectoralTweets	SVM	SRL improves emotion detection; stimuli often implicit; dataset is a valuable resource	Informal language, code-switching, lack of Arabic SRL tools

(ECE) by using the Chinese microblogs.

Only few research have been introduced for English, Ghazi et al. [103] focus on identifying the causes behind expressed emotions in text, they develop baseline systems which work with intuitive features. other studies was presented, we cite Neviarouskaya and Aono [110], Mohammad et al. [43], Kim and Klinger [104], Bostan et al. [1].

Russo et al. [111] worked on a corpus for Italian news texts, whereas Yada et al. [112] annotated Japanese phrases from news stories and question-and-answer websites.

Table 6.6: A Comparative Overview of Existing Studies on Emotion Cause Extraction (ECE)

Ref	Year	Language	Source	Scale
[107]	2010	Chinese	Sinica	6058
[111]	2011	Italian	Newspaper	6000
[106]	2014	Chinese	Weibo	1300
[158]	2014	Chinese	Weibo	16500
[159]	2015	Chinese	Sina	16300
[103]	2015	English	social networks	820
[160]	2016	Chinese	Sina news	2000
[161]	2017	English & Chinese	English novel & Sina news	5000
[112]	2017	Japenese	Newspaper, web news	300.000
[104]	2018	English	Fictional texts	1720
[1]	2019	English	News headlines	5000
[162]	2020	Chinese	Sina news	2085

6.4 ChatGPT in Cross-lingual projection

Cross-lingual projection consists of transferring information from a source language to a target language. Cross-lingual projection has garnered considerable attention from multilingual communities, due to its importance in NLP, particularly for creating new datasets for low-resourced languages. In this section, we introduce the most contributing articles in the field of cross-lingual projection.

Pado et al. [163] introduced a framework designed to induce high-precision semantic role annotation in German through the projection of English FrameNet annotations.

Moreover, Khondaker et al. [164] investigated ChatGPT's performance on 44 distinct Arabic NLP tasks over 60 different datasets under various shot settings. The authors suggest that ChatGPT performs well in English, but gives poor results in smaller Arabic models but struggles with Arabic dialects. Additionally, ChatGPT's comparison with BLOOMZ in summarizing and news title generation underscores the need for rigorous evaluations across various languages. However, concerns arise from the exclusion of massive language models like PaLM (540B) for efficiency and poor performance on open-domain dialectal dialogue tasks.
In the same context, Lai et al. [165] evaluated ChatGPT's performance across various languages for NLP tasks. They observed reduced performance for languages other than English and highlighted the necessity for task-specific models to improve multilingual learning. Regarding cross-lingual summarization (CLS) task, Wang et al. [166] evaluated the zero-shot performance of ChatGPT models on three widely-used CLS datasets, highlighting GPT-4's leading performance in this domain. Meanwhile, Zhang et al. [167] explored the multilingual abilities of LLMs, employing prompt translation and response back-translation methods to assess response accuracy across different languages, revealing the nuanced challenges of subordinate multilingualism faced by models like GPT.

6.5 Conclusion

This literature review chapter serves as a foundation for understanding the current landscape of SRL in the context of emotion classification. By systematically reviewing and categorizing existing approaches (from traditional learning algorithms to modern transformer-based and LLM-powered models), we provide a basis for appreciating current trends and identifying research gaps. The final synthesis underscores the growing importance of integrating SRL with emotion analysis, especially in subtasks like Emotion Cause Extraction, which are crucial for developing more interpretable and context-aware emotion analysis systems.

Part II

Contributions

CHAPTER 7

ARABIC EMOTION ROLE LABELING CORPUS

7.1 Introduction

Modern full-scale Natural Language Processing (NLP) applications rely on large sets of annotated data. However, such essential resources are available primarily for highly studied languages such as English. In contrast, for Arabic—considered a low-resourced language—there is still a significant need for relevant datasets and tools. Only a few studies in the literature have attempted to integrate SRL with emotion analysis. Therefore, exploring different techniques and approaches to overcome this limitation remains an important research question.

To the best of our knowledge, no existing study has developed resources or models for identifying emotion role labeling in Arabic. We aim to address this gap by adopting the FrameNet approach to enrich emotional text with semantic roles. Our choice is motivated by the widespread use of FrameNet in NLP. In this contribution, we present in detail the construction and annotation of an Arabic emotional text corpus, with particular emphasis on the relationship between semantic roles and emotional arguments. Figure 7.1 illustrates an example of an annotated tweet, highlighting the cue (in grey), the experiencer (in purple), and the cause (in red).

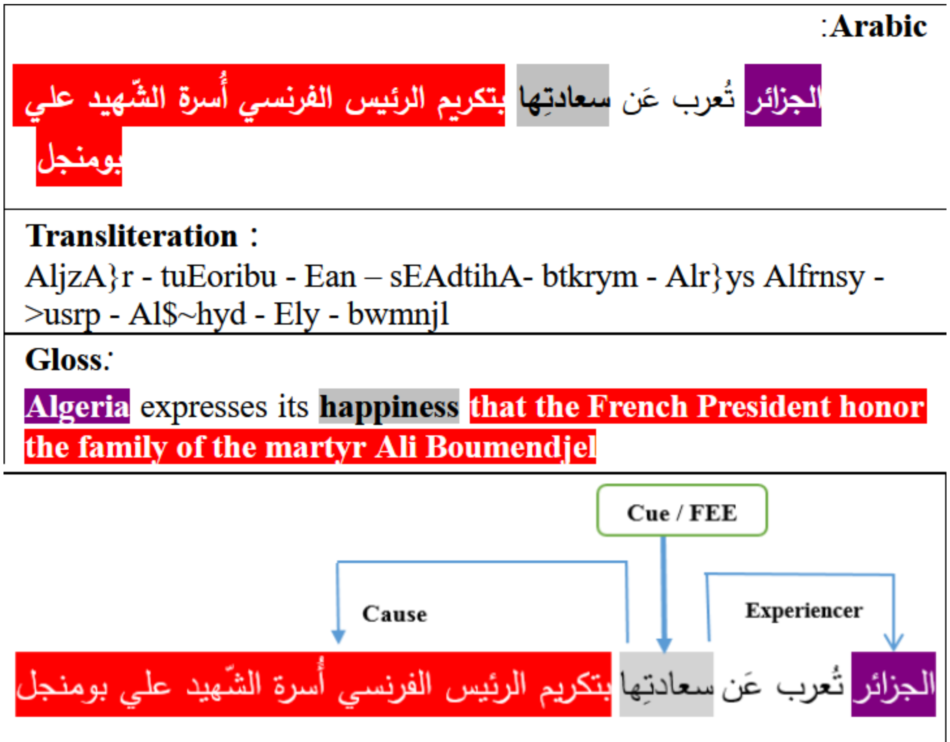


Figure 7.1: Example of an annotated tweet with emotion arguments.

Despite the progress made in computational resources for Arabic, emotion analysis of Arabic text still lacks annotated corpora, especially those offering a fine-grained description that captures the full structure of an emotion.

Our main contribution is a multi-level annotation procedure. In the first stage, we annotate semantic roles based on the FrameNet framework, using a FrameNet-like annotation tool for the Arabic language [39], [152], [151]. In the second stage, once the cue words are detected, we annotate sentence polarity and define emotion categories. Finally, we annotate the structure of the emotion expression with relevant arguments such as stimulus, experiencer, cause, topic, degree, and so on. These arguments are largely described through semantic frames.

Our contributions in terms of annotation can be summarized as follows:

- Annotation of words denoting emotion categories.

- Annotation of expressions denoting semantic roles using FrameNet frame elements.

- Annotation of the relationship between expressions denoting semantic roles and the words denoting the emotion (cues), with labels such as stimulus, experiencer, cause, etc.
- Inter-annotator validation to assess the quality of the annotation results.

7.2 Research methodology

The main goal of this study is to provide a corpus annotated with semantic role labels and emotion-related arguments to improve Arabic NLP tasks. As a first step, we focus on short texts from Twitter posts. Tweets are generally less formal (often colloquial), contain grammatical errors and typos, and require cleaning before use as raw data.

7.2.1 Corpus creation and annotation process

7.2.1.1 Data collection

The data collection phase consists of selecting representative expressions from Twitter that meet a set of predefined criteria and cover diverse Arabic language phenomena.
Our methodology for constructing the dataset of tweets is described as follows: First, we focus on four emotion classes selected from Ekman's list (joy, sadness, anger, and fear), in addition to a neutral class that contains tweets not belonging to any of the aforementioned categories. For each emotion class, we select a set of terms associated with different intensities of that emotion. For example, the joy class (فرح، سعادة، ابتهاج، سرور، مرح، بهجة، غبطة) includes multiple related terms.
Similarly, for the fear class (خوف، ذعر، رعب، وجل، هلع، خشية، وجس، فزع، وهل، ارتياع), we select a range of representative terms. The identification of the query words for each emotion is carried out using the Oxford English–Arabic Dictionary.
The Arabic tweets were collected using the Twitter Search API during September and October 2021, based on the hashtags of the selected query words. To obtain newly posted tweets, we queried the Twitter Search API every six hours. We then applied standard preprocessing techniques, including removing hashtags, user mentions (@user),

numbers, special characters and punctuation, stickers and emoticons, and eliminating hypertext links (URLs) using regular expressions and text-cleaning functions.

7.2.1.2 Annotation tool

Figure 7.2 shows a screenshot of the annotation tool, which was developed using the Java programming language to address specific challenges posed by the Arabic language [39], [152]. We use a multi-level annotation tool [151] to generate different linguistic levels of information, ranging from fine-grained POS tags and syntactic relations to semantic roles (SR). In addition to POS tags, words are annotated with their semantic class using SUMO, their English equivalent, and WordNet SYNSET. Following the FrameNet annotation methodology, we annotate arguments with the corresponding SR (frame elements), grammatical function, and functional phrase type.

Figure 7.2: A screenshot of the annotation tool.

7.2.1.3 Emotions category

The annotation phase lasted several days using the annotation tool. The annotation team consisted of four annotators (three PhD students) and one expert native Arabic speaker. They independently annotated tweets with the fundamental emotion classes (joy, sadness, anger, fear, and neutral) by selecting the target word (cue) and the corresponding emotion class, in addition to the associated polarity of the emotion. Each tweet was assigned one of the five emotion labels. If a tweet could fall under more than one category, the annotators chose the most appropriate one. For example:

يوم إعلان النتيجة **بكيت**(Sadness) من شدة **الفرح** (Joy)
On the day the results were announced, I was so delighted that I cried.

In the example above, there are two words that indicate emotions: ”بكيت bakaytu – cried-I” and ”الفرح Alfaraho – Joy,” which convey opposite senses, while the sentence as a whole belongs to the joy category. The collected data comprised 45,000 tweets, but the annotators selected only 3,010 emotional Arabic tweets that are meaningful and correctly spelled.

7.2.1.4 Semantic role labeling

The annotation procedure of the corpus is based on semi-automatic annotation of tweet sentences, considering several layers of description. In addition to emotion categories, our annotation tool provides multiple layers of linguistic information.
We select relevant text spans from each tweet that can be used in the annotation process. After selecting the corresponding entities proposed by the tool, we annotate semantic roles such as experiencer, cause, stimulus, and target to describe the relationships between entities representing semantic arguments and the emotion word (cue).

7.2.2 Validation and results

7.2.2.1 Validation of the Dataset

The annotation tool records three distinct annotations. For emotion annotation, if all three annotators indicate the same category for a sentence, the tool accepts and validates it. If two annotations are identical while the third differs, the tool validates the majority category. Otherwise, the decision is deferred to the expert. For SRL annotation, the system proposes a set of possible solutions from which the annotator selects the most probable one. In cases where two out of three annotators agree, the

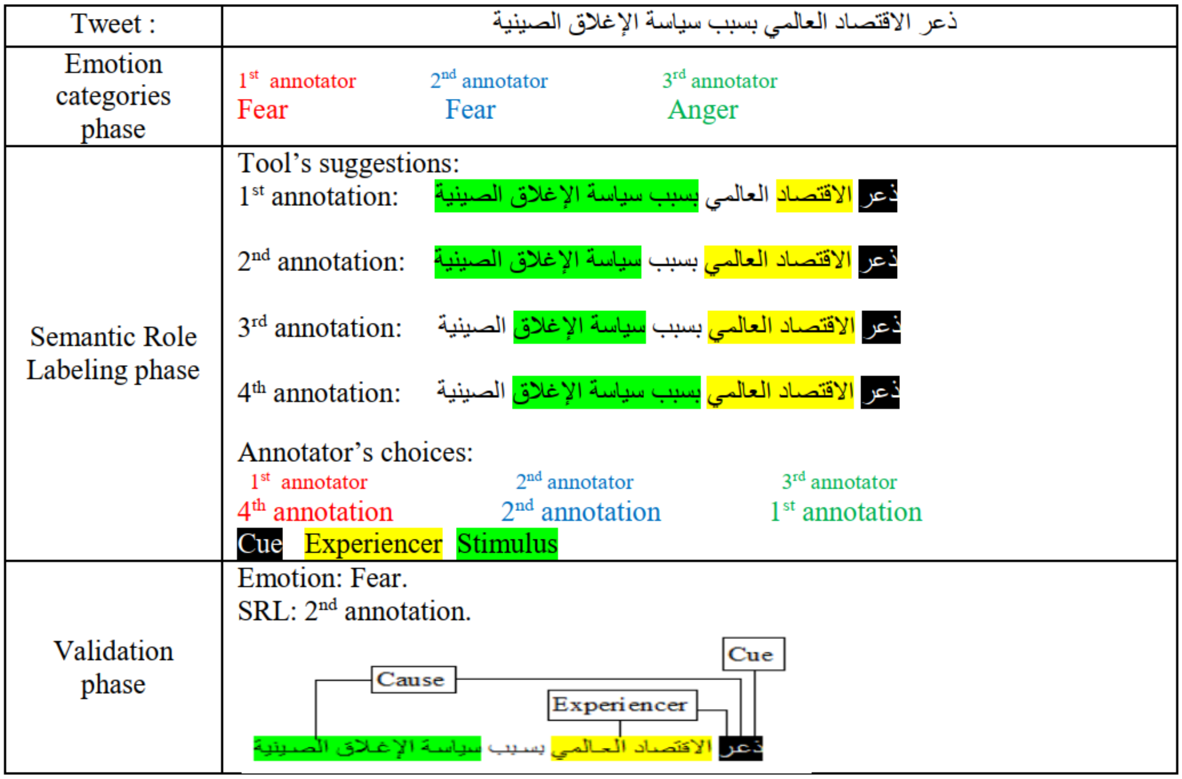


Figure 7.3: Example of an annotated tweet from our corpus.

annotation is validated; otherwise, the tweet is referred to the expert, as illustrated in Figure 7.3.

7.2.2.2 Corpus statistics

At the end of the annotation process, our dataset comprised 3,000 sentences, semi-automatically annotated. Table 7.1 presents general corpus statistics, including the distribution of emotion categories and the total number of annotations for each semantic role with respect to the dominant emotion.

Our results show that the highest number of annotations corresponds to the Joy class (32%), followed by Sadness, Anger, Neutral, and Fear. We also observe that the target word (cue) dominates the dataset. This is due to the use of query words (cues) as hashtags when querying the Twitter Search API, ensuring that at least one cue appears in each tweet.

In contrast to the results of GoodNewsEveryone [1], where the role of the target was among the most frequently annotated, our analysis shows that it had the fewest annotations (13%).

Table 7.1: Corpus Statistics for Tweets Frequency and Role Annotations in Each Emotion Category

Emotion Category	**Arabic Tweets Frequency**	**Roles**			
		Experiencer	***Cue***	***Cause***	***Target***
Joy	952	894	1062	798	403
Sadness	801	750	791	600	322
Anger	542	483	542	371	180
Fear	253	213	253	106	85
Neutral	462	420	524	284	276
All	3010	2760	3172	2159	1266

7.2.2.3 Data storage format

The data generated by our tool are stored in an XML file, organized into layers of description (Figure 7.4). Each layer contains annotated information corresponding to a specific level of annotation. The XML file produced by our tool can be utilized in various NLP tasks.


```xml
<text>الجزائر تُعرب عَن سعادتِها بتكريم الرئيس الفرنسي أسرة الشّهيد علي بومنجل</text>
<annotationSet cDate="14/09/2022" status="SM" ID="116">
  <layer rank="1" name="Target">
    <label cBy="A" end="24" start="17" name="Target"/>
  </layer>
  <layer rank="1" name="polarity">
    <label cBy="A" end="24" start="17" name="positive"/>
  </layer>
  <layer rank="1" name="spans">
    <label cBy="A" spID="1" end="72" start="26" name="Cause"/>
    <label cBy="A" spID="2" end="6" start="0" name="Experiencer"/>
    <label cBy="A" spID="3" end="24" start="17" name="Cue"/>
    <label cBy="A" spID="3" end="24" start="17" name="joy"/>
    <label cBy="A" spID="2" end="6" start="0" name="Character"/>
  </layer>
  <layer rank="1" name="relations">
    <label cBy="A" relID="1" sspID="3" tspID="1" name="Experiencer"/>
    <label cBy="A" relID="2" sspID="3" tspID="2" name="Cause"/>
  </layer>
  <layer rank="1" name="FE">
    <label cBy="A" feID="197" end="72" start="26" name="Stimulus"/>
    <label cBy="A" feID="194" end="6" start="0" name="Experiencer"/>
  </layer>
```


Figure 7.4: A screenshot of the annotation output

7.3 Conclusion

In this work, we presented the process of building an Arabic corpus annotated with semantic role labels and emotion categories to advance Arabic NLP tasks. We developed a tool to annotate a corpus of 3,000 Arabic tweets with emotions and semantic arguments, based on the theory of frame semantics. The corpus we provide facilitates future research on the recognition of emotions and their associated entities in text.

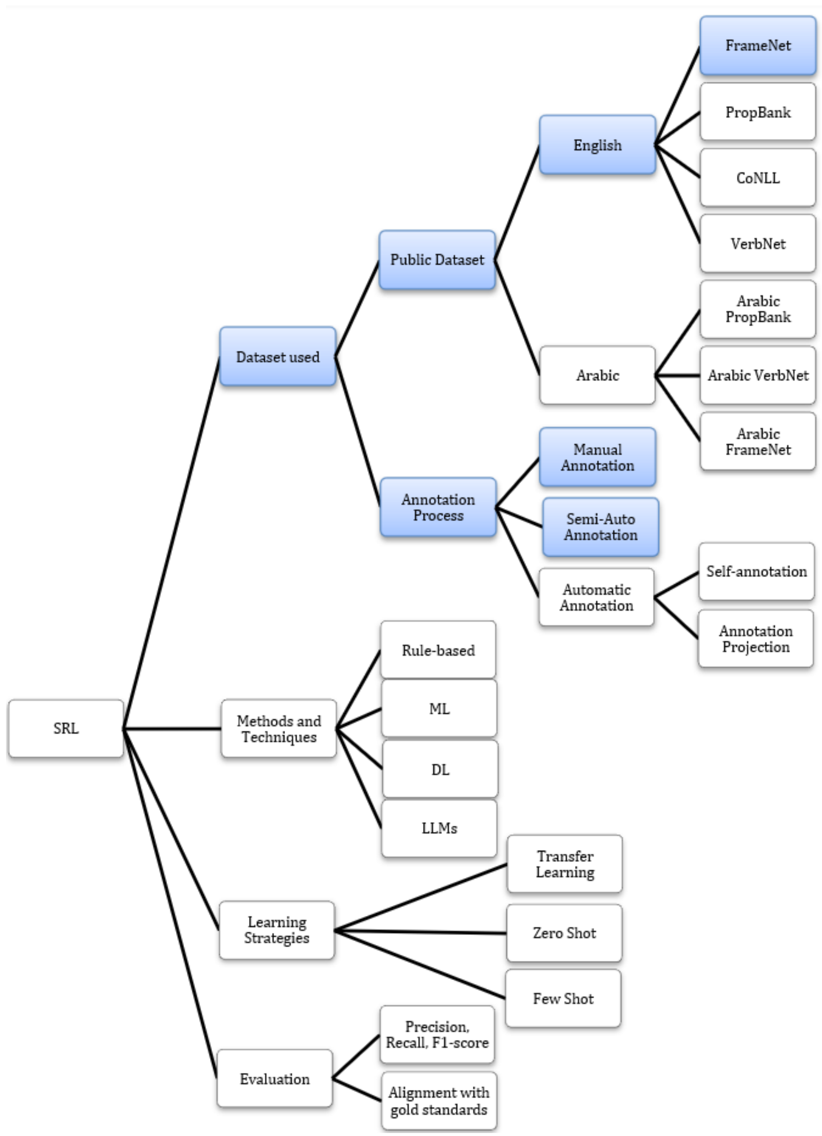


Figure 7.5: Scope of the first contribution (Arabic ERL corpus) within the SRL taxonomy

The diagram 7.5 represents the scope of the first contribution of our work within the broader context of Semantic Role Labeling, as outlined in the complete taxonomy shown in Figure 3.5.

While the full SRL taxonomy (Figure 3.5) includes multiple components (methods, learning strategies, and evaluation), the colored nodes in diagram 7.5 highlight the components addressed in the creation of the Arabic ERL corpus, focusing specifically on the "Dataset used" branch. More precisely, it details the annotation methodologies employed to construct our ERL dataset, which combines manual and semi-automatic annotation approaches. The underlying resource used for defining semantic roles was FrameNet.

The following chapter extends this corpus by applying several models.

CHAPTER 8

SRL AND EMOTION DETECTION USING TRANSFORMER BASED MODELS

8.1 Introduction

Several datasets for emotion classification have been provided across various domains and applications, including Reddit comments (Goemotion [90]), Self-reported narratives (International Survey on Emotion Antecedents and Reactions (ISEAR) [77]), tweets (Affect in Tweets [114]), YouTube video reactions (EmoReact [168]), and news (EmoNet [169], Affective text [170]).

When it comes to integrate emotions and SRL, the task revolves around the emotional cue (word or expression), rather on an action [42].

A representation of the relationship between emotion category and roles is shown in Figure 8.1.

The main research question is: How does the inclusion of semantic role information impact the accuracy and interpretability of emotion classification?. We aim to investigate the influence of semantic roles on emotion recognition in Arabic text. In first stage, we use an augmented version of our dataset [46] to improve the efficiency of DL based emotion detection models. This expansion facilitates the use of more advanced neural architectures, allowing the models to identify more intricate patterns and attain en-

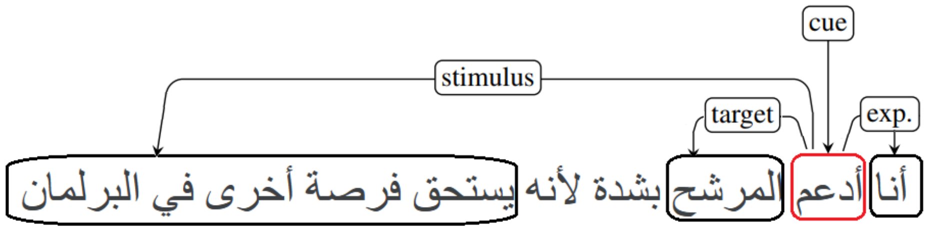


Figure 8.1: A sentence from our dataset represent a visual representation of the relationship between CUE, emotion category and roles.

hanced accuracy and generalization. in second stage, We systematically concealed each semantic role to evaluate the individual contribution of each emotional argument on the overall emotion detection efficacy. In the last stage, we applied the fine-tuned Arabic Bidirectional Encoder Representations from Transformers (AraBERT) pre-trained BERT-based model for our task, which has exhibited exceptional performance in diverse Natural Language Processing (NLP) tasks, such as text classification and named entity recognition.

The key contributions can be summarized as follows:

- Expanding our previous corpus by collecting new data by incorporating newly collected data,
- Masking the emotional arguments individually to study their impact on the performance of emotion identification,
- fine-tuning AraBERT for special tasks.

8.2 Methodology

8.2.1 Dataset

In NLP fields, data assume an essential and indispensable role. As there are no existing annotated Arabic corpora combined SRL with emotion identification, we have augmented our original dataset of 3,000 sentences [46] building an extended dataset named **Arabic Emotion Role Labeling (AraERL)**. We manually collected an additional

3,000 sentences from several sources, resulting in a total sample of 6,000 sentences. The sources for these new sentences included news headlines, Koran, academic books, Arabic novels, children's stories, and publicly available corpora, guaranteeing extensive coverage of numerous contexts and domains (see table 8.1).

To ensure consistency with our previous dataset, we used the same annotation tool and followed the same annotation scheme, which relied on a team of annotators and labeled sentences based on semantic roles (cue, target, experiencer, cause) and emotion categories (sadness, anger, fear, joy, disgust, love, surprise, and hate). We carefully reviewed each sentence, ensuring coherence across the expanded dataset. This expanded corpus provides a broader foundation for emotion detection tasks and enhances the diversity and generalizability of the model applied to the text.

Table 8.1: Corpus Statistics for Number of Sentences in Each Source

Source	Koran	Twit.	Fb	News	Acad. books	Novels	Stories	Corpora
NB sentences	150	2300	570	700	187	1028	931	214

The table 8.1 summarizes the number of sentences annotated with semantic roles and emotion labels in each sources represents a different domain or corpus utilized for evaluating the contribution of semantic roles to emotion detection tasks.

8.2.2 Data preprocessing

Data cleaning process was applied to eliminate unnecessary noise, redundant characters, and whitespace. This phase was essential for improving the quality of the data used in classification. In Arabic texts, certain orthographic variations can introduce noise, various normalization techniques were applied. Letters such as The letter "Alif" with different "hamza" positions, were consolidated into a singular representation for text standardization. Diacritic marks (tashkeel), frequently superfluous in analysis, were eliminated to facilitate tokenization. Furthermore, the "tatweel" character, employed for word elongation, was removed.

8.2.3 Masking roles

In the analysis of emotion classification from text, masking particular semantic roles in sentences is a crucial method for discerning the impact of each role (target, cue, experiencer, or cause) on the emotional significance expressed by the sentence. This method is methodically substituting the words or phrases associated with a specific semantic role with a placeholder, such as "[Target]" or "[Experiencer]," and subsequently assessing how the omission of this role influences the classification of the sentence's emotions.

	Sentences	Cue	Experiencer	Cause	Target	EmotionARA	Emotion
0	لكن زوجة أبيها عنفتها،[Cue] بقسوه	عنف وبخ	زوجة ابيها	NaN	ها	غضب	anger
1	[Cue] ت الملك و الملكة بشده	وبخ	ت	NaN	الملك الملكة	غضب	anger
2	فحدقت فيه بنظره [Cue]	حقد كراهية	ت	NaN	NaN	غضب	anger
3	[Cue]هم الطاهية	وبخ	الطاهية	NaN	هم	غضب	anger
4	زوجته لم تعطه فرصة للجدال، وأخذت [Cue] ه وتلومه	توبخ	زوجته	NaN	ها	غضب	anger

Figure 8.2: Dataset representation after masking roles.

This strategy aims to isolate the impact of specific roles on the emotion classification process. By concealing, for example, the "cue" of the emotion, the classifier can evaluate the significance of this role in determining the overall emotional category of the text. By systematically repeating this procedure for each semantic role, we can determine which roles have the most significant influence on emotion recognition. If the removal of a specific role leads to a significant decrease in classification accuracy, it indicates that the role is crucial for transmitting emotion. If the masking has minimal impact, it suggests that the function may not be crucial in shaping the sentence's emotional content. This method enhances the model by focusing on the most significant features. It offers insights into the contribution of various phrase components to emotional interpretation, so improving both theoretical comprehension and practical implementations in NLP tasks, including sentiment analysis and affective computing.

8.2.4 Dataset splitting

The primary aim of dataset splitting is to partition the given dataset into subsets that fulfill distinct roles in the model development process, table 8.7 show the statistic of splitting dataset. The data is generally divided into three parts; (70%) for training set, (15%) for validation set, and (15%) for testing set.

Table 8.2: Dataset distribution over the eight emotions

Emotions	**Training set**	**Validation set**	**Testing set**
Joy	990	223	219
Sadness	1032	219	259
Disgust	201	25	30
Hate	99	47	27
Fear	561	139	131
Anger	308	47	69
Surprise	220	47	54
Love	770	150	124
Overall	4197	897	897

8.2.5 AraBERT Pre-trained language model

In order to enhance comprehension of word meanings, the intricate syntactic and semantic information, a pre-trained language models for word representation was used. AraBERT is a pre-trained language model that supports Modern Standard Arabic (MSA) and several dialects, proposed by Antoun et al. (2020) [171]. The Bidirectional Encoder Representations from Transformers (BERT) architecture serves as the foundation for this multi-layer bidirectional transformer. AraBERT was fine-tuned with a substantial corpus of Arabic text (77 GB) derived from various sources. The model seeks to tackle the complexities associated with the Arabic language, including its complex morphology and diacritics. It is applicable for several NLP tasks, including sentiment analysis, text categorization, question answering, and named entity recognition.
We utilized the *BERT-base-arabertv02* model, which has been pre-trained to achieve optimal performance in the emotion classification task on Arabic text. We chose this model due to its pre-training on an extensive and varied Arabic corpus, as well as its

integration of language-specific tokenization methods.

8.2.6 Hyperparameter for AraBERT

The role of hyperparameters is to control how the model is trained. For learning rate, we set it to 2e-5. while, the batch size is set to 16. The per-device-eval-batch-size is set to 8, controlling how many samples are evaluated at a time. Finally, We train the model for 7 epochs, which means that the entire dataset is passed through the model seven times.

8.3 Results

8.3.1 Masking cue

Table 8.3 shows the results of masking the cue role, which indicates the explicit emotion words or phrases in a sentence.

Table 8.3: Experimental results of masking the cue role.

Emotions	**Precision**	**Recall**	**F1-score**
Joy	0.52	0.71	0.60
Sadness	0.20	0.23	0.11
Disgust	0.45	0.59	0.51
Hate	1.00	0.11	0.19
Fear	0.44	0.53	0.48
Anger	0.48	0.42	0.45
Surprise	0.14	0.29	0.20
Love	1.00	0.20	0.33
weighted avg	0.48	0.47	0.44

The highest F1-score (0.60) was provide by Joy category, followed by disgust and fear with 0.51 and 0.48 respectively. While sadness and surprise were poorly predicted, which suggests a complete failure to classify these emotions.
The elimination of this role substantially affected the model's performance reducing its weighted avg rate to 0.44, as the cue frequently functions as an indicator of the expressed emotion.

8.3.2 Masking experiencer

The table 8.4 summarizes the detailed performance across eight emotion classes, highlighting categories where the model performed well, as well as those where it encountered challenges.

Table 8.4: Experimental results of masking the experiencer role.

Emotions	**Precision**	**Recall**	**F1-score**
Joy	0.86	0.88	0.87
Sadness	0.14	0.23	0.40
Disgust	0.77	0.86	0.81
Hate	0.61	0.74	0.67
Fear	0.87	0.90	0.89
Anger	0.81	0.94	0.87
Surprise	0.33	0.07	0.11
Love	1.00	0.20	0.33
weighted avg	0.76	0.79	0.76

The results showed that the best performance was in fear followed by anger and joy category with 0.89, 0.87 and 0.87 F1-score respectively.
The weighted average across all categories for precision, recall, and F1-score is around 0.76. These results demonstrate that while the model performs well for certain emotions, it faces difficulty in predicting others such as sadness and surprise.

8.3.3 Masking cause

The table 8.5 presents the results of emotion classification when the cause role is masked, showing different metrics for each emotion class.

Table 8.5: Experimental results of masking the cause role.

Emotions	**Precision**	**Recall**	**F1-score**
Joy	0.91	0.88	0.90
Sadness	0.33	0.17	0.22
Disgust	0.80	0.89	0.84
Hate	0.79	0.79	0.79
Fear	0.81	0.90	0.85
Anger	0.85	0.94	0.89
Surprise	0.10	0.41	0.27
Love	1.00	0.80	0.89
weighted avg	0.77	0.82	0.79

Joy achieves strong performance across all metrics, with an F1-score of 0.90, suggesting that masking the cause role has little impact on this emotion's classification. while sadness shows poor results with an F1-score of 0.22.
The weighted averages for precision, recall, and F1-score are 0.77, 0.82, and 0.79, respectively. This means that the masking of the cause role's influence on the emotion classification is weak.

8.3.4 Masking target

The table 8.6 presents the experimental results of emotion classification when the target role is masked.

Table 8.6: Experimental results of masking the target role.

Emotions	**Precision**	**Recall**	**F1-score**
Joy	0.88	0.88	0.88
Sadness	0.17	0.17	0.17
Disgust	0.75	0.87	0.81
Hate	0.78	0.74	0.76
Fear	0.87	0.88	0.88
Anger	0.85	0.94	0.89
Surprise	0.33	0.07	0.11
Love	1.00	0.20	0.33
weighted avg	0.78	0.80	0.78

Anger, joy, and fear achieves highest performance, with an F1-score of 0.89, 0.88, and 0.88 respectively, showing small impact from the masking of the target role. The weighted average precision, recall, and F1-score of 0.78, 0.80, and 0.78, respectively. This suggests that masking the target role has a modest impact on different emotions.

8.3.5 Emotion Classification with all semantic roles

Table 8.7 illustrate the result of fine-tuning AraBERT model for emotion classification on the entire dataset, without masking any emotional arguments.

Table 8.7: Emotion classification result on the AraBERT model

Emotions	**Precision**	**Recall**	**F1-score**
Joy	0.86	0.92	0.89
Sadness	0.78	0.90	0.84
Disgust	1.00	0.92	0.96
Hate	0.50	0.57	0.53
Fear	0.97	0.82	0.89
Anger	0.83	0.77	0.80
Surprise	0.43	0.23	0.30
Love	0.93	0.93	0.93
Weighted avg	0.84	0.85	0.84

The evaluation metrics include Precision, Recall, and F1-score for different emotion categories. These metrics provide a comprehensive view of the model's performance in distinguishing between diverse emotions (Joy, Sadness, Disgust, Hate, Fear, Anger, Surprise, and Love). the model achieve high performance for both Disgust and Love (0.96, 0.93 respectively). Fear and Joy are also well-classified with 0.89. The worst performance is observed for Hate and Surprise, indicating difficulty in distinguishing these emotions.

8.4 Discussion

The set of experiments on masking semantic roles in emotion categorization underscores the substantial impact of each role in identifying emotions in Arabic texts is shown in table 8.8, w/o "Role" represent the test set results rate without emotional arguments, and w/ All roles represents the test set accuracy rate containing all emotional arguments.

We notice that certain roles exhibit a more significant impact than others. The model demonstrated substantial performance declines across all categories when masked cue role (which indicates the existence of emotion in a sentence), particularly in those that depend significantly on explicit cues to indicate emotional content. The masking of this role resulted in significant reductions in accuracy reach 0.47 , indicating that cues are crucial in anchoring emotion prediction.

The masking of the experiencer and target had varied effects on the classification performance (0.79 and 0.80 reaspectively) , with emotions such as fear and anger demonstrating more robustness, while emotions like love and sadness shown diminished performance. This indicates that certain emotions are closely linked to the presence of an experiencer, whereas others can be deduced from the context without direct information about the experiencer. Finally we find that the most modest impact is shown in masking cause (0.80).

Table 8.8: Overall experimental results.

Roles	**Precision**	**Recall**	**F1-score**	**Accuracy**
w/o Cue	0.48	0.47	0.44	0.47
w/o Experiencer	0.76	0.79	0.76	0.79
w/o Cause	0.77	0.82	0.79	0.82
w/o Target	0.78	0.80	0.78	0.80
w/ All roles	**0.84**	**0.85**	**0.84**	**0.85**

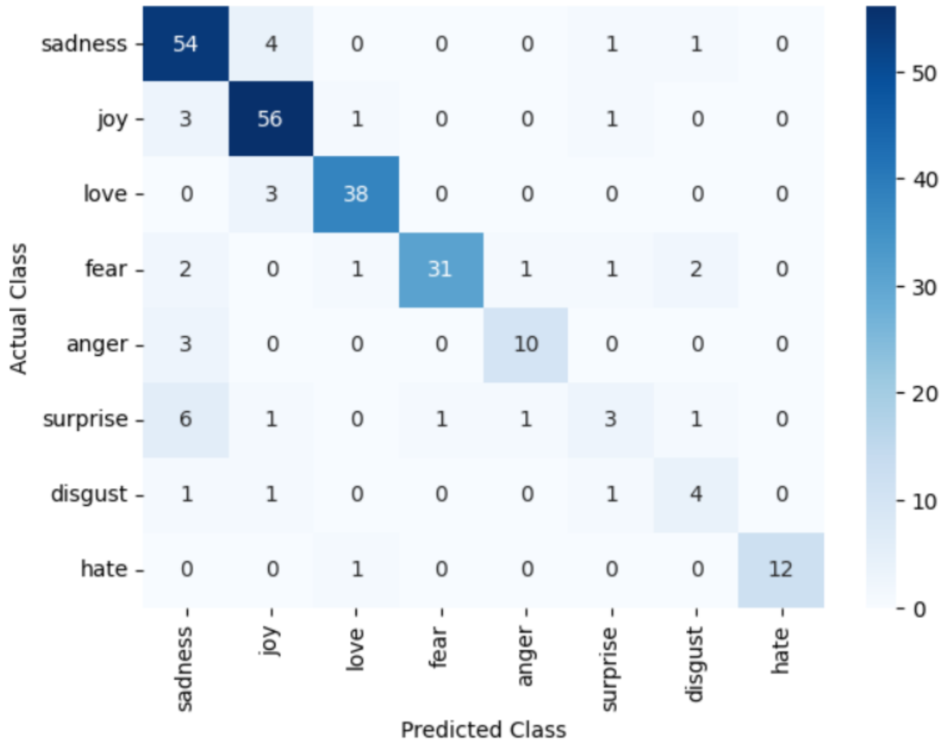


Figure 8.3: confusion matrix of emotion classification across different categories.

8.5 Comparison of Our Dataset with Existing Datasets

The development of our dataset was motivated by the limitations observed in existing SRL and emotion classification datasets, particularly for Arabic texts. While several benchmark datasets have contributed significantly to advancing SRL and emotion-related tasks, they are either language-specific (mostly English), lack emotion annotations, or do not provide fine-grained semantic roles linked to emotions.

Our dataset addresses this gap by offering a **manually annotated corpus of Arabic sentences**, in which each sentence is labeled with emotion-specific roles such as experiencer, cause, target, and cue. Furthermore, each instance is associated with one of the basic emotion categories, inspired by Ekman's models.

Table 8.9: Comparison of Datasets with Emotion Role Annotations

Dataset	Size	Source	Language	Emotions	Cue	Exp.	Cause	Target
GNE [1]	5000	News headline	English	8	✓	✓	✓	✓
REMAN [104]	1720	Fictional text	English	9	✓	✓	✓	✓
GERSTI [105]	2006	News headline	German	10	✓	✓	✓	
NTCIR(EN) [172]	1826	Novels	English	6	✓		✓	
NTCIR(ZH) [172]	2022	News	Chinese	6	✓		✓	
Elections [43]	1385	Tweets	English	8	✓	✓	✓	✓
AraERL [46]	**6000**	**Tweets, news headlines, Koran, academic books, Arabic novels, children's stories, corpora**	**Arabic**	**6**	✓	✓	✓	✓

8.6 Conclusion

This contribution highlight the intricate relationships between semantic roles and emotion detection efficiency. The study, examine the influence of masking semantic roles on the performance of emotion detection and classification, using AraBERT model. To this end, we performed our BERT based model on an existing dataset. A selected of 6000 emotional expressions were manually annotated with semantic roles and emotion categories. The experimental results indicated that various roles contribute differently to the model's ability to identify distinct emotions. Cues and experiencer roles provide clear indications and explanations for emotions, were recognized as crucial for accurate classification. Roles like target and cause, while significant, showed a more moderate effect.

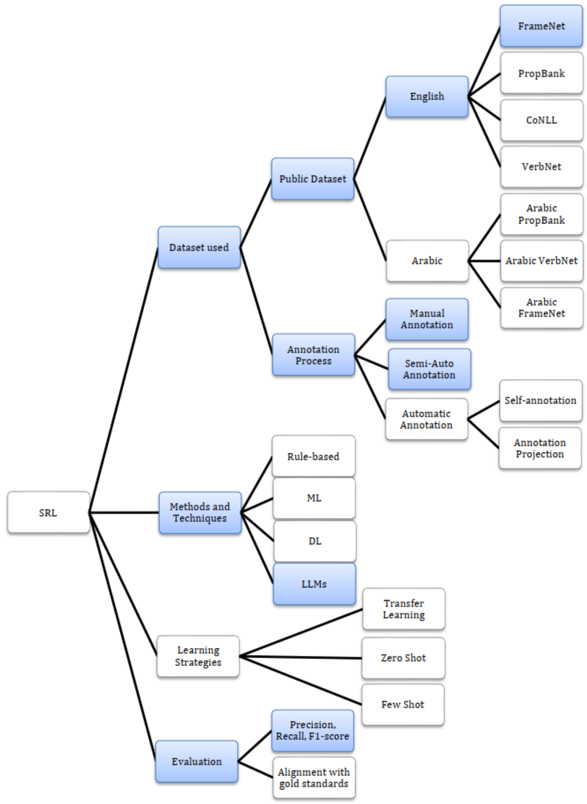


Figure 8.4: Scope of the second contribution within the SRL taxonomy

The diagram 8.4 represents the scope of the second contribution of our work in the broader context of Semantic Role Labeling, as outlined in the complete taxonomy shown in Figure 3.5.
The highlighted (colored) components represent the elements addressed in the current contribution. Having addressed dataset development and annotation strategies in the first contribution, the current contribution extends the analysis to provide more detailed coverage of SRL tasks, including modeling techniques and evaluation methods and the use of LLM.

Future studies can build on these findings to enhance emotion detection and classification task and further enhance the accuracy and reliability of emotion recognition, in Arabic language.

CHAPTER 9

LEVERAGING CHATGPT FOR ENHANCING ARABIC NLP: APPLICATION FOR SRL AND CROSS-LINGUAL ANNOTATION PROJECTION

9.1 Introduction

Approaches based on cross-lingual projection techniques, could be a potential solution for low-resource languages, to create initial annotated datasets. Cross-Lingual Annotation Projection (CLAP) leverages the existing well-annotated resource in the source language (English) to overcome the rarity of resources in other languages based on the translational and structural equivalences present in the aligned data [50]. This technique can help reduce the effort in terms of time, and both material resources and cost required to generate annotations or models for a new target language [45].

Recently, Large Language Models (LLM), such as GPT-3/4 [173], LaMDA [174], Google Bard, and PaLM [175], [176], have demonstrated high performance in diverse NLP tasks. The growing use of internet-based services such as cloud computing, social networks, and forums have facilitated the creation and training of LLMs, due to their broad accessibility and huge amount of data used for self-learning. These models can produce human-like language text with a high level of expertise, precision, and consistency without the need for fine-tuning [177].

We examine the use of Chat Generative Pre-trained Transformer (ChatGPT) as a tool for performing two main sub-contributions for Arabic language processing (1) creating an Arabic annotated resource with emotional semantic roles from an English corpus, using cross-lingual projection approach, and (2) annotating the Arabic corpus of emotional text with emotion categories and semantic roles. The reference English corpus GoodNewsEveryone (GNE) [1]. Furthermore, we evaluate the potential of ChatGPT in translating English sentences to Arabic.

In summary, we assess the capabilities of using ChatGPT in challenging human-centered NLP tasks, requiring understanding and competences, such as translation and annotation, by following six steps evaluation pipeline. Finally, we evaluate the impact of sentence complexity on the performance of ChatGPT in Semantic Role Labeling (SRL) , cross-lingual annotation projection, and zero-shot annotation accuracy obtained by ChatGPT compared to human annotators.
From a perspective of generalization, our study was extended to explore open-LLMs models for automatic SRL, and CLAP. To elucidate our objectives, we investigate the following related questions:

- To what extent ChatGPT can serve as a viable alternative to human expert annotators in performing SRL task?
- To what extent ChatGPT can accurately map argument and SRL labels from a source language to a target language?
- To what extent open-LLMs can perform automatic SRL, and CLAP?

Ultimately, these questions have been addressed through a series of evaluations and assessments, based on a benchmark of manual annotations. ChatGPT have been evaluated through three main steps, the first step consist of two tasks, where the first task consists of SRL of translated sentences from English to Arabic. Whereas, the second task involves projecting semantic role labels from an English corpus (the source language) into Arabic (target language). Subsequently, the performance of the projected SRL labels is assessed in comparison with manual annotation. In the following phase, a comparison was made between the translation of ChatGPT and that of human annotators and translators. Similarly, from the perspective of generalization, we evaluate the performance of open-LLMs in SRL and CLAP tasks, using Multilingual BERT (mBERT) and Multilingual Bidirectional and Auto-Regressive Transformers (mBART), respectively. The contributions can be summarized as follows:

- The assessment of ChatGPT's performance in several tasks, namely, emotion category identification, intensity prediction, and SRL with cue, experiencer, cause, and target arguments.
- The investigation of ChatGPT's capability in projecting semantic role labels from English datasets to Arabic.
- The construction of two benchmark Arabic data sets, one consisting of human-translated and annotated versions, while the other is generated using ChatGPT.
- The investigation of Open-LLMs capability, in automatic SRL and CLAP tasks.

9.2 Methodology

In this contribution we provide a comparative analysis pipeline to evaluate the effectiveness of ChatGPT, a proprietary LLM (Figure 9.1), in annotating Arabic sentences. Our objective is to assess ChatGPT's proficiency in SRL and CL annotation projection tasks, with the goal of leveraging its capabilities to enhance NLP tasks.
As shown in Figure 9.1.(A), the study is divided into six steps, the first step represented in data collection and preparation, where 500 sentences were selected from the GNE database. In the second step, a proficient specialists translated sentences from English to Arabic. In the subsequent steps, we conducted two separate experiments. In the first experiment, we provided ChatGPT the translated sentences along with a questionnaire consisting of six questions in order to semantically annotate them. The questions were presented in a single prompt to be collectively analyzed by ChatGPT. While, in the

second experiment, ChatGPT was used to translate an English annotated dataset of news headlines GNE, into Arabic while adhering to the SRL requirements. By using a single prompt that interrogate ChatGPT in a manner that adhere to SRL requirements. It is important to note that we have employed the 'gpt-3.5-turbo' version of the ChatGPT model. We followed zero-shot classifications to submit the headlines to ChatGPT over the time-frame of February 20th to March 3rd, 2024. Where, we established a separate chat session for each headline to ensure that the model's outputs are not affected by previous annotation histories. It is important to note also that any response generated by ChatGPT were neither adjusted nor excluded (except specific cases stated in section 9.4.2), preserving their original state to better evaluate ChatGPT's efficiency in automatic annotation. The outputs were validated by comparing them to human annotations. For validation, a team of annotators manually annotated the translated phrases from the first phase to assess each pipeline, as outlined in the fifth phase. The manual annotations were subsequently compared to those produced by ChatGPT in SRL and CLAP phases.

Furthermore, translated sentences are also compared with the sentences translated by ChatGPT in CL annotation projection. Finally, the last step aims to further evaluate the performance of ChatGPT in handling language complexities. Based on language rules and experts, the headlines we categorized into three complexity levels: easy, medium, and difficult, for both Arabic and English. Afterwards, a statistical analysis is conducted for each category to assess the CL annotation projection-SRL projection between Arabic and English.

In contrast, as shown in 9.1.(B), the study implements open-LLMs, namely mBERT, and mBART, in order to further explore the use of different open-LLMs models in automatic SRL, and CLAP.

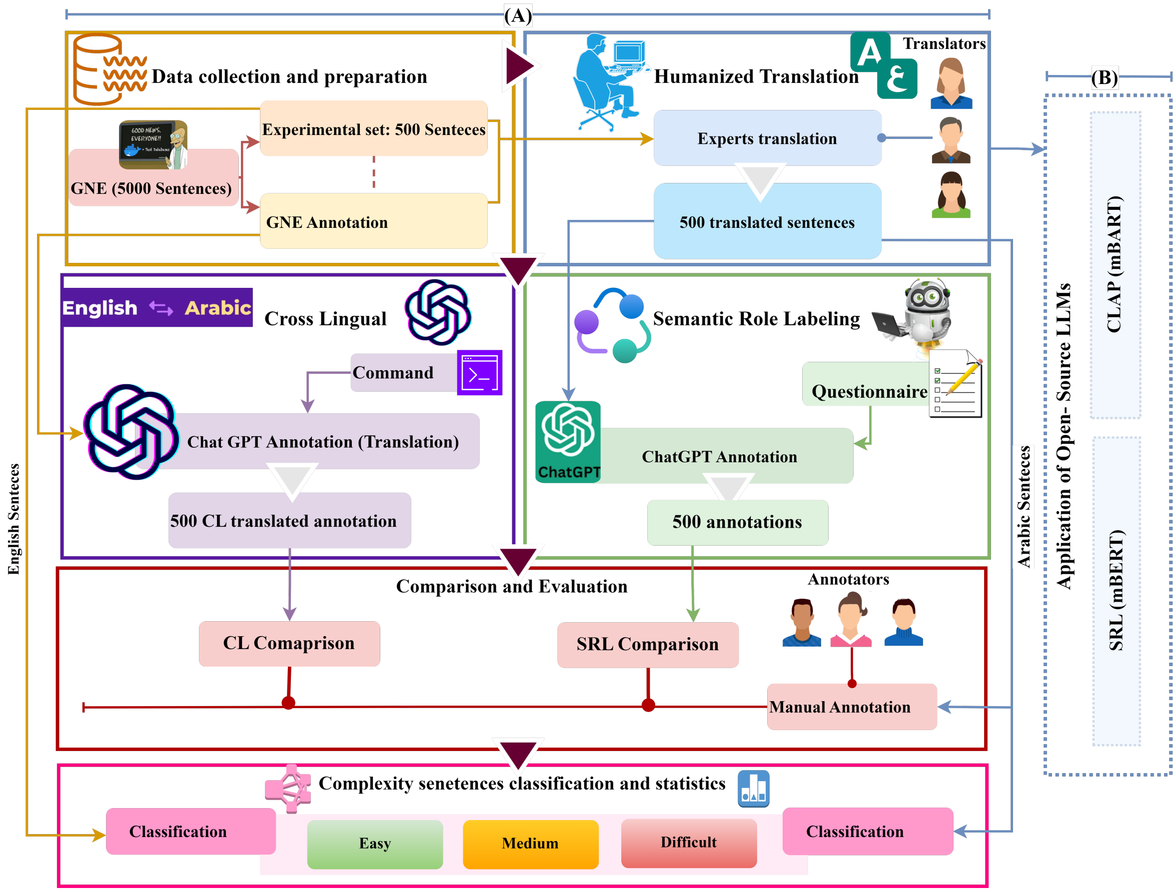


Figure 9.1: A. Illustrative workflow of the proposed methodology for evaluating ChatGPT in Arabic annotation; B. Application of Open-LLMs in SRL and CLAP.

9.2.1 Data collection and preparation

In our research, we utilized the GoodNewsEveryone (GNE) dataset [1] as a case study. The dataset consists of 5,000 English headlines collected from 82 different sources. Each headline has been annotated with intensity, emotion, cues, experiencers, causes, and targets (see Table 9.1).

This dataset has been chosen due to its significance, as it encompasses the essential emotional arguments and takes into account news headlines characterized by conciseness and informative words.

From the original dataset, a subset of 500 headlines, are randomly selected along with their English annotation from the entire dataset, as a sample to evaluate the output of ChatGPT (Figure 9.1).

Table 9.1: GNE Dataset description

General Information	Year	2020
	Citations	60
Statistics	Sentences	5000
	Words	56612
	Characters	354173
Characteristics	Experiencer	3466
	Cue	4814
	Cause	4798
	Target	4477

9.2.2 Humanized translation

The process of humanized translation mechanism by a native speaker goes beyond a simple word-for-word exchange across languages. It is basically based on a deep comprehension and safeguarding of the cultural nuances, context, and intent conveyed by the original text. The difficulty of Arabic sentences stated in chapter 2

Hence, initially, a team of translators attempts to understand the content and then produces an initial translation. Afterwards, these translations are refined and revised, based on cultural and linguistic information and rules to accurately capture the finest details in the sentence context. Finally, an expert reviews the translator's output and selects the translations that ensure accuracy and fluency. This process ensures that the translated content maintains the essence of the original text.

9.2.3 Emotional Semantic role labeling

9.2.3.1 Dominant emotion and intensity

To start the process of annotating Arabic headlines, we initiate a chat session for each sentence that is submitted by asking a series of leading questions. In which, the first question is, "Which feeling is most prominent in this sentence?". The answer is based on a predetermined set of Plutchik's emotion classes: anger, fear, disgust, joy, love, irritation, pessimism, negative surprise, optimism, positive surprise, guilt, pride, sadness, shame, and trust. This set refers to the same set of emotions employed for annotating

GNE.

The second question that follows is, ”How intense is the emotion?”. This question allows for one of three possible answers: high, medium, or low. A detailed explanation of the asked questions is illustrated in Table 9.2.

9.2.3.2 ChatGPT semantic role labeling pipeline

The main task in the third step (Figure 9.1) is the Arabic annotation of sentences using ChatGPT, while preserving the SRL. This task is achieved by following a series of structured steps, guided by a set of questions to ensure that the chat process will not diverge into another context. The overall process is summarized in algorithm below.

```
Require: Dataset of input sentences
  Define a list of questions (see Table 9.2)
  for each sentence in dataset do
    Initialize new chat session
    responses ← {}
    for each question in questions do
      Send question with sentence to ChatGPT
      Receive response
      if response is valid then
        Store response in responses
      else
        Adjust the query to obtain a correct answer.
      end if
    end for
    Save responses as JSON
  end for
```

In order to facilitate the annotation of Arabic headlines with ChatGPT, we begin by initializing a chat session for each input sentence. Subsequently, we display the questions sequentially within the same chat session. Continuously, we iterate the question-answering procedure for each sentence in our data. Table 9.2 includes the various questions we posed to ChatGPT.

Table 9.2: Questionnaires for the emotion analysis and SRL.

Questions	Parameter
Which emotion predominates in this sentence? (in one word)	Emotion categorie
How intensely is the emotion? (high, medium, low)	Intensity (Low, Med, High)
Which words identify the emotion? (in one word)	Cue
Who is the experiencer of the emotion?	Experiencer
Select the words that caused the expressed emotion.	Cause
Who or what is the emotion directed at?	Target

A number of tests were carried out before the method of querying ChatGPT was finally fixed as follows:

- **Test 1 :** We provide only the sentence and questions, without any further components. As illustrated in Figure 9.2, we note that the sentence is in Arabic, while the response is in English.

- **Test 2:** Prior to the questions, we include a remark to specify that ChatGPT must provide the answers in Arabic. In cases when there is no response, this leads to the generation of random roles that are not even mentioned in the sentence, as illustrated in Figure 9.3.

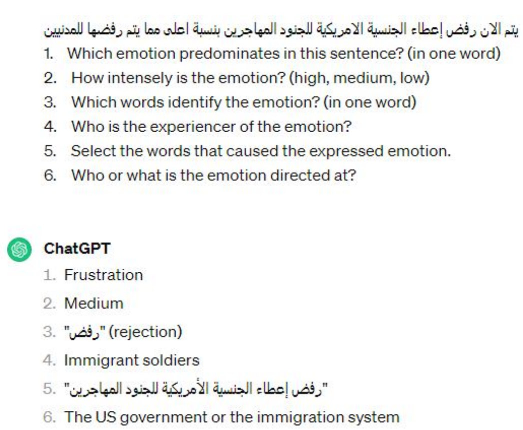


Figure 9.2: Test 1: the sentence in Arabic but the response in English

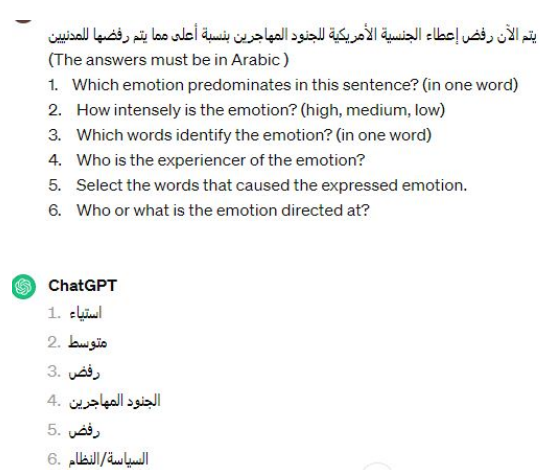


Figure 9.3: Test 2: generation of random response.

Ultimately, we instructed ChatGPT to respond as follow: *"The answers must be in Arabic and taken from the sentence. If there is no answer, do not formulate it"*. Following every ChatGPT generated output, we extract the annotation and save it in a JSON file. Figure 9.4 displays an instance of SRL performed by ChatGPT.

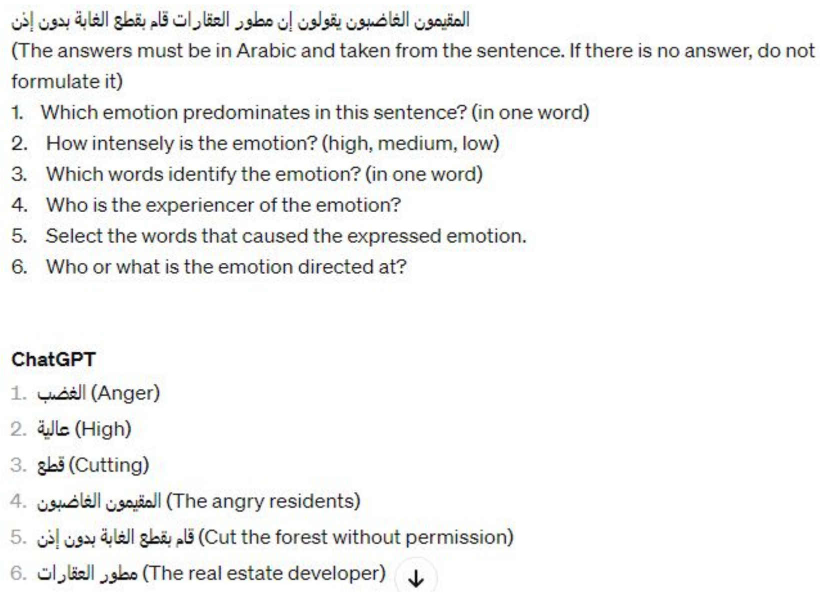


Figure 9.4: Example of an SRL generated by ChatGPT.

9.2.4 Cross-lingual Annotation Projection

9.2.4.1 Translation mechanism of ChatGPT

ChatGPT's translation aims to preserve the semantic meaning of the text, taking into consideration the input sequence. The translation process is shown in Figure 9.5. First, the English sentence is tokenized into words, and then embeddings are used to translate each token into a numerical representation. ChatGPT makes use of self-attention processes to understand the relationships between several input linguistic elements. Then, it generates the Arabic tokens one by one until the translated sentence is completed. The translated sentence can be subjected to post-processing to improve the quality and format of the sentence [178].

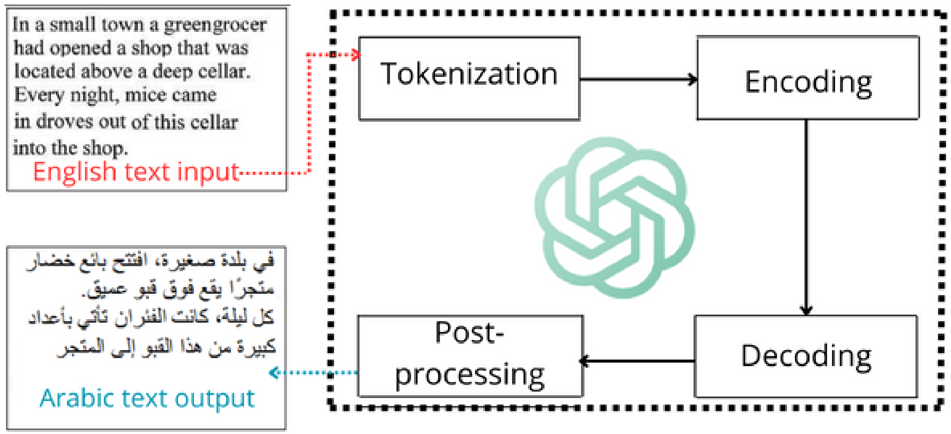


Figure 9.5: The translation mechanism of ChatGPT.

9.2.4.2 ChatGPT's cross-lingual annotation projection pipeline

Given the lack of Arabic resources, we utilize the ChatGPT's computational cross-lingual features to create an Arabic corpus from the GNE corpus. This objective is accomplished by utilizing ChatGPT's linguistic expertise to ensure precise semantic role alignment with the semantic context of Arabic (see Figure 9.7). The result is the creation of a new Arabic corpus that possesses identical characteristics and annotations as the GNE corpus, all achieved in an easy, efficient, and cost-effective manner. To achieve this milestone we adhered to the steps illustrated in Figure 9.6

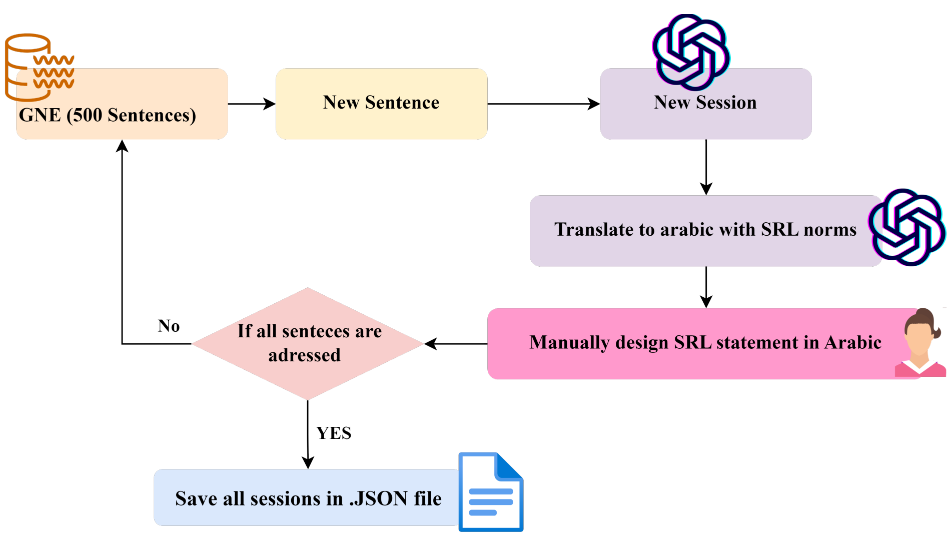


Figure 9.6: ChatGPT's cross-lingual projection flowchart

As illustrated in Figure 9.6 for each headline in our dataset, we initiated a new chat session with ChatGPT. Our goal was to translate English statements from the GNE corpus into Arabic while maintaining the integrity of their semantic roles. To achieve this, we provided ChatGPT with a specific command: ”The following sentence is with its SRL; translate it to Arabic by preserving its semantic meaning and by respecting the Arabic rules of SRL.”

After receiving the translated sentences, we took additional steps to ensure accuracy. We manually aligned the semantic roles projected onto the Arabic translations with the corresponding Arabic statements. This manual alignment is a crucial step, as it guarantees that the semantic roles in the translated text accurately reflect the intended meanings from the original English statements.

At the end of the process, we store the generated Arabic annotations in a JSON file. This structured storage allows for easy access and future use of the translated and annotated data.

the following sentence is with its semantic role labeling, translate it to Arabic by preserving it's semantic meaning and by respecting the Arabic rule of semantic role labeling (keep the same JSON forma please).

{"headline": "David Beckham gets six-month driving ban for using phone at wheel", "annotations": {"dominant_emotion": "negative_surprise", "other_emotions": "reader_emotions": "intensity": "medium", "readerperception": true, "cause": [["gets six-month driving ban for using phone at wheel"]], "cue": [["ban"]], "experiencer": [["david beckham"]], "target": [["driving ban"]]}}

ChatGPT

{"headline": "دافيد بيكهام يحصل على حظر القيادة لمدة ستة أشهر بسبب استخدام الهاتف أثناء القيادة", "annotations": {"dominant_emotion": "negative_surprise", "other_emotions": "reader_emotions": "intensity": "medium", "readerperception": true, "cause": [["يحصل على حظر القيادة لمدة ستة أشهر بسبب استخدام الهاتف أثناء القيادة"]], "cue": [["حظر"]], "experiencer": [["دافيد بيكهام"]], "target": [["حظر القيادة"]]}}

Figure 9.7: Example of CL annotation projection generated by ChatGPT

9.2.5 Annotation tasks and comparison

As illustrated in Figure 9.1, three annotators independently annotate the Arabic sentences consecutively, indicating the dominating emotion, intensity, cue, experiencer, cause, and target. These annotators had a prior training. Whereas, we constructed a series of organized questionnaires (Table 9.2), where each participant is obliged to answer all questions in order, resulting in the creation of three separate and unrelated annotations.

The validation method for each headline item involves the consensus of three annotators. If all three annotators agree on a single annotation, the expert approves and validates the headline. If there is a situation where two annotations align while the third one differs, the expert will confirm the category that the majority agrees on. However, when the annotators can not reach a consensus, the expert decides which annotation to be retained. Upon completion of the process, we obtain a gold-standard dataset that has been manually annotated. This dataset is now prepared for use as a benchmark to assess the accuracy of the SRL and CLAP findings of ChatGPT. We gathered the outcomes provided by ChatGPT in both experiment, and we contrasted them with the annotations conducted by human annotators.

9.2.6 Classification of sentence complexity

During this phase, we categorize the sentences based on their level of complexity, ranging from easy to medium to difficult. The purpose of this approach is to evaluate ChatGPT's ability to comprehend the characteristics of sentences, and to see to what extent does the compelxity of statements impact the performance of ChatGPT's annotation. To this end, the classification was established based on two fundamental principles for both languages.

9.2.6.1 English sentences complexity

For the classification of English sentences, we utilized two reading formulas: the Flesch Kincaid method and the Gunning Fog Index formula.

Flesch Kincaid: Rudolph Flesch created the formula for measuring text readability, taking into account sentence and word counts. The given score is between 0 and 100; [90–100] shows that the text is easier to read; [60–70] average difficulty; and [0–30] indicates that the sentence is difficult to read [**32**]. The mathematical formula used is as follows:

$$0.39(\frac{totalwords}{totalsentences}) + 11.8(\frac{totalsyllables}{totalwords}) - 15.59 \tag{9.1}$$

Gunning Fog Index Formula: Set up by Robert Gunning, the Gunning Fog Index readability score focuses its calculations on complex words (three or more syllables), the grade level produced by this formula between 0 and 20 [**32**]. The formula is as follows:

$$0.4(\frac{totalwords}{totalsentences}) + 100(\frac{complexwords}{totalwords}) \tag{9.2}$$

We selected these two readability formulas based on distinct criteria: the first method relies on sentence length, while the second takes into account the usage of complex terms.

9.2.6.2 Arabic sentences complexity

Due to the intricacy of the Arabic language, and the lack of standardized norms and metrics for measuring sentence difficulty, we enlisted the help of a human expert to create a three-class categorization system for the set of sentences.

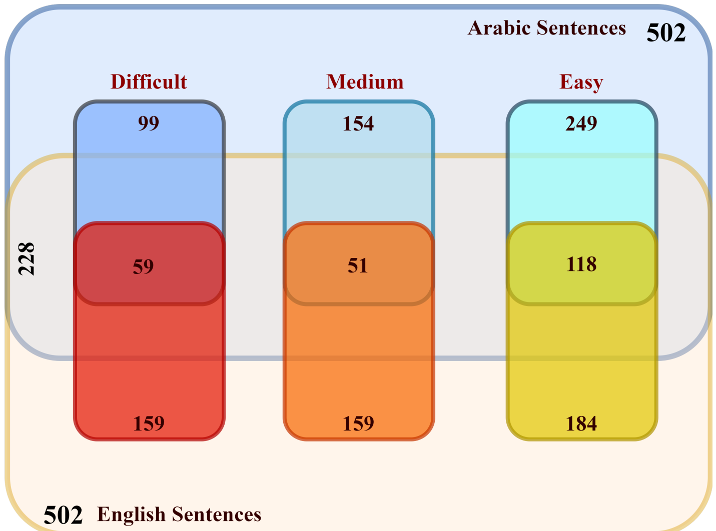


Figure 9.8: Comparison between Arabic and English complexity

The Venn diagram (see Figure 9.8) depicts the connections between Arabic and English sentence complexity. The two large sets consist of Arabic and English sentences, each comprising 502 instances. Each instance is then divided into three subsets: "difficult, medium, and easy." The intersection between them corresponds to the number of sentences that exhibit equivalent complexity in both Arabic and English. Within the set of sentences, 59 were considered very tough, 51 sentences were fairly comprehensible, and 118 sentences were considered easy.

9.2.7 Open-LLMs in SRL and Cross-lingual annotation projection

At this level, we expand our contribution to another front, in which we investigate the versatility of our proposal to Open-LLMs based models, and to explore their effectiveness in annotation tasks. As demonstrated in recent researches, LLMs have the ability to understand the context, allowing them to interpret and annotate text more accurately based on the surrounding content. Also, LLMs models can be fine-tuned on specific datasets or tasks, allowing for customized annotation that meets particular requirements. In this context, we apply the same methodology used so far (in sections 9.2.3, and 9.2.4) based on some well known open-LLM, so that the obtained result will be further compared to ChatGPT performance.

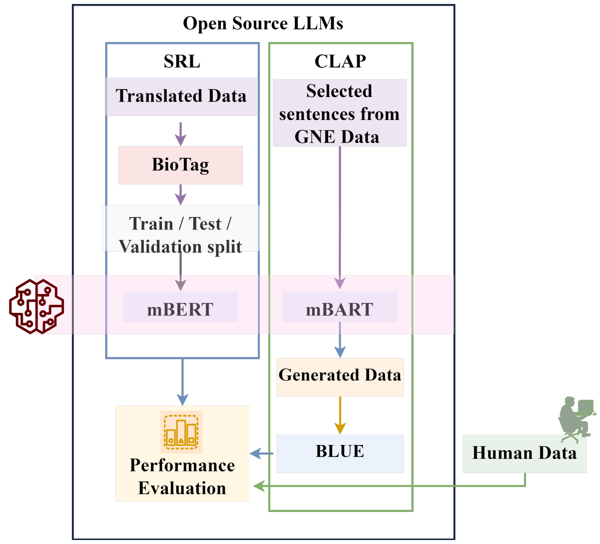


Figure 9.9: Workflow of the used Open-LLMs in SRL and CLAP tasks

As chosen LLMs, two state-of-the-art models have been selected to replace the ChatGPT task, namely mBERT, and mBART for SRL and cross-lingual annotation projection respectively:

- **mBERT** (Multilingual Bidirectional Encoder Representations from Transformers): A multilingual version of BERT pre-trained on 104 languages, encompassing Arabic. mBERT is extensively utilized for multilingual NLP tasks due to its ability to successfully capture cross-lingual representations [179]. In addition mBERT is conceptualise based on the Transformer architecture, that uses attention mechanisms to generate deep, contextualized representations for every word, which is a very relevant skill in performing SRL tasks.

- **mBART** (Multilingual Bidirectional and Auto-Regressive Transformers): It is a sequence-to-sequence model created by Facebook Artificial Intelligence (AI) for multilingual natural language generation tasks, it's an extension of the BART model [180]. The particularity of this model is it's wide and diverse languages training approach, also it can capture both forward and backward dependencies, making it powerful for translation and text generation tasks. All this characteristics qualified mBART to be adopted by state-of the-art works in cross- lingual tasks [181], [182].

9.2.7.1 mBERT for SRL and emotion analysis

As shown in figure 9.1, we applied a three steps model to evaluate the performance of mBERT in SRL and emotion analysis. Where, we have converted the dataset to CoNLL format, with each token presented on a separate line alongside its Begin-Inside-Outside Tagging Scheme (BIO) as illustrated in Figure 9.10. Furthermore, the sentences are delineated by a blank line. The dataset used in emotion analysis is in csv format, each sentence has its corresponding emotion category in the corresponding column. The experiments are conducted by fine-tuning mBERT model for token classification tasks (SRL).

```
الديمقراطيون B-Experiencer
يدافعون B-Cue
عن O
النائبات O
اللواتي O
يقول O
ترامب B-Target
أنهن O
يجب O
ان O
يرجعن B-Cause
إلى I-Cause
بلدانهن I-Cause
الفاسدة I-Cause
```

Figure 9.10: Example of Arabic SRL from our dataset in BIO format

To train the mBERT on SRL task, we processed the dataset of sentences with BIO semantic role labels, by creating a lists of words and their corresponding labels. The Arabic text was tokenized with BertTokenizerFast, and maps the labels to numerical IDs. The dataset was partitioned into training (80%), validation (10%), and test (10%) sets.

The mBERT model (BertForTokenClassification) is initialized with the number of labels needed for SRL. The hyperparameters used in this experiment have been tuned through a set of executions where it stabalised to the following setting : a learning rate of 2e-5, a batch size of 8 per device, and 3 epochs of training in addition to the Adam optimizer.

9.2.7.2 mBART for Cross-lingual Annotation Projection

In the same context, we applied a three steps experiment in the CLAP task, where we translated fields in an English JSON dataset to Arabic utilizing mBART and stores the translated annotation in a new JSON file. The mBART initializes the model and tokenizer for multilingual translation, sets English as the source language, and Arabic as the target. Each entry's "headline" and relevant fields within "annotations" (such as "dominant emotion," "cue," "experiencer," etc.) are translated to Arabic. These entries are skipped in case the JSON structure is invalid or indexing issues arise. Afterwards, we use Bilingual Evaluation Understudy (BLEU) score to compare mBART result's with expert translation.

9.3 Results and evaluation

9.3.1 Comparison of ChatGPT translation with Expert translation

In order to evaluate the translation output of ChatGPT, we measure the similarity between the results obtained by ChatGPT and those produced by human translators and expert. The examples bellow, presents a samples of selected GNE sentences in English form, and sheds light on the varieties and similarities in translations given by human and those provided by ChatGPT.

Example : 01

Sentence :

BA and Lufthansa suspend Cairo flights citing security.

Expert :

الخطوط الجوية البريطانية ولوفتهانزا تعلقان رحلاتهما إلى القاهرة لأسباب أمنية

ChatGPT :

اي . إيه ولوفتهانزا قررتا تعليق رحلاتهما إلى القاهرة نظراً لأسباب أمنية.

Example 02

Sentence :

Stephen Colbert Hits Long-Shot Dems for Spewing "Republican Talking Points" at CNN Debate.

Expert :

ستيفن كولبيرت ينتقد المرشحين الديمقراطيين البعيدين عن الفوز لاستخدام نقاط تحدث عنها الجمهوريون خلال مناظرة على شبكة سي إن إن.

ChatGPT :

ستيفن كولبيرت يهاجم الديمقراطيين الذين طال انتظارهم بسبب إلقاء نقاط الحديث الجمهوري في مناظرة سي ان ان

In the first sentence, both translations presents the same meaning, each of them chose a different way to express the translation. Consequently, we address both translations as accurate. Whereas, the second sentence presents variations in the selection of words as well as the writing style. Where, the expert translation tends to be more formal and precise, and the used expression is maintained. While the translation offered by ChatGPT is more straightforward and concise, yet, it may not completely deliver the complex meanings of the original Arabic language.

To verify ChatGPT's performance in English to Arabic translation, we performed used standard to perform a thorough analysis and evaluation. Where, we presented an overall statistic on the number of sentences in which ChatGPT accurately translated and the number of sentences in which it made errors.

We present the results obtained by comparing the translations generated by ChatGPT with expert references using BLEU ratings. In which, it offer the assessment of translation precision and correspondence with expert opinions, ranging from 0 to 1, where higher values indicate a higher degree of similarity.

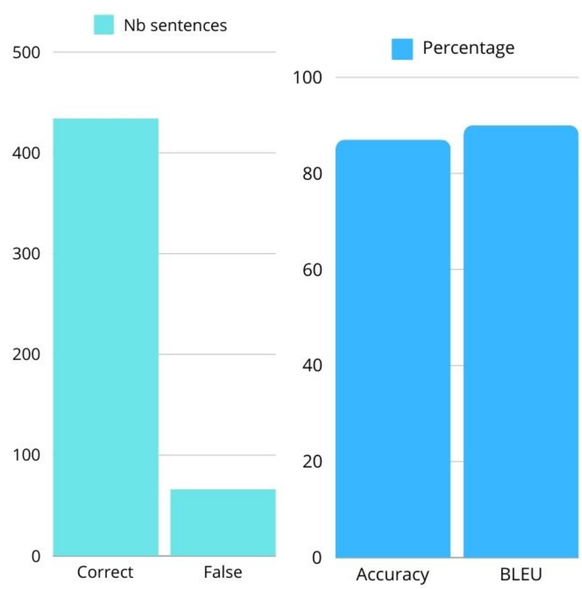


Figure 9.11: Comparison of Arabic Translations: Machine (ChatGPT) compared by human

The Figure 9.11, exhibit the results histograms in terms of correct, and false translations, as well as accuracy and BLUE score. Where we observe that a total of 434 translations out of 500 sentences were considered correct, a statement is considered correct in two cases : the sentence translated by the machine is the same as the sentence provided by the expert, or there are some differences in the choice of phrases but they convey the same meaning. Whereas, 66 translations provided by ChatGPT is considered as incorrect or inaccurate compared to human translations.

Based on this statistics, we obtained an accuracy rate of 0.87, and a high BLUE score of ≈ 0.90, indicating that the model succeeded in capturing n-grams efficiently, and that there is a strong similarity between the translations generated by the model and the reference human translations.

9.3.2 Comparison with manual annotation

The annotation task through either ChatGPT or human is considered as classification task. In our problem, the human annotations represent the reference labels of the data set, and ChatGPT is the predictive model to be tested. Therefore metrics such accuracy, F1-score, precision and recall are widely used factors to evaluate the performance of any predictive model outputs in comparison to the reference ones. Yet, applying a binary distribution of outcomes based on true prositive(TP), true negative(TN), false positive(FP), fasle negative (FN) and their decision when evaluating our model using such metrics is challenging. In which according to the human reference labeling conflicts may arise in the ChatGPT output when there is a mistaken judgment or when the case does not exist. In such scenario, the expert has two decision classes : one for when the role exists and another for when it does not exist. Therefore we can assign a value of 1 to indicate the existence of the role and a value of 0 to indicate its absence. Moreover, ChatGPT has the ability to accurately forecast the presence or absence of a role, yet it is also prone to failure. If we classify the ChatGPT findings as 1 for ”exists” and 0 for ”does not exist,” we still need a label in the case of an incorrect prediction. Hence, we suggest the use of our personalized labeling evaluator as follows (Figure 9.12).

- Expert:
  - The existence of the role is labeled as 1.
  - The absence of the role is labeled as 0.
- ChatGPT:
  - The correct prediction of an existent role is labeled as 1 (TP).

- The correct prediction of an absent role is labeled as 0 (TN).
- The wrong prediction of an existant role is labeled as -1 (FP).
- The wrong prediction of an absent role is labeled as -1 (FN).

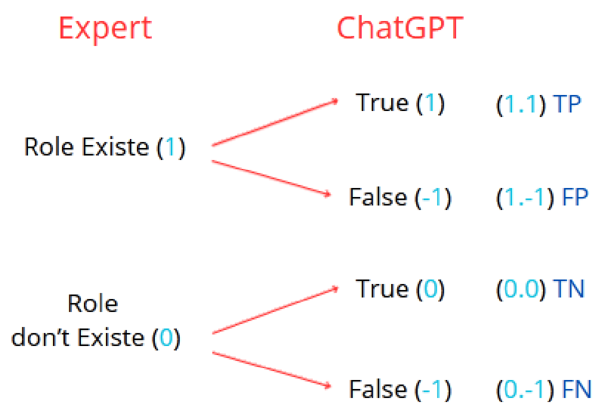


Figure 9.12: Proposed binary notation for prediction evaluation metrics

The metrics are calculated with respect to the new notation (Figure 9.12). Where we can refere to the illustrative example depicted in Figure 9.13 highlighting the case of ChatGPT role prediction, and how notations are distributed.

اليابان تعود للصيد للمرة الاولى بعد 31 عاما

Al{yAbAn taEwd lilSyod lilmar~ap Al}uwlY bEod 31 EAmA

Expert				ChatGPT			
Target	Cause	Experiencer	Cue	Target	Cause	Experiencer	Cue
		اليابان Al{yAbAn	صيد lilSyod		للمرة الاولى lilmar~ap Al}uwlY	اليابان Al{yAbAn	تعود taEwd
0	0	1	1	0	- 1	1	-1

Figure 9.13: Example of binary notation distribution of TP, TN, FP,FN

9.3.2.1 Evaluation of ChatGPT's in SRL

The initial research inquiry we posed and sought to address is whether ChatGPT may replace or assist SR annotators. To this end, we conducted a comparison between ChatGPT SRL and the annotations provided by our annotators. Specifically, we are examining the following aspects : emotion category, emotion intensity, experiencer, cue, cause, and target.

Table 9.3: Accuracy, precicion, recall and F1-score metrics of agreement between SR-ChatGPT and annotator in category, intensity and each role

		Accuracy	Precision	Recall	F1
emotion	Category	0.85	/	/	/
	intensity	0.87	/	/	/
Roles	Experiencer	0.72	0.79	0.87	0.82
	Cue	0.72	0.64	1	0.78
	Cause	0.67	0.66	0.93	0.77
	Target	0.70	0.69	0.83	0.74
	Overall	**0.76**	**0.70**	**0.91**	**0.78**

According to Table 9.3, the accuracy of emotion intensity achieved the highest level of precision with a value of 0.87. While, the caption of the main emotion achieved a remarkable accuracy rate of 0.85. This results showcase the ChatGPT's ability in recognizing the level of emotions and the expressed predominant emotions. When observing emotional arguments, an analysis of ChatGPT's effectiveness reveals that the categories 'Experiencer' and 'cue' are the most detectable, achieving a combined score of 0.72. These two categories are crucial in determining the meaning of the statement and the emotional context. Furthermore, target , achieved a precision rate of 0.70. Finally, 'Cause' has an accuracy rate of 0.67. Overall, the final F1-score of SRL is approximately 0.78, with an accuracy of 0.76.

9.3.2.2 Evaluation of ChatGPT's in cross-lingual annotation projection

The ultimate objective of the experiment is to evaluate the projection of emotion categories and SRL related to emotion arguments from English to Arabic, while maintaining the meaning, semantic roles, and grammar of the Arabic language. The Table 9.4 displays the precise measurements of ChatGPT's accuracy in correctly expressing emotion categories, intensity, and semantic roles in the Arabic language.

Table 9.4: Accuracy, precision, recall and F1-score of agreement between CL annotation projection-ChatGPT and annotators in category, intensity and emotion roles

		Accuracy	Precision	Recall	F1
emotion	Category	0.95	/	/	/
	intensity	**0.96**	/	/	/
Roles	Experiencer	0.80	0.8	0.96	0.87
	Cue	0.73	0.74	0.98	0.84
	Cause	0.70	0.72	0.98	0.82
	Target	0.64	0.69	0.84	0.76
	Overall	**0.94**	**0.74**	**0.94**	**0.82**

Where, emotion intensity achieves the highest accuracy of 0.96, followed by dominating emotion with 0.95. The accuracy of semantic roles varies from 0.80 for the experiencer to 0.64 for the target, which is the lowest. The SRL model achieved a final F1-score of around 0.82 and an accuracy of 0.94.

9.3.3 Complexity classification

9.3.3.1 ChatGPT as a tool for SRL

We randomly introduced the sentences to ChatGPT and monitored its responses. Subsequently, we compiled these responses to build the Arabic corpus annotated with SRL. We classified the sentences and counted each category separately. Table 9.5 shows statistics of diverse metrics for each complexity category.

Table 9.5: ChatGPT's SRL accuracy, precision, recall, and F1-score of emotion category, intensity and each role in three levels

		Easy				Medium				Difficult			
		Accuracy	P	R	F1	Accuracy	P	R	F1	Accuracy	P	R	F1
Emotion	Category	0.91	/	/	/	**0.94**	/	/	/	0.71	/	/	/
	Intensity	0.91	/	/	/	0.90	/	/	/	0.79	/	/	/
Role	Experiencer	0.71	0.73	0.86	0.79	0.82	0.78	0.93	0.85	0.62	0.85	0.81	0.83
	Cue	0.77	0.77	1	0.87	0.56	0.54	1	0.7	0.65	0.61	1	0.76
	Cause	0.74	0.73	0.96	0.83	0.71	0.63	0.83	0.72	0.65	0.61	1	0.76
	Target	0.74	0.63	0.96	0.8	0.56	0.69	0.67	0.7	0.71	0.74	0.85	0.72

The accuracy values vary slightly among the groups, with the emotion showing the highest value in the medium category of 0.94, and the difficult level the lowest at 0.71. While we find the intensity exhibiting the highest accuracy (0.91) in easy complexity.

For different role the results shows that the experiencer role, reach the highest accuracy in Medium level (0.82), and F1-score (0.85). However, the cue, cause and target exhibit the highest accuracy with 0.77, 0.74, 0.74 and F1-score 0.87, 0.83, 0.8 respectively in easy complexity. These results prove that the difficulty of the sentence affects the ChatGPT's efficiency; the more difficult the sentence, the less accurate the ChatGPT is in determining the semantic roles.

9.3.3.2 ChatGPT for cross-lingual Annotation Projection

The Table 9.6 presents the performance of ChatGPT in projecting SRL from English to Arabic, using different metrics within complexity levels. The accuracy of ChatGPT's translation ranged between 0.83 and 0.89 in the three levels, where it represented the highest score in easy sentences. The same performance is observed for emotion category and intensity that reached an accuracy of 1.

Table 9.6: ChatGPT's CLAP accuracy, precision, recall, and F1-score of translation, emotion category, intensity, and each role in three levels

		Easy				Medium				Difficult			
		Accuracy	P	R	F1	Accuracy	P	R	F1	Accuracy	P	R	F1
Lang.	Translation	0.85	/	/	/	**0.89**	/	/	/	0.83	/	/	/
Emotion	Category	**1**	/	/	/	0.94	/	/	/	0.90	/	/	/
	Intensity	**1**	/	/	/	0.96	/	/	/	0.94	/	/	/
Role	Experiencer	0.84	0.81	0.95	0.87	0.82	0.84	0.99	0.91	0.74	0.75	0.94	0.83
	Cue	0.87	0.86	1	0.92	0.7	0.72	1	0.84	0.62	0.65	0.95	0.77
	Cause	0.81	0.83	0.99	0.9	0.69	0.71	0.98	0.82	0.59	0.61	0.98	0.75
	Target	0.77	0.79	0.91	0.85	0.64	0.69	0.86	0.77	0.52	0.58	0.74	0.65

Among the roles, the highest accuracy of 0.87 is registered for the Cue role in the Easy level. Conversely, the lowest accuracy is observed for the Target role in the Difficult level, with an accuracy of 0.52. In terms of F1-score, the highest value is again for the Cue role in the Easy level, achieving an F1-score of 0.92. The lowest F1-score is noted for the Target role in the Difficult level, with a value of 0.65. This indicates that while ChatGPT's CLAP performs optimally for easy complexity sentences, and for the Cue role, its performance decrease significantly for more complex sentences, and for the target role.

9.3.4 Evaluating Open-LLM Performance

The table 9.7 shows the side by side performance of mBERT and ChatGPT on emotion detection and SRL tasks within the experimental settings. These experimental values assess the adaptability of our approach in both service-based LLMs, and Open-LLMs. As we observe the results, we notice that mBERT reaches an accuracy of 0.83 in detecting emotion categories and 0.84 in identifying emotion intensity achieving a very close performance of ChatGPT that scored 0.85 and an accuracy of 0.87 in emotion intensity. When analyzing SRL, ChatGPT surpasses mBERT in experiencer, cause, and target with an accuracy of (0.72, 0.67, and 0.70), respectively, while mBERT achieves an accuracy of (0.68, 0.58, and 0.49), respectively. However, mBERT shows higher performance in cue identification, with 0.85 accuracy. The obtained results exhibit the prominency and versatility of the proposed approach in enhancing LLMs performances in recognizing emotions and their intensities.

Table 9.7: A side by side analysis of mBERT and ChatGPT in Emotion Detection and SRL Tasks

Model	Aspect	Accuracy	Precision	Recall	F1
mBERT	Emotion	0.83	/	/	/
	Intensity	0.84	/	/	/
	Experiencer	0.68	0.67	0.86	0.8
	Cue	0.85	0.84	0.98	0.88
	Cause	0.58	0.51	0.72	0.64
	Target	0.49	0.55	0.67	0.70
ChatGPT	Emotion	0.85	/	/	/
	Intensity	0.87	/	/	/
	Experiencer	0.72	0.79	0.87	0.82
	Cue	0.72	0.64	1	0.78
	Cause	0.67	0.66	0.93	0.77
	Target	0.70	0.69	0.83	0.74

In the same quest, the figure 9.14 represents a bar chart comparing the BLEU scores of the mBART and ChatGPT. Highlighting their performance regarding CLAP quality. ChatGPT Achieves a Higher BLEU Score of 0.90 indicating a high-quality output compared to those of mBART 0.72 aligned with human references. This score suggests that mBART provides quality output and shows notable improvement in cross-lingual alignment, contributing valuable insights despite its slightly lower BLEU score compared to ChatGPT.

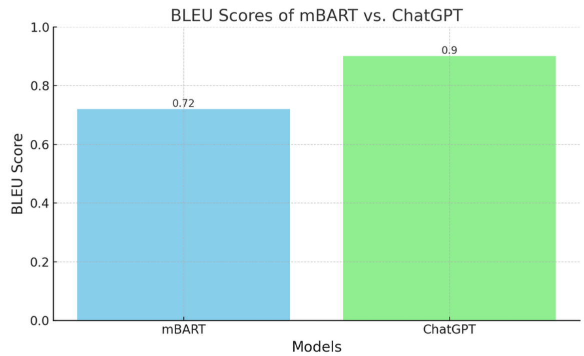


Figure 9.14: A comparison of BLEU Scores between mBART and ChatGPT

9.4 General discussion

9.4.1 Key challenges, linguistic characteristics, and driving remarks

Based on the preceding results, the likely causes for the relative low accuracy of Arabic SRL and CLAP can be summarized in a few key points. Firstly, Arabic is a linguistically intricate language when compared to English, possessing plenty of morphological and syntactic complexities. Furthermore, Arabic, being a Semitic language, in most cases, follows a Verb–Subject–Object (VSO) word order. However, English follows a Subject–Verb–Object (SVO) word order. Therefore, precisely identifying the roles might be problematic.

Furthermore, the cultural element can also exert influence, as demonstrated in the following headline: "*Greta Gerwig and Noah Baumbach Welcome Their First Child*". The annotators of the GNE categorize the birth of a baby as a "*negative surprise*", perceiving it as a burden for two parents to bear. Conversely, Arabic annotators view it as a "*positive surprise*".

Additionally, during the translation process from English to Arabic, some words may be omitted in order to preserve the overall meaning, fluency, and coherence of the text. The following words can function as a semantic role: The translation of "*Scarred By Sinema's Senate Win, Team Trump Makes Early Moves To Keep AZ Red*" is as follows:

بعد فوز سينيما في مجلس الشيوخ، قام فريق ترامب بتحركات مبكرة للحفاظ على اللون الأحمر من الألف إلى الياء .

The word "**scarred**" represented the "**Cue**" disappears in translation to maintain sentence consistency.

Moreover, while observing the obtained results, we have noticed that the Experiencer and cue play the most accurate roles. This is likely because cue is often expressed via verbal phrases that are easy to identify, and Experiencer in news headlines are direct words such as public figures, politics, or literature figures. On the other hand, detecting the target is challenging as it can be difficult, vague, or expressed through convoluted relative clauses.

In the same context, we observe that the difficulty of the sentence directly affects the ChatGPT's performance, as we notice that the ChatGPT obtained the highest results in sentences classified as easy.

The comparison between SRL and CL annotation projection, enables us to construct the following graphic curve, Figure 9.15, which show clearly that ChatGPT's CL annotation projection outperform ChatGPT's SRL.

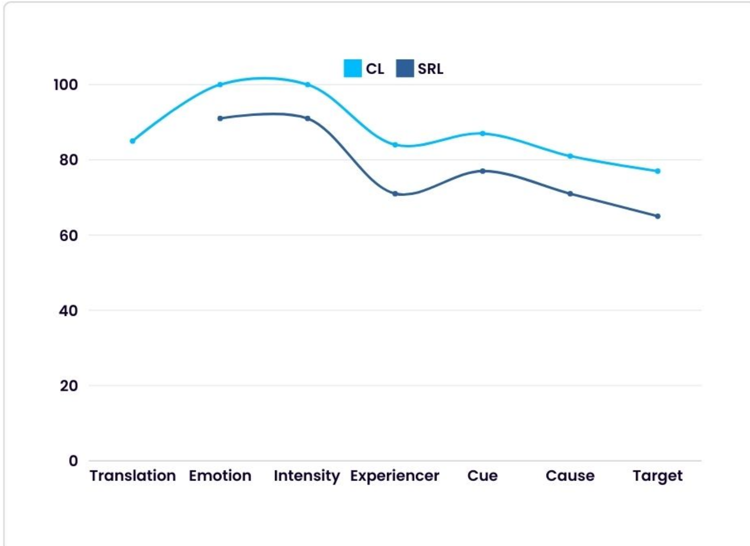


Figure 9.15: SRL vs CL Annotation projection

Moreover, the obtained results from using open-LLMs in validating our approach, affirm, our proposal in both SRLs, and CLAP automatic generation, also it suggests how that combining both proprietary and open-LLMs can be a promising approach, helping language models better capture and analyze emotions and roles across languages and contexts.

9.4.2 ChatGPT limitation

During our experiments on SRL and CLAP, we observed some limitations of ChatGPT. Here, we discuss the encountered key challenges.
As the evaluation of ChatGPT was zero-shot, we find that the conversational agent allows 37 sessions to be opened per hour; this limitation makes classification slow. Figure 9.16 shows the error message.

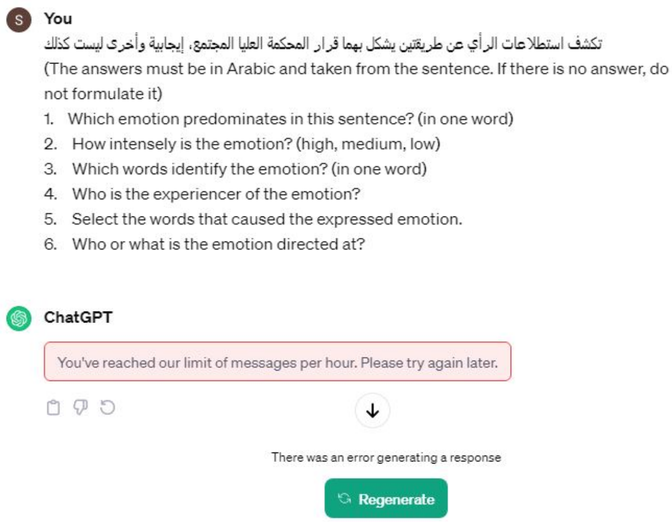


Figure 9.16: Limit of messages

Moreover, ChatGPT adds additional information to the annotation, such as the second name of proper nouns, which may be misleading and distort the semantic roles. an example in Figure 9.17.

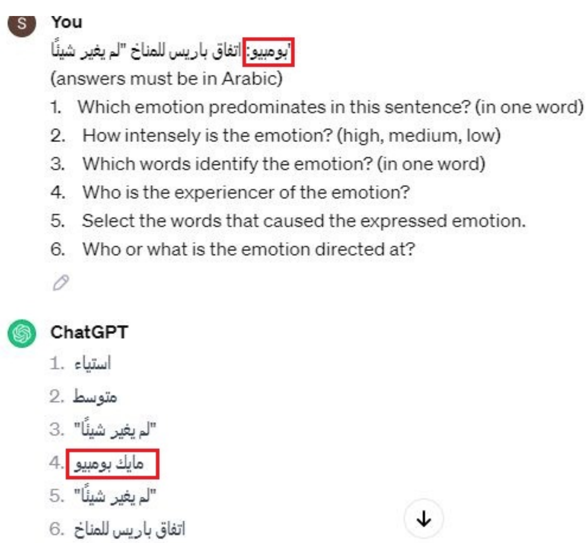


Figure 9.17: An example of adding terms to the annotation

Furthermore, in some situations, ChatGPT understands the headline itself as a direct question. Figure 9.18 clarify such situation.

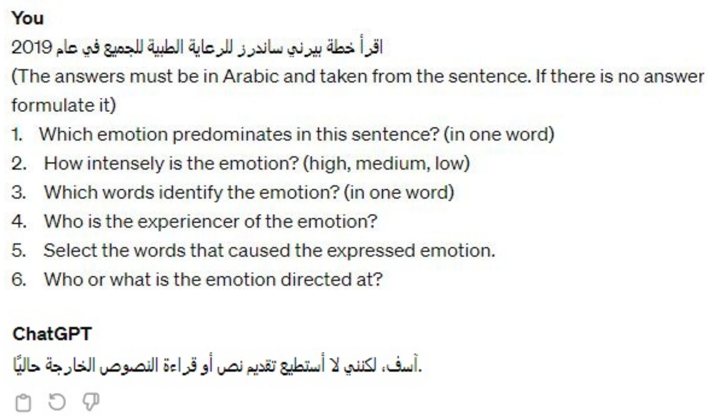


Figure 9.18: ChatGPT consider the headline as a question

9.5 Conclusion

This contribution aims to investigate alternative solutions for building new NLP resources and semantic roles labeling in low-resourced languages, that reduce efforts and time needed for manual annotation. To explore LLMs capability in some NLP tasks, we have conducted three experiments regarding to the transfer of English NLP resources to Arabic. The experiments were performed by human experts and ChatGPT and were concerned with the translation of an English corpus, the cross-lingual annotation projection of semantic roles, and finally the annotation of the translated corpus with emotion category and semantic roles.

Our experiments has assessed the accuracy of ChatGPT in performing translation, SRL and CLAP of semantic roles from English to Arabic. Our experimental results validate the research findings from earlier studies and support the hypothesis that ChatGPT could serve as a collaborative tool with humans in handling complex NLP tasks. ChatGPT's capabilities for such tasks may ensure adaptation across linguistic divergence, saves time, and reduces manual labor needed for developing new resources.

Besides validating our approach through ChatGPT's performance, the results from using open-LLMs enhance the effectiveness of our proposal in SRL and CLAP tasks. This opens up new perspectives on combining proprietary and open-LLMs to enhance models' ability to interpret emotions and roles across different languages and contexts. In conclusion, the results suggest the development of an automated SRL and cross-lingual annotation projection tool, leveraging either proprietary or open-LLMs, to assist both human and NLP experts in these tasks. Additionally, this approach could help generate and construct large Arabic datasets for future experiments and studies.

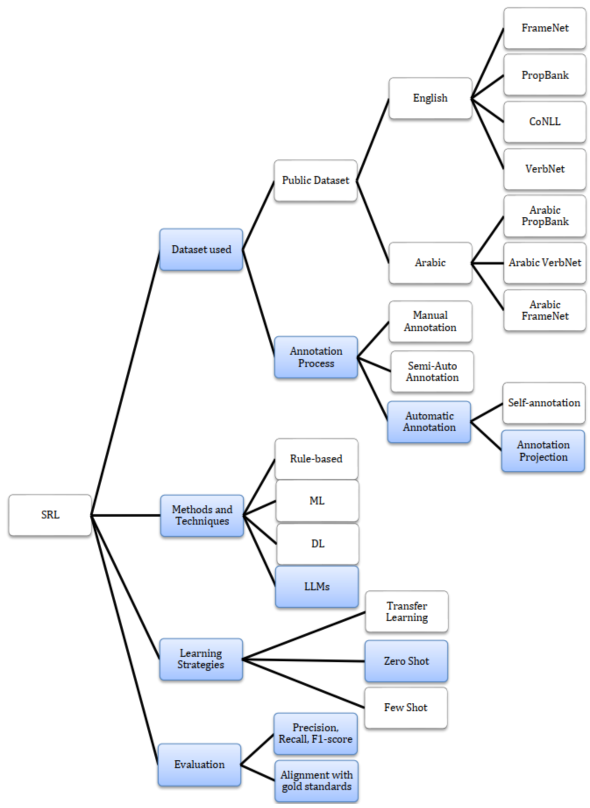


Figure 9.19: Scope of the third contribution within the SRL taxonomy

Building upon the foundations established in the first two contributions, which focused on dataset development and annotation strategies, as well as model construction using LLMs, the third contribution expands the scope further into an alternative approach to producing annotated corpora, especially for low-resource languages, cross-lingual projection and automatic evaluation by incorporating zero-shot learning strategy and comprehensive evaluation methodologies for SRL. This stage demonstrates how multilingual LLMs can be leveraged to improve SRL for Arabic with minimal resources.

Part III

Conclusion and Future Perspectives

This thesis set out to address key challenges related to the structural analysis of emotions in Arabic text. The first major challenge concerned the absence of resources annotated with emotional semantic roles, while the second involved the limited exploration of AI-based chat-bots and LLMs as tools for enhancing models and applications designed for Arabic emotion analysis. To overcome these limitations, this thesis investigated alternative solutions for building new NLP resources and supporting SRL in low-resourced languages, reducing the substantial effort and time required for manual corpus construction and annotation.

With the rapid advancement of LLMs, AI-based text generation systems have emerged as promising collaborators—if not competitors—to human annotators in several NLP tasks. To evaluate these capabilities, we conducted three experiments focused on transferring English NLP resources to Arabic. These experiments, performed by both human experts and ChatGPT, examined three tasks: translating an English corpus, performing Cross-Lingual Annotation Projection of semantic roles, and annotating the translated corpus with both emotion categories and semantic roles.

The results of our experiments assessed ChatGPT's accuracy in translation, SRL, and semantic-role CLAP, demonstrating that ChatGPT can serve as an effective collaborative tool for supporting humans in complex annotation tasks. Our findings align with previous research and reinforce the potential of LLM-based systems to ensure adaptation across linguistic divergence, save time, and reduce manual labor needed for developing new resources. Besides validating our approach through ChatGPT's performance, the results from using open-LLMs enhance the effectiveness of our proposal in SRL and CLAP tasks.

This final diagram provides a comprehensive overview of our contributions, illustrating the progressive integration of resources, models, and evaluation strategies across the different stages of the thesis. Beginning with the construction and annotation of the first Arabic ERL corpus (Contribution 1, Chapter 7), the research advanced to the application of transformer-based models on an expanded dataset (Contribution 2, Chapter 8),

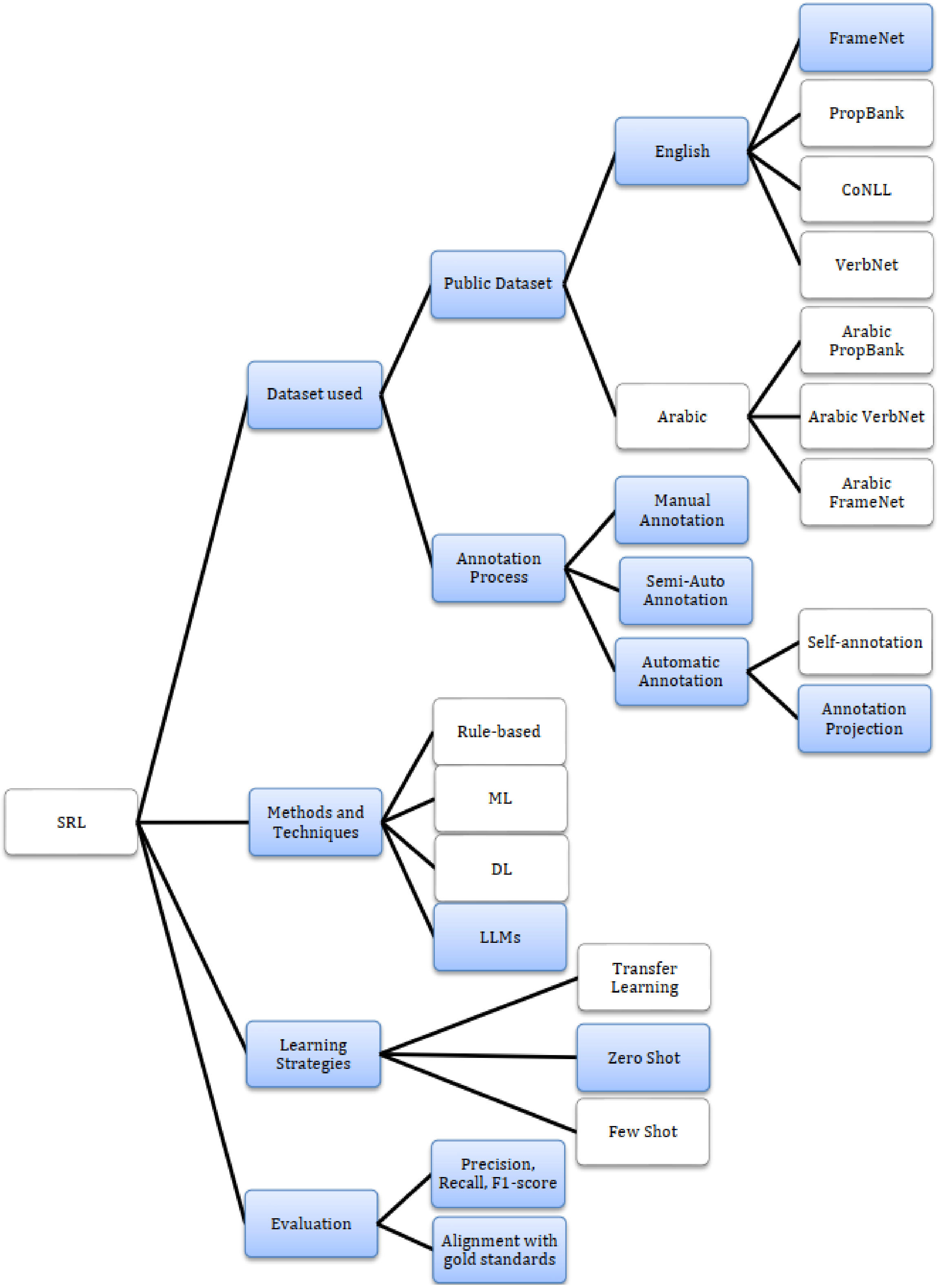


Figure 9.20: Scope of the Contributions Presented in This Thesis

and ultimately to the integration of Large Language Modelss, cross-lingual annotation techniques, zero-shot and few-shot learning strategies, and rigorous evaluation protocols (Contribution 3, Chapter 9).

The highlighted nodes in the figure summarize the full scope of the contributions, including:

- The development of an Arabic dataset based on FrameNet-inspired annotations.
- A multi-layered annotation methodology combining manual, semi-automatic, and automatic procedures.
- The use of modern computational techniques, with an emphasis on leveraging LLMs.
- The implementation of learning strategies tailored to low-resource settings, including zero-shot learning.
- A detailed evaluation framework based on both standard performance metrics and alignment with gold-standard annotations.

By progressively addressing these core dimensions of SRL for emotion analysis, this work offers a robust and comprehensive framework for advancing SRL in Arabic and other low-resource languages.
Through the successive contributions presented in this work, we have addressed nearly all key aspects of the SRL task.

Altogether, the contributions demonstrate how the combination of linguistic analysis, machine learning techniques, and multilingual capabilities can be effectively leveraged to move the field forward.
While this thesis has made significant progress in NLP field, future research offers the possibility for additional improvement. This opens up new perspectives on:

- Extending the study to other Semitic or morphologically rich languages.
- Investigating multimodal approaches by integrating text with speech, images, or video to enhance semantic interpretation.

- Incorporating more nuanced emotion categories and pragmatic factors (e.g., intensity, polarity, sarcasm) into the annotation framework.

Part IV

Bibliography

BIBLIOGRAPHY

[1] L. Bostan, E. Kim, and R. Klinger, "Goodnewseveryone: A corpus of news headlines annotated with emotions, semantic roles, and reader perception," *arXiv preprint arXiv:1912.03184*, 2019.

[2] B. Xu, H. Lin, Y. Lin, Y. Diao, L. Yang, and K. Xu, "Extracting emotion causes using learning to rank methods from an information retrieval perspective," *IEEE Access*, vol. 7, pp. 15 573–15 583, 2019.

[3] M. L. Chan, "Learning to read in hebrew and arabic: Challenges and pedagogical approaches," *Education Sciences*, vol. 14, no. 7, p. 765, 2024.

[4] C. Khairallah, R. Marzouk, S. Khalifa, M. Nassar, and N. Habash, "Computational morphology and lexicography modeling of modern standard arabic nominals," *arXiv preprint arXiv:2402.00385*, 2024.

[5] I. Hamed, C. Sabty, S. Abdennadher, N. T. Vu, T. Solorio, and N. Habash, "A survey of code-switched arabic nlp: Progress, challenges, and future directions," *arXiv preprint arXiv:2501.13419*, 2025.

[6] Y. Al Moaiad, M. Alobed, M. Alsakhnini, and A. M. Momani, "Challenges in natural arabic language processing," *Edelweiss Applied Science and Technology*, vol. 8, no. 6, pp. 4700–4705, 2024.

[7] M. Maamouri, A. Bies, T. Buckwalter, and W. Mekki, "The penn arabic treebank: Building a large-scale annotated arabic corpus," in *Proceedings of the NEMLAR International Conference on Arabic Language Resources and Tools*, Cairo, Egypt, 2004. [Online]. Available: https://www.researchgate.net/publication/228693973_The_penn_arabic_treebank_Building_a_large-scale_annotated_arabic_corpus.

[8] K. Dukes and N. Habash, "Morphological annotation of quranic arabic," in *Proceedings of the 7th International Conference on Informatics and Systems (INFOS)*, Cairo, Egypt, 2010. [Online]. Available: https://aclanthology.org/L10-1190/.

[9] T. Buckwalter, "Buckwalter arabic morphological analyzer version 1.0," *Linguistic Data Consortium, University of Pennsylvania*, pp. 86–93, 2002.

[10] S. ElKateb, W. Black, H. Rodríguez, M. Alkhalifa, P. Vossen, A. Pease, C. Fellbaum, *et al.*, "Building a wordnet for arabic.," in *LREC*, 2006, pp. 29–34.

[11] B. B. Greene and G. M. Rubin, *Automatic grammatical tagging of English*. Department of Linguistics, Brown University, 1971.

[12] R. Grishman and B. M. Sundheim, "Message understanding conference-6: A brief history," in *COLING 1996 volume 1: The 16th international conference on computational linguistics*, 1996.

[13] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? sentiment classification using machine learning techniques," *arXiv preprint cs/0205070*, 2002.

[14] B. F. Green Jr, A. K. Wolf, C. Chomsky, and K. Laughery, "Baseball: An automatic question-answerer," in *Papers presented at the May 9-11, 1961, western joint IRE-AIEE-ACM computer conference*, 1961, pp. 219–224.

[15] D. Gildea and D. Jurafsky, "Automatic labeling of semantic roles," *Computational linguistics*, vol. 28, no. 3, pp. 245–288, 2002.

[16] R. M. Bakker, A. J. Schoevers, R. A. van Drie, M. P. Schraagen, and M. H. de Boer, “Semantic role extraction in law texts: A comparative analysis of language models for legal information extraction,” *Artificial Intelligence and Law*, pp. 1–35, 2025.

[17] J. Persson, R. Johansson, and P. Nugues, “Text categorization using predicate-argument structures,” in *Proceedings of the 17th Nordic Conference of Computational Linguistics (NODALIDA 2009)*, 2009, pp. 142–149.

[18] A. Klein, J. Mamou, V. Pyatkin, D. Stepanov, H. He, D. Roth, L. Zettlemoyer, and I. Dagan, “Qanom: Question-answer driven srl for nominalizations,” in *Proceedings of the 28th international conference on computational linguistics*, 2020, pp. 3069–3083.

[19] P. Nakov, D. Hoogeveen, L. Màrquez, A. Moschitti, H. Mubarak, T. Baldwin, and K. Verspoor, “Semeval-2017 task 3: Community question answering,” *arXiv preprint arXiv:1912.00730*, 2019.

[20] K. Chakma, A. Das, and S. Debbarma, “Deep semantic role labeling for tweets using 5w1h: Who, what, when, where, why and how,” *Computación y Sistemas*, vol. 23, no. 3, pp. 751–763, 2019.

[21] V. Miśra, *The Descriptive Technique of Pānini*. Mouton, 1966, vol. 18.

[22] C. J. Fillmore, “Frame semantics and the nature of language,” *Annals of the New York Academy of Sciences*, vol. 280, no. 1, pp. 20–32, 1976.

[23] J. Ruppenhofer, M. Ellsworth, M. Schwarzer-Petruck, C. R. Johnson, and J. Scheffczyk, “Framenet ii: Extended theory and practice,” International Computer Science Institute, Tech. Rep., 2016.

[24] C. Subirats and H. Sato, “Spanish framenet and framesql,” in *4th International Conference on Language Resources and Evaluation. Workshop on Building Lexical Resources from Semantically Annotated Corpora. Lisbon (Portugal)*, 2004.

[25] O. Lyashevskaya and E. Kashkin, “Framebank: A database of russian lexical constructions,” in *International Conference on Analysis of Images, Social Networks and Texts*, Springer, 2015, pp. 350–360.

[26] K. H. Ohara, S. Fujii, T. Ohori, R. Suzuki, H. Saito, and S. Ishizaki, “The japanese framenet project: An introduction,” in *Proceedings of LREC-04 Satellite Workshop “Building Lexical Resources from Semantically Annotated Corpora”(LREC 2004)*, 2004, pp. 9–11.

[27] M. Palmer, D. Gildea, and P. Kingsbury, “The proposition bank: An annotated corpus of semantic roles,” *Computational linguistics*, vol. 31, no. 1, pp. 71–106, 2005.

[28] Q. Zhou, Z. Jiang, and F. Yang, “Sentences similarity based on deep structured semantic model and semantic role labeling,” in *2020 International Conference on Asian Language Processing (IALP)*, IEEE, 2020, pp. 40–44.

[29] N. Xue and M. Palmer, “Annotating the propositions in the penn chinese treebank,” in *Proceedings of the second SIGHAN workshop on Chinese language processing*, 2003, pp. 47–54.

[30] M. Palmer, O. Babko-Malaya, A. Bies, M. T. Diab, M. Maamouri, A. Mansouri, and W. Zaghouani, “A pilot arabic propbank.,” in *LREC*, 2008.

[31] G. G. Şahin and E. Adalı, “Annotation of semantic roles for the turkish proposition bank,” *Language Resources and Evaluation*, vol. 52, pp. 673–706, 2018.

[32] K. Haverinen, J. Kanerva, S. Kohonen, A. Missilä, S. Ojala, T. Viljanen, V. Laippala, and F. Ginter, “The finnish proposition bank,” *Language Resources and Evaluation*, vol. 49, pp. 907–926, 2015.

[33] X. Carreras and L. Màrquez, “Introduction to the conll-2005 shared task: Semantic role labeling,” in *Proceedings of the ninth conference on computational natural language learning (CoNLL-2005)*, 2005, pp. 152–164.

[34] M. Surdeanu, R. Johansson, A. Meyers, L. Màrquez, and J. Nivre, "The conll 2008 shared task on joint parsing of syntactic and semantic dependencies," in *CoNLL 2008: Proceedings of the Twelfth Conference on Computational Natural Language Learning*, 2008, pp. 159–177.

[35] J. Hajic, M. Ciaramita, R. Johansson, D. Kawahara, M. A. Martí, L. Màrquez, A. Meyers, J. Nivre, S. Padó, J. Štěpánek, *et al.*, "The conll-2009 shared task: Syntactic and semantic dependencies in multiple languages," in *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL 2009): Shared Task*, 2009, pp. 1–18.

[36] R. Cai and M. Lapata, "Syntax-aware semantic role labeling without parsing," *Transactions of the Association for Computational Linguistics*, vol. 7, pp. 343–356, 2019.

[37] C. Fellbaum, "Wordnet: An electronic lexical database," *MIT Press google schola*, vol. 2, pp. 678–686, 1998.

[38] C. B. Khelil, D. Duchier, C. B. O. Zribi, and Y. Parmentier, "A new syntactic-semantic interface for arabtag an arabic tree adjoining grammar," in *The 18th International Arab Conference on Information Technology (ACIT'2017)*, 2017.

[39] A. Lakhfif and M. T. Laskri, "A frame-based approach for capturing semantics from arabic text for text-to-sign language mt," *International Journal of Speech Technology*, vol. 19, no. 2, pp. 203–228, 2016.

[40] T. Underwood, D. Bamman, and S. Lee, "The transformation of gender in english-language fiction," *Journal of Cultural Analytics*, vol. 3, no. 2, 2018.

[41] H. Zhang, F. Boons, and R. Batista-Navarro, "Whose story is it anyway? automatic extraction of accounts from news articles," *Information processing & management*, vol. 56, no. 5, pp. 1837–1848, 2019.

[42] C. Campagnano, S. Conia, and R. Navigli, "Srl4e–semantic role labeling for emotions: A unified evaluation framework," in *Proceedings of the 60th Annual*

Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2022, pp. 4586–4601.

[43] S. Mohammad, X. Zhu, and J. Martin, "Semantic role labeling of emotions in tweets," in *Proceedings of the 5th workshop on computational approaches to subjectivity, sentiment and social media analysis*, 2014, pp. 32–41.

[44] E. Bassignana, V. Basile, V. Patti, *et al.*, "Hurtlex: A multilingual lexicon of words to hurt," in *CEUR Workshop proceedings*, CEUR-WS, vol. 2253, 2018, pp. 1–6.

[45] R. Cai and M. Lapata, "Alignment-free cross-lingual semantic role labeling," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 3883–3894.

[46] F. Senator, H. Boutouta, A. Lakhfif, and C. Mediani, "Semantic role labeling of arabic emotional text in tweets," in *2023 International Conference on Advances in Electronics, Control and Communication Systems (ICAECCS)*, IEEE, 2023, pp. 1–6.

[47] S. Zhang, O. Jafari, and P. Nagarkar, "A survey on machine learning techniques for auto labeling of video, audio, and text data," *arXiv preprint arXiv:2109.03784*, 2021.

[48] Y. Xu, J. Zhang, R. He, L. Ge, C. Yang, C. Yang, and Y. N. Wu, "Sas: Self-augmentation strategy for language model pre-training," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, 2022, pp. 11 586–11 594.

[49] J. Liu, C. Liang, and J. Xu, "Document-level event argument extraction with self-augmentation and a cross-domain joint training mechanism," *Knowledge-Based Systems*, vol. 257, p. 109 904, 2022.

[50] S. Padó and M. Lapata, "Cross-lingual annotation projection for semantic roles," *Journal of Artificial Intelligence Research*, vol. 36, pp. 307–340, 2009.

[51] S. Gunasekara, D. Chathura, C. Jeewantha, and G. Dias, "Using annotation projection for semantic role labeling of low-resourced language: Sinhala," in *2020 International Conference on Asian Language Processing (IALP)*, IEEE, 2020, pp. 98–103.

[52] Z. M. Kyu and N. L. Wah, "Level-wise segment parser for semantic role labelling for myanmar text," in *2018 IEEE/ACIS 17th International Conference on Computer and Information Science (ICIS)*, IEEE, 2018, pp. 286–290.

[53] B. Mahesh, "Machine learning algorithms-a review," *International Journal of Science and Research (IJSR).[Internet]*, vol. 9, no. 1, pp. 381–386, 2020.

[54] P. P. Shinde and S. Shah, "A review of machine learning and deep learning applications," in *2018 Fourth international conference on computing communication control and automation (ICCUBEA)*, IEEE, 2018, pp. 1–6.

[55] S. Dong, P. Wang, and K. Abbas, "A survey on deep learning and its applications," *Computer Science Review*, vol. 40, p. 100 379, 2021.

[56] J. Zhou and W. Xu, "End-to-end learning of semantic role labeling using recurrent neural networks," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2015, pp. 1127–1137.

[57] L. He, K. Lee, M. Lewis, and L. Zettlemoyer, "Deep semantic role labeling: What works and what's next," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2017, pp. 473–483.

[58] Z. Tan, M. Wang, J. Xie, Y. Chen, and X. Shi, "Deep semantic role labeling with self-attention," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, 2018.

[59] J. Bae and C. Lee, "Korean semantic role labeling with bidirectional encoder representations from transformers and simple semantic information," *Applied Sciences*, vol. 12, no. 12, p. 5995, 2022.

[60] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[61] K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," *Journal of Big data*, vol. 3, pp. 1–40, 2016.

[62] A. Subburathinam, D. Lu, H. Ji, J. May, S.-F. Chang, A. Sil, and C. Voss, "Cross-lingual structure transfer for relation and event extraction," in *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (emnlp-ijcnlp)*, 2019, pp. 313–325.

[63] F. Pourpanah, M. Abdar, Y. Luo, X. Zhou, R. Wang, C. P. Lim, X.-Z. Wang, and Q. J. Wu, "A review of generalized zero-shot learning methods," *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 4, pp. 4051–4070, 2022.

[64] A. Pelicon, M. Pranjić, D. Miljković, B. Škrlj, and S. Pollak, "Zero-shot learning for cross-lingual news sentiment classification," *Applied Sciences*, vol. 10, no. 17, p. 5993, 2020.

[65] M. Bombieri, M. Rospocher, S. P. Ponzetto, and P. Fiorini, "Machine understanding surgical actions from intervention procedure textbooks," *Computers in Biology and Medicine*, vol. 152, p. 106 415, 2023.

[66] A. Parnami and M. Lee, "Learning from few examples: A summary of approaches to few-shot learning," *arXiv preprint arXiv:2203.04291*, 2022.

[67] I. Jindal, A. Rademaker, K.-N. Tran, H. Zhu, H. Kanayama, M. Danilevsky, and Y. Li, "Primesrl-eval: A practical quality metric for semantic role labeling systems evaluation," *arXiv preprint arXiv:2210.06408*, 2022.

[68] H. Boutouta, A. Lakhfif, F. Senator, and C. Mediani, "Enhancement of implicit emotion recognition in arabic text: Annotated dataset and baseline models," *IEEE Access*, 2025.

[69] H. Boutouta, A. Lakhfif, F. Senator, and C. Mediani, “A transformer-based hybrid model for implicit emotion recognition in arabic text,” *Engineering, Technology & Applied Science Research*, vol. 15, no. 3, pp. 23 834–23 839, 2025.

[70] Oxford University Press. “Emotion.” Accessed: 2025-04-14. (2024), [Online]. Available: https://www.oxfordlearnersdictionaries.com/definition/american_english/emotion.

[71] Cambridge University Press. “Emotion - definition in the cambridge english dictionary.” Accessed: 2025-04-14. (2024), [Online]. Available: https://dictionary.cambridge.org/dictionary/english/emotion.

[72] American Psychological Association. “Emotion.” Accessed: 2025-04-14. (2024), [Online]. Available: https://dictionary.apa.org/emotion.

[73] S. Peng, L. Cao, Y. Zhou, Z. Ouyang, A. Yang, X. Li, W. Jia, and S. Yu, “A survey on deep learning for textual emotion analysis in social networks,” *Digital Communications and Networks*, vol. 8, no. 5, pp. 745–762, 2022.

[74] P. Ekman, “An argument for basic emotions,” *Cognition & emotion*, vol. 6, no. 3-4, pp. 169–200, 1992.

[75] R. Plutchik, *A psychoevolutionary theory of emotions*, 1982.

[76] J. A. Russell, “A circumplex model of affect,” *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1161–1178, 1980.

[77] K. R. Scherer, “What are emotions? and how can they be measured?” *Social Science Information*, vol. 44, no. 4, pp. 695–729, 2005.

[78] R. A. Calvo, D. N. Milne, M. S. Hussain, and H. Christensen, “Natural language processing in mental health applications using non-clinical texts,” *Natural Language Engineering*, vol. 23, no. 5, pp. 649–685, 2017.

[79] N. Alswaidan and M. E. B. Menai, “A survey of state-of-the-art approaches for emotion recognition in text,” *Knowledge and Information Systems*, vol. 62, no. 8, pp. 2937–2987, 2020.

[80] F. A. Acheampong, C. Wenyu, and H. Nunoo-Mensah, "Text-based emotion detection: Advances, challenges, and opportunities," *Engineering Reports*, vol. 2, no. 7, e12189, 2020.

[81] M. L. Loureiro and M. Alló, "Sensing climate change and energy issues: Sentiment and emotion analysis with social media in the uk and spain," *Energy Policy*, vol. 143, p. 111 490, 2020.

[82] M. Sykora, S. Elayan, I. R. Hodgkinson, T. W. Jackson, and A. West, "The power of emotions: Leveraging user generated content for customer experience management," *Journal of Business Research*, vol. 144, pp. 997–1006, 2022.

[83] Q. Wang, T. Su, R. Y. K. Lau, and H. Xie, "Deepemotionnet: Emotion mining for corporate performance analysis and prediction," *Information Processing & Management*, vol. 60, no. 3, p. 103 151, 2023.

[84] S. M. Srinivasan, R. Sangwan, C. Neill, and T. Zu, "Power of predictive analytics: Using emotion classification of twitter data for predicting 2016 us presidential elections," *The Journal of Social Media in Society*, vol. 8, no. 1, pp. 211–230, 2019.

[85] H. Rashkin, E. M. Smith, M. Li, and Y.-L. Boureau, "Towards empathetic open-domain conversation models: A new benchmark and dataset," *arXiv preprint arXiv:1811.00207*, 2019.

[86] Y. Ma, K. L. Nguyen, F. Z. Xing, and E. Cambria, "A survey on empathetic dialogue systems," *Information Fusion*, vol. 64, pp. 50–70, 2020.

[87] N. M. Razali, A. Harun, V. Michaael, N. A. A. Hishamuddin, and A. Ayob, "Emote, buy, enjoy: Mapping the intensity of emotions for virtual skins with plutchik's wheel," *Idealogy Journal*, vol. 10, no. 2, 2025.

[88] W. G. Parrott, *Emotions in social psychology: Essential readings*. psychology press, 2001.

[89] S. M. Mohammad and P. D. Turney, "Nrc emotion lexicon," *National Research Council, Canada*, vol. 2, p. 234, 2013.

[90] D. Demszky, D. Movshovitz-Attias, J. Ko, A. Cowen, G. Nemade, and S. Ravi, "Goemotions: A dataset of fine-grained emotions," *arXiv preprint arXiv:2005.00547*, 2020.

[91] E. Öhman, M. Pàmies, K. Kajava, and J. Tiedemann, "Xed: A multilingual dataset for sentiment analysis and emotion detection," *arXiv preprint arXiv:2011.01612*, 2020.

[92] Y. Li, H. Su, X. Shen, W. Li, Z. Cao, and S. Niu, "Dailydialog: A manually labelled multi-turn dialogue dataset," *arXiv preprint arXiv:1710.03957*, 2017.

[93] S. Buechel and U. Hahn, "Emobank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis," *arXiv preprint arXiv:2205.01996*, 2022.

[94] H. Rashkin, E. M. Smith, M. Li, and Y.-L. Boureau, "Towards empathetic open-domain conversation models: A new benchmark and dataset," *arXiv preprint arXiv:1811.00207*, 2018.

[95] E. Troiano, S. Padó, and R. Klinger, "Emotion ratings: How intensity, annotation confidence and agreements are entangled," *arXiv preprint arXiv:2103.01667*, 2021.

[96] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, "Meld: A multimodal multi-party dataset for emotion recognition in conversations," *arXiv preprint arXiv:1810.02508*, 2018.

[97] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, pp. 335–359, 2008.

[98] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the recola multimodal corpus of remote collaborative and affective interactions," in *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)*, IEEE, 2013, pp. 1–8.

[99] O. Araque, L. Gatti, J. Staiano, and M. Guerini, "Depechemood++: A bilingual emotion lexicon built through simple yet powerful techniques," *IEEE transactions on affective computing*, vol. 13, no. 1, pp. 496–507, 2019.

[100] P.-L. Huguet-Cabot, D. Abadi, A. Fischer, and E. Shutova, "Us vs. them: A dataset of populist attitudes, news bias and emotions," *arXiv preprint arXiv:2101.11956*, 2021.

[101] S. G. Tesfagergish, J. Kapočiūtė-Dzikienė, and R. Damaševičius, "Zero-shot emotion detection for semi-supervised sentiment analysis using sentence transformers and ensemble learning," *Applied Sciences*, vol. 12, no. 17, p. 8662, 2022.

[102] G. Guibon, M. Labeau, L. Lefeuvre, and C. Clavel, "Few-shot emotion recognition in conversation with sequential prototypical networks," *Software Impacts*, vol. 12, p. 100 237, 2022.

[103] D. Ghazi, D. Inkpen, and S. Szpakowicz, "Detecting emotion stimuli in emotion-bearing sentences," in *Computational Linguistics and Intelligent Text Processing: 16th International Conference, CICLing 2015, Cairo, Egypt, April 14-20, 2015, Proceedings, Part II 16*, Springer, 2015, pp. 152–165.

[104] E. Kim and R. Klinger, "Who feels what and why? annotation of a literature corpus with semantic roles of emotions," in *Proceedings of the 27th International Conference on Computational Linguistics*, 2018, pp. 1345–1359.

[105] B. M. D. Dang, L. Oberländer, and R. Klinger, "Emotion stimulus detection in german news headlines," *arXiv preprint arXiv:2107.12920*, 2021.

[106] L. Gui, L. Yuan, R. Xu, B. Liu, Q. Lu, and Y. Zhou, "Emotion cause detection with linguistic construction in chinese weibo text," in *Natural Language Processing and Chinese Computing: Third CCF Conference, NLPCC 2014, Shenzhen, China, December 5-9, 2014. Proceedings 3*, Springer, 2014, pp. 457–464.

[107] S. Y. M. Lee, Y. Chen, S. Li, C.-R. Huang, *et al.*, "Emotion cause events: Corpus construction and analysis.," in *LREC*, 2010.

[108] L. Gui, R. Xu, D. Wu, Q. Lu, and Y. Zhou, "Event-driven emotion cause extraction with corpus construction," in *Social Media Content Analysis: Natural Language Processing and Beyond*, World Scientific, 2018, pp. 145–160.

[109] X. Cheng, Y. Chen, B. Cheng, S. Li, and G. Zhou, "An emotion cause corpus for chinese microblogs with multiple-user structures," *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, vol. 17, no. 1, pp. 1–19, 2017.

[110] A. Neviarouskaya and M. Aono, "Extracting causes of emotions from text," in *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, 2013, pp. 932–936.

[111] I. Russo, T. Caselli, F. Rubino, E. Boldrini, P. Martínez-Barco, *et al.*, "Emocause: An easy-adaptable approach to emotion cause contexts," Association for Computational Linguistics (ACL), 2011.

[112] S. Yada, K. Ikeda, K. Hoashi, and K. Kageura, "A bootstrap method for automatic rule acquisition on emotion cause extraction," in *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, IEEE, 2017, pp. 414–421.

[113] A. D. P. Ariyanto, D. Purwitasari, and C. Fatichah, "A systematic review on semantic role labeling for information extraction in low-resource data," *IEEE Access*, 2024.

[114] S. Mohammad, F. Bravo-Marquez, M. Salameh, and S. Kiritchenko, "Semeval-2018 task 1: Affect in tweets," in *Proceedings of the 12th international workshop on semantic evaluation*, 2018, pp. 1–17.

[115] S. Lazemi, H. Ebrahimpour-Komleh, and N. Noroozi, "Improving persian dependency-based semantic role labeling using semantic and structural relations," in *2019 4th International Conference on Pattern Recognition and Image Analysis (IPRIA)*, IEEE, 2019, pp. 163–167.

[116] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, “Unsupervised cross-lingual representation learning at scale,” *arXiv preprint arXiv:1911.02116*, 2019.

[117] C. J. Fillmore, “The case for case.,” 1967.

[118] D. Gildea and D. Jurafsky, “Automatic labeling of semantic roles,” *Computational linguistics*, vol. 28, no. 3, pp. 245–288, 2002.

[119] S. Pradhan, K. Hacioglu, W. Ward, J. H. Martin, and D. Jurafsky, “Semantic role chunking combining complementary syntactic views,” in *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)*, 2005, pp. 217–220.

[120] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, “Natural language processing (almost) from scratch,” *Journal of machine learning research*, vol. 12, pp. 2493–2537, 2011.

[121] A. Björkelund, L. Hafdell, and P. Nugues, “Multilingual semantic role labeling,” in *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL 2009): Shared Task*, 2009, pp. 43–48.

[122] L. Yuan, “Semantic role labeling based on valence structure and deep neural network,” *IETE Journal of Research*, vol. 70, no. 5, pp. 5044–5052, 2024.

[123] H. Fei, M. Zhang, B. Li, and D. Ji, “End-to-end semantic role labeling with neural transition-based model,” in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, 2021, pp. 12 803–12 811.

[124] K. Munir, H. Zhao, and Z. Li, “Neural unsupervised semantic role labeling,” *Transactions on Asian and Low-Resource Language Information Processing*, vol. 20, no. 6, pp. 1–16, 2021.

[125] R. Devianti and Y. Miyao, “Transferability of syntax-aware graph neural networks in zero-shot cross-lingual semantic role labeling,” in *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024, pp. 20–42.

[126] D. Marcheggiani, A. Frolov, and I. Titov, “A simple and accurate syntax-agnostic neural model for dependency-based semantic role labeling,” *arXiv preprint arXiv:1701.02593*, 2017.

[127] J. Cheng, “Long short-term memory-networks for machine reading,” *arXiv preprint arXiv:1601.06733*, 2016.

[128] E. Strubell, P. Verga, D. Andor, D. Weiss, and A. McCallum, “Linguistically-informed self-attention for semantic role labeling,” *arXiv preprint arXiv:1804.08199*, 2018.

[129] S. V. Mehta, J. Y. Lee, and J. Carbonell, “Towards semi-supervised learning for deep semantic role labeling,” *arXiv preprint arXiv:1808.09543*, 2018.

[130] A Vaswani, “Attention is all you need,” *Advances in Neural Information Processing Systems*, 2017.

[131] Y. Zhang, R. Wang, and L. Si, “Syntax-enhanced self-attention-based semantic role labeling,” *arXiv preprint arXiv:1910.11204*, 2019.

[132] Z. Lin, M. Feng, C. N. d. Santos, M. Yu, B. Xiang, B. Zhou, and Y. Bengio, “A structured self-attentive sentence embedding,” *arXiv preprint arXiv:1703.03130*, 2017.

[133] R Paulus, “A deep reinforced model for abstractive summarization,” *arXiv preprint arXiv:1705.04304*, 2017.

[134] T. Shen, T. Zhou, G. Long, J. Jiang, S. Pan, and C. Zhang, “Disan: Directional self-attention network for rnn/cnn-free language understanding,” in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, 2018.

[135] J. Park, “Selectively connected self-attentions for semantic role labeling,” *Applied Sciences*, vol. 9, no. 8, p. 1716, 2019.

[136] C. Qin, A. Zhang, Z. Zhang, J. Chen, M. Yasunaga, and D. Yang, “Is chatgpt a general-purpose natural language processing task solver?” *arXiv preprint arXiv:2302.06476*, 2023.

[137] B. D. Lund and T. Wang, “Chatting about chatgpt: How may ai and gpt impact academia and libraries?” *Library hi tech news*, vol. 40, no. 3, pp. 26–29, 2023.

[138] V. Goar, N. S. Yadav, and P. S. Yadav, “Conversational ai for natural language processing: An review of chatgpt,” *International Journal on Recent and Innovation Trends in Computing and Communication*, vol. 11, pp. 109–117, 2023.

[139] Y. Tan, D. Min, Y. Li, W. Li, N. Hu, Y. Chen, and G. Qi, “Can chatgpt replace traditional kbqa models? an in-depth analysis of the question answering performance of the gpt llm family,” in *International Semantic Web Conference*, Springer, 2023, pp. 348–367.

[140] J. Wang, Y. Liang, F. Meng, Z. Sun, H. Shi, Z. Li, J. Xu, J. Qu, and J. Zhou, “Is chatgpt a good nlg evaluator? a preliminary study,” *arXiv preprint arXiv:2303.04048*, 2023.

[141] A. Hendy, M. Abdelrehim, A. Sharaf, V. Raunak, M. Gabr, H. Matsushita, Y. J. Kim, M. Afify, and H. H. Awadalla, “How good are gpt models at machine translation? a comprehensive evaluation,” *arXiv preprint arXiv:2302.09210*, 2023.

[142] J. Ye, X. Chen, N. Xu, C. Zu, Z. Shao, S. Liu, Y. Cui, Z. Zhou, C. Gong, Y. Shen, *et al.*, “A comprehensive capability analysis of gpt-3 and gpt-3.5 series models,” *arXiv preprint arXiv:2303.10420*, 2023.

[143] B. Ding, C. Qin, L. Liu, Y. K. Chia, S. Joty, B. Li, and L. Bing, “Is gpt-3 a good data annotator?” *arXiv preprint arXiv:2212.10450*, 2022.

[144] J. Gao, H. Zhao, C. Yu, and R. Xu, “Exploring the feasibility of chatgpt for event extraction,” *arXiv preprint arXiv:2303.03836*, 2023.

[145] N. Cheng, Z. Yan, Z. Wang, Z. Li, J. Yu, Z. Zheng, K. Tu, J. Xu, and W. Han, “Potential and limitations of llms in capturing structured semantics: A case study on srl,” in *International Conference on Intelligent Computing*, Springer, 2024, pp. 50–61.

[146] X. Sun, L. Dong, X. Li, Z. Wan, S. Wang, T. Zhang, J. Li, F. Cheng, L. Lyu, F. Wu, *et al.*, "Pushing the limits of chatgpt on nlp tasks," *arXiv preprint arXiv:2306.09719*, 2023.

[147] H. Boutouta, A. Lakhfif, F. Senator, and C. Mediani, "From context to emotion: Leveraging llms for recognizing implicit emotions," in *Proceedings of the 8th International Conference on Natural Language and Speech Processing (ICNLSP-2025)*, 2025, pp. 399–409.

[148] M. Diab, A. Moschitti, and D. Pighin, "Cunit: A semantic role labeling system for modern standard arabic," in *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, 2007, pp. 133–136.

[149] M. Diab, M. Alkhalifa, S. ElKateb, C. Fellbaum, A. Mansouri, and M. Palmer, "Semeval-2007 task 18: Arabic semantic labeling," in *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, 2007, pp. 93–98.

[150] M. Diab, A. Moschitti, and D. Pighin, "Semantic role labeling systems for arabic using kernel methods," in *Proceedings of ACL-08: HLT*, 2008, pp. 798–806.

[151] A. Lakhfif, M. T. Laskri, and E. Atwell, "Multi-level analysis and annotation of arabic corpora for text-to-sign language mt," *arXiv preprint arXiv:1605.07346*, 2016.

[152] A. Lakhfif and M. T. Laskri, "Framenet based analysis and annotation: Contribution to the contrastive study of motion events in arabic and english," *Traitement automatique des langues*, vol. 58, no. 3, pp. 67–95, 2017.

[153] A. Lakhfif, "Design and implementation of a virtual 3d educational environment to improve deaf education," *arXiv preprint arXiv:2006.00114*, 2020.

[154] A. Lakhfif and M. T. Laskri, "A frame semantics based approach to comparative study of digitized corpus," *arXiv preprint arXiv:2006.00113*, 2020.

[155] H. Meguehout, T. Bouhadada, and M. T. Laskri, "Semantic role labeling for arabic language using case-based reasoning approach," *International Journal of Speech Technology*, vol. 20, pp. 363–372, 2017.

[156] L. Oberländer, K. Reich, and R. Klinger, "Experiencers, stimuli, or targets: Which semantic roles enable machine learning to infer the emotions?" *arXiv preprint arXiv:2011.01599*, 2020.

[157] L. Oberländer and R. Klinger, "Token sequence labeling vs. clause classification for english emotion stimulus detection," *arXiv preprint arXiv:2010.07557*, 2020.

[158] W. Li and H. Xu, "Text-based emotion classification using emotion cause extraction," *Expert Systems with Applications*, vol. 41, no. 4, pp. 1742–1749, 2014.

[159] K. Gao, H. Xu, and J. Wang, "A rule-based approach to emotion cause detection for chinese micro-blogs," *Expert Systems with Applications*, vol. 42, no. 9, pp. 4517–4528, 2015.

[160] L. Gui, Y. Zhou, Q. Liu, Y. Zhou, M. Zhang, and X. Huang, "Event-driven emotion cause extraction with corpus construction," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2016, pp. 1639–1649.

[161] R. Gao, Z. Jia, X. Zhang, and R. Wang, "Emotion cause extraction using emotion cue and semantic roles," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2017, pp. 1635–1644.

[162] Y. Chen, W. Hou, S. Li, C. Wu, and X. Zhang, "End-to-end emotion-cause pair extraction with graph convolutional network," in *Proceedings of the 28th International Conference on Computational Linguistics (COLING)*, Barcelona, Spain (Online): International Committee on Computational Linguistics, 2020, pp. 198–207.

[163] S. Padó and M. Lapata, "Cross-lingual annotation projection for semantic roles," vol. 36, 2009, pp. 307–340.

[164] M. T. I. Khondaker, A. Waheed, E. M. B. Nagoudi, and M. Abdul-Mageed, "Gptaraeval: A comprehensive evaluation of chatgpt on arabic nlp," *arXiv preprint arXiv:2305.14976*, 2023.

[165] V. D. Lai, N. T. Ngo, A. P. B. Veyseh, H. Man, F. Dernoncourt, T. Bui, and T. H. Nguyen, "Chatgpt beyond english: Towards a comprehensive evaluation of large language models in multilingual learning," *arXiv preprint arXiv:2304.05613*, 2023.

[166] J. Wang, Y. Liang, F. Meng, B. Zou, Z. Li, J. Qu, and J. Zhou, "Zero-shot cross-lingual summarization via large language models," in *Proceedings of the 4th New Frontiers in Summarization Workshop*, Singapore: Association for Computational Linguistics, Dec. 2023, pp. 12–23. DOI: 10.18653/v1/2023.newsum-1.2.

[167] X. Zhang, S. Li, B. Hauer, N. Shi, and G. Kondrak, "Don't trust chatgpt when your question is not in english: A study of multilingual abilities and types of llms," in *Conference on Empirical Methods in Natural Language Processing*, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:258947405.

[168] B. Nojavanasghari, T. Baltrusaitis, C. E. Hughes, and L. Morency, "EmoReact: a multimodal approach and dataset for recognizing emotional responses in children," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction, ICMI 2016, Tokyo, Japan, November 12-16, 2016*, ACM, 2016, pp. 137–144. DOI: 10.1145/2993148.2993168. [Online]. Available: https://doi.org/10.1145/2993148.2993168.

[169] M. Abdul-Mageed and L. Ungar, "Emonet: Fine-grained emotion detection with gated recurrent neural networks," in *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers)*, 2017, pp. 718–728.

[170] C. Strapparava and R. Mihalcea, "Semeval-2007 task 14: Affective text," in *Proceedings of the fourth international workshop on semantic evaluations (SemEval-2007)*, 2007, pp. 70–74.

[171] W. Antoun, F. Baly, and H. Hajj, "Arabert: Transformer-based model for arabic language understanding," *arXiv preprint arXiv:2003.00104*, 2020.

[172] Q. Gao, J. Hu, R. Xu, L. Gui, Y. He, K.-F. Wong, and Q. Lu, "Overview of ntcir-13 eca task.," in *NTCIR*, 2017.

[173] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[174] R. Thoppilan, D. De Freitas, J. Hall, N. Shazeer, A. Kulshreshtha, H.-T. Cheng, A. Jin, T. Bos, L. Baker, Y. Du, *et al.*, "Lamda: Language models for dialog applications," *arXiv preprint arXiv:2201.08239*, 2022.

[175] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, *et al.*, "Palm: Scaling language modeling with pathways," *Journal of Machine Learning Research*, vol. 24, no. 240, pp. 1–113, 2023.

[176] R. Anil, A. M. Dai, O. Firat, M. Johnson, D. Lepikhin, A. Passos, S. Shakeri, E. Taropa, P. Bailey, Z. Chen, *et al.*, "Palm 2 technical report," *arXiv preprint arXiv:2305.10403*, 2023.

[177] O. Zheng, M. Abdel-Aty, D. Wang, Z. Wang, and S. Ding, "Chatgpt is on the horizon: Could a large language model be suitable for intelligent traffic safety research and applications?" *arXiv preprint arXiv:2303.05382*, 2023.

[178] R. Li, A. M. Nawi, and M. S. Kang, "Human-machine translation model evaluation based on artificial intelligence translation," *EMITTER International Journal of Engineering Technology*, vol. 11, no. 2, pp. 145–159, 2023.

[179] J. Devlin, “Bert: Pre-training of deep bidirectional transformers for language understanding,” *arXiv preprint arXiv:1810.04805*, 2018.

[180] Y Liu, “Multilingual denoising pre-training for neural machine translation,” *arXiv preprint arXiv:2001.08210*, 2020.

[181] K. K. Maurya, M. S. Desarkar, Y. Kano, and K. Deepshikha, “Zmbart: An unsupervised cross-lingual transfer framework for language generation,” *arXiv preprint arXiv:2106.01597*, 2021.

[182] C. Tran, Y. Tang, X. Li, and J. Gu, “Cross-lingual retrieval for iterative self-supervised training,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 2207–2219, 2020.