



**Arab American University**  
**Faculty of Graduate Studies**

**AraBERT-based Approach to Arabic Cyberbullying  
Detection in Facebook comments**

By

**Rania Ibrahim Hithnawi**

Supervisor

**Dr. Mohammad M. N. Hamarsheh**

**This thesis was submitted in partial fulfillment of the  
requirements for the Master`s degree in Cybersecurity**

**July/ 2024**

**© Arab American University- 2024. All rights  
reserved.**

## Thesis Approval

### **AraBERT-based Approach to Arabic Cyberbullying Detection in Facebook comments**

By

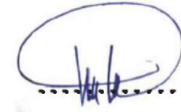
**Rania Ibrahim Hithnawi**

This thesis was defended successfully on 10 / 7 / 2024 and approved by:

Committee members

Signature


1. Dr. Mohammad M.N. Hamarsheh: Supervisor



2. Dr. Mohammad Maree: Internal Examiner



3. Dr. Bashar Tahaynaeh: External Examiner



## Declaration


Rania Ibrahim Hithnawi with a university ID 202112685 declares that this thesis, "AraBERT-based Approach to Arabic Cyberbullying Detection in Facebook Comments" is her own original work and that all informational and materials sources have been appropriately acknowledged.

I confirm that the material, facts, and concepts utilized in this thesis have all been properly cited and recognized in accordance with Arab American University citation guidelines.

I provide permission for this thesis to be archived and made accessible by Arab American University for research and educational purposes. Any additional uses or copies of this thesis must have my permission.

Name of Student: Rania Ibrahim Hithnawi.

University ID: 202112685

Signature: 

Date: 3/02/2025

## **Dedication**

To my parents, who still care for me as I were a child; to my husband, who always supports me; to my sons, the source of my smile; to my brothers and sisters; to all of them I dedicate this thesis work.

Rania Ibrahim Hithnawi

## **Acknowledgment**

I am sincerely grateful to my supervisor Dr. Mohammed Hamarsheh for his help, assistance, suggestions, and patience. I would like to thank the Department of Cyber security for their help and guidance. Thanks to all my friends and colleagues at Arab American University.

## Abstract

Cyberbullying is one of the significant issues with communication platforms like Facebook. It is especially alarming that, in contrast to traditional bullying, it can have serious emotional consequences and follow victims all the time, thus concern regarding cyberbullying is growing on websites like Facebook. Pre-trained language models have achieved significant success in a variety of natural language processing tasks. Although research has been conducted on optimizing BERT-based models for detecting cyberbullying and creation of various pre-trained Arabic models, limited attention has been paid to Arabic cyberbullying detection and low resource available in Arabic language.

This thesis aims to investigate the effectiveness of using AraBERT, a pre-trained Arabic language model, in detecting Arabic cyberbullying comments. We create a balanced dataset out of 20,000 Facebook comments in Arabic language and manually label it as either bullying or non-bullying. We employed fine-tuning techniques to adapt AraBERTv2 to the cyberbullying detection task. Through experimentation with freezing layers' technique and unfreezing different layers of the model, we explored the trade-off between leveraging pre-trained knowledge and adapting the model to the specific task. Our findings demonstrate that the tuning of all layers in AraBERTv2 achieved the highest performance, resulting in an accuracy of 91.9% and an F1 score of 93.0%.

## Table of Contents

Thesis Approval.....	i
Declaration.....	ii
Dedication.....	iii
Acknowledgment.....	iv
Abstract.....	v
Table of Contents .....	vi
List of Tables .....	vii
List of Figures.....	viii
Chapter 1: Introduction.....	1
1.1 Introduction to Cyberbullying .....	1
1.2 Definition of the problem .....	2
1.3 Research objectives.....	5
1.4 Structure of the Thesis .....	6
1.5 Significance of study .....	7
Chapter 2: Literature Review .....	8
2.1 Keyword-based approaches .....	8
2.2 Machine Learning Models.....	9
2.3 Deep Learning Models:.....	10
2.4 Transformer .....	15
2.5 BERT models.....	17
2.6 AraBERT Model.....	21
Chapter 3: Research Methodology .....	25
3.1 Dataset Collection.....	25
3.2 Dataset Preprocessing .....	27
3.3 AraBERT Model Selection and Fine-tuning.....	30
3.3.1 AraBERT Model Setup .....	32
3.3.2 Training and Evaluation .....	33
Chapter 4: Experiments and Results.....	36
4.1 Experimentation process.....	36
4.2 Results and error analysis .....	38
4.3 Discussion.....	41
Chapter 5: Conclusion and Future work.....	43
References .....	44
الملخص.....	51

## List of Tables

Table 1. Cyberbullying Detection Methods in Arabic Text.....	24
Table 2. AraBERT model versions differences.....	31
Table 3. Computational load and training time.....	38
Table 4. Freezing strategies used in the model. ....	38



## List of Figures

Figure 1. Lifetime Cyberbullying Victimization Rates [5].	2
Figure 2. Cyberbullying detection papers [10].	4
Figure 3. Differences in the architecture of pre-training models[14].	14
Figure 4. The architecture of the Transformer model[43].	16
Figure 5. Pre-training and Fine-tuning stages for BERT [50].	19
Figure 6. Methodology.	25
Figure 7. The percentage of bully and not bully comments in the dataset.	27
Figure 8. A sample of the collected dataset.	27
Figure 9. Tokenization forms[69].	29
Figure 10. The proposed architecture using AraBERT.	30
Figure 11. BERT Layers Transferability by Task[81].	37
Figure12. The evaluation matrix of our model results.	39
Figure13. Confusion matrix for the model experiments.	40

## **Chapter 1: Introduction**

### **1.1 Introduction to Cyberbullying**

The home environment, neighborhood social circles, and the structured world of school all have an important role in shaping adolescence, which is an age of major growth for individuals. Bullying, a tragically prevalent type of youth violence, was once limited to the physical environment. It could happen in public spaces such as playgrounds, bus stops, or during school hours or during the journeys to and from the school. However, the explosion of modern technology has transformed bullying, empowers bullies with the ability to harass and intimidate their targets beyond the limitations of physical space by utilizing electronic devices like computers and cell phones, allowing tech-savvy aggressors to relentlessly target their victims day and night, this new form, known as cyberbullying (Patchin and Hinduja, 2006). Children's homes usually have a considerable number of computers, which are used for social, entertainment, educational, and productivity. Furthermore, younger age groups are increasingly using and consuming cell phones since they are seen as status symbols, facilitating communications with peers in various physical settings.

Globally, the number of people using social networking technologies is growing as the field of communication and information technology continues to expand quickly. The percentage of individuals who own smartphone and Internet access continues to grow in Palestine, reaching 88.6% among individuals aged 18 and above (Statistics, 2018). Social networks are online communication platforms that allow interaction through the sharing of textual, visual, and video content. This fosters both connection and conflict, as diverse individuals with varying ideas engage with each other (Boyd and Ellison, 2007). Expressing diverse ideas can have positive impacts like fostering acceptance, tolerance, and learning, but if handled improperly, it can lead to negative outcomes like bullying, intolerance, and harassment. Social networks, especially anonymous platforms with the Misunderstanding in the concept of freedom in speech, have the potential for escalating these issues by amplifying hate speech and targeting victims (Cortis and Handschuh, 2015). Hence, it is important to take seriously this exploitation of social media for malicious purposes.

## 1.2 Definition of the problem

Cyberbullying victimization is a noticeably big problem that has been increasing steadily over the years of studying this phenomenon, as depicted in Figure 1. It affects both victims and societies (Patchin, J. W. & Hinduja, 2023). Accordingly, action is needed from all major stakeholders, including parents, teenagers, kids, school administrators, and other responsible individuals in positions of influence and responsibility. Often, the best approaches to reduce and even eliminate the widespread problem of cyberbullying in our culture are public education and intervention. All parents, kids, and teenagers need to be aware of cyberbullying, its costs and implications, and the people who are most likely to participate in it before they become as a victim (Mitsu and Dawood, 2022). Many attempts have been offered toward prevention, detection or mitigation of cyberbullying.



Figure 1. Lifetime Cyberbullying Victimization Rates (Patchin, J. W. & Hinduja, 2023).

The benefits of being online are vast. It does come with some risks, which you have to guard against, just like many other things in life. You might want to remove specific applications or take a temporary break from the internet if you've been the victim of cyberbullying so you have time to heal. However, turning off the Internet is not a permanent fix. Since you did nothing wrong, you should not suffer any consequences, and it can even give the bullies the incorrect message, supporting their inappropriate actions (unicef, 2023).

Facebook occupies a dominant position in the social media landscape, it is arguably the most popular social media platform worldwide with 2.74 billion monthly active users (Dreamgrow, 2024). Facebook and Instagram are using tools for managing online interactions, they actively develop technology to encourage positive interactions and remove harmful content by AI-enabled prompts that alert users to reconsider potentially offensive captions or comments before posting. Other tools are also used like an automatic comment filters that hide content containing offensive words, phrases, or emojis. In addition, custom word lists are used to allow users to filter specific unwanted terms from comments and messages and blocking, muting, and restricting accounts to avoid unwanted interactions with specific users (Meta, 2023). Different tools are available on each social media platform to limit who can view or comment on your posts, who can friend you automatically, and how you can report bullying incidents. Many of them have easy actions to report, mute, or restrict cyberbullying, but it is recommended to implement an automated technique to detect or eliminate cyberbullying on social media.

Managing cyberbullying on social media can be extremely difficult because it may take many different forms, including the use of images, harmful comments, videos, and more. Here in our thesis, it is limited to detect the act of bullying via Arabic language text. Even AI technology is designed to automatically detect and remove cyberbullying, the detection of these bullying comments is still a hard task, as their classification is very subjective. Certain comments made by friends informally may appear to be bullying, even though they are not.

Many efforts for developing and implementing a cyberbullying detection model using Machine Learning (ML) and Deep Learning (DL) algorithms, which enable the identification or accurate classification of new relevant instances after training them

with enough data, has become one of the significant approaches to automate the detection of cyberbullying. More and more data has lately been available to enable the development of such systems which works better with large data sets. A small amount of data is available for low-resource languages like Arabic ,Bengali, Hindi, or Urdu; the majority of the data is available for high-resource languages like English or German (Mahmud *et al.*, 2023) as it is shown in figure 2. Although more than 500 million in 21 countries speaking Arabic language in the world (Boudad *et al.*, 2018), there are limited researches in the field of natural processing language because the morphology of Arabic language is rich and more complex compared to English and the presence of different dialects increases the complexity.

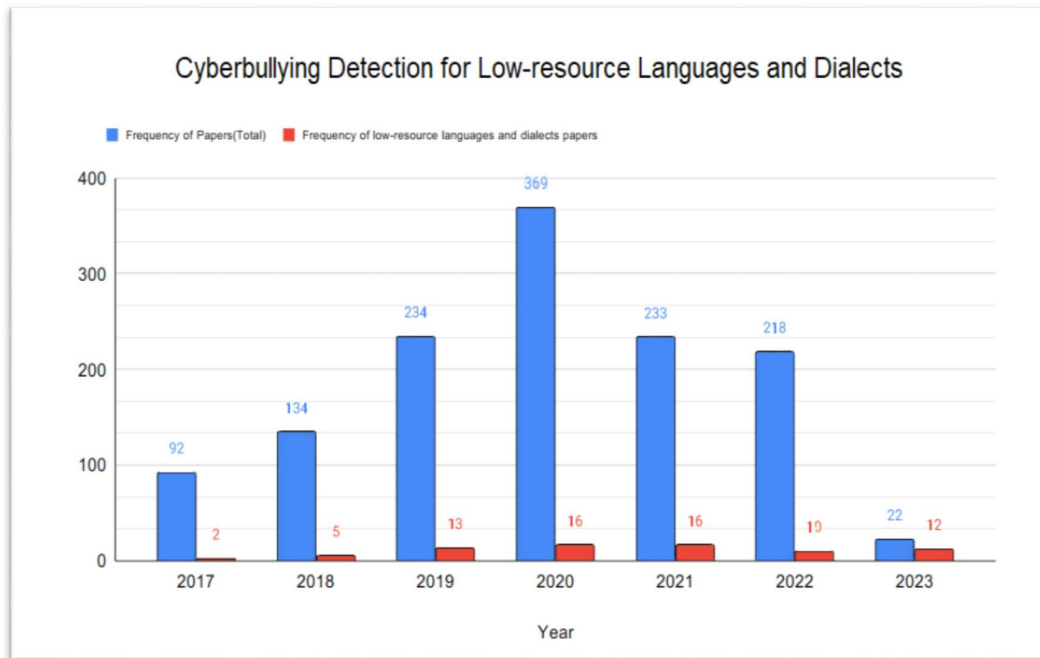


Figure 2. Cyberbullying detection papers (Mahmud *et al.*, 2023).

Sentiment analysis studies to identify cyberbullying in Arabic online environments is impeded by several challenges. First, it is difficult to build an effective foundation for detection algorithms due to a major lack of research specifically focused to sentiment analysis of Arabic text. This is made more difficult by the inherent complexity of Arabic. Sentiment analysis technologies have significant obstacles due to the

complicated word structure of the language and the diversity of dialects. Furthermore, it is more difficult to detect cyberbullying on Arabic social media platforms due to a lack of well-developed pre-processing techniques made specifically for Arabic content and a general lack of resources. Arabic's specific language characteristics make it difficult to use even promising machine learning approaches, such as deep learning algorithms. Building reliable algorithms that can automatically identify cyberbullying in Arabic online environments requires addressing these challenges. The recent studies on Arabic cyberbullying detection use available datasets that are collected from Twitter mostly and YouTube platform (Mahmud *et al.*, 2023).

### **1.3 Research objectives**

Recently, research work focused on using transfer learning and fine-tuning for large pre-trained language models in order improve performance of NLP tasks with small datasets. In our thesis, we explore the effectiveness of using the Arabic bidirectional encoder representation transformer (AraBERT) based on BERT model for detecting Arabic cyberbullying comments on Facebook.

Arabic language is a morphologically rich language with few resources' datasets and a less explored syntax compared to English language. However, Arabic Natural Language Processing (NLP) tasks as Sentiment Analysis and Named Entity Recognition have proven to be very challenging to tackle (Antoun, Baly and Hajj, 2020). With the advancement of transformer-based models in NLP, the pre-training of language-specific models, such as BERT, has shown to be very effective in achieving high performance on various NLP tasks. By pre-training BERT specifically for the Arabic language, researchers aim to overcome the challenges faced by Arabic NLP and achieve similar success to what BERT has achieved for the English language. The newly developed model, AraBERT, has shown to achieve state-of-the-art results on various Arabic NLP tasks, such as sentiment analysis, named entity recognition, and question answering, compared to multilingual BERT models and other state-of-the-art approaches (WissamAntoun, 2022). We employed AraBERT, a variant of the BERT model, as our approach. It is a popular model for a variety of NLP tasks across a number of languages. AraBERT is an Arabic language pre-trained model developed on the Google BERT architecture(Antoun, Baly and Hajj, 2020). There are versions for

AraBERT: AraBERT v0.2/v2, Large and base versions with more training data, larger vocabulary, and more data and the other versions AraBERT v0.1/v1, the original (WissamAntoun, 2022) (El Moubtahij, Abdelali and Tazi, 2022).

This research contributes to the field of cyberbullying detection by creating a large balanced dataset of Arabic Facebook comments. The dataset is labeled as bullying and non-bullying content, it was collected from popular pages that have large number of followers. The previous researches used datasets from Twitter and YouTube in many, in our work we gather Facebook comment using Apify tool to create Arabic bullying and not bullying dataset, because Facebook is the most popular social media platform that are accessed by user (Dreamgrow, 2024). Our balanced dataset contains 20K Arabic comments and labeled manually as bully (1) and not bully (0).

The research aims to contribute to the field of cyberbullying detection by:

- Building an Arabic dataset contains Facebook comments.
- Developing and evaluating AraBERT-based models for cyberbullying detection on Facebook comments.
- Investigating the impact of fine-tuning AraBERT model on the created dataset to achieve and improve the performance of the model.
- Evaluating the effectiveness of the proposed models on real-world.

#### **1.4 Structure of the Thesis**

The thesis begins by presenting the literature reviews and explores existing research on Arabic cyberbullying detection automatically. There are many approaches used to detect Arabic cyberbullying, they begin using Keyword-based approaches, Machine learning, Deep learning and recently BERT transformer and using AraBERT. Moving to the methodology section stating with creating dataset that contains bully and not bully comments of the size 20K from Facebook pages, it prepared as balanced dataset and labeled manually, in addition to preprocessing and tokenization process followed to choose the AraBERT model v2 to implement our model. The next chapter explores the experimentations and results evaluation. We use multiple strategies for freezing layers in the model to get the best performance and achieve high accuracy in the evaluation process, followed by the discussion stage to analyst our obtained results.

Finally, the conclusion that have been achieved so far in the research and the suggested planning of future work.

### **1.5 Significance of study**

In this study, we identify and outline the key components and procedures of choosing layers and fine tuning the Arabic bidirectional encoder representations transformer (AraBERT) model. The second significant aspect of our study, is collecting a 20K-record dataset of Facebook comments using online tools that scrape the comments and saved it as CSV file before it has been cleaned and labeled manually. The manual work result in precise classification of the comments into bully and not bully categories. Hence, we evaluated our model on the dataset, the experiments show that our approach achieves very good results and high accuracy score.



## Chapter 2: Literature Review

The goal of artificial intelligence (AI) is to develop machines that can emulate human thought and behavior. AI systems are very competent at specific tasks, such as medical diagnosis, search engines, and chatbots, even if they are not yet able to completely imitate human intelligence. Even with the enormous achievements made thus far, realizing full, global human intelligence is still a difficult task. In order to create intelligent machines that are capable of performing tasks that typically require human intelligence, such as speech recognition, visual perception, decision-making, and language translation, physiological researchers have simulated the various neural network models extracted from the human brain processing systems (Copeland, 2024). Overall, the structure and function of the brain have prompted many of the AI techniques, and the study of the brain and its functioning has had a major effect on the development of AI.

The use of artificial intelligence (AI) in fighting against cyberbullying has become more efficient. By employing its ability for a large-scale data analysis, pattern recognition, and learning through. This literature review explores existing research on cyberbullying detection automatically, with a focus on utilizing AraBERT, a pre-trained Arabic language model, for analyzing comments.

### 2.1 Keyword-based approaches

These methods identify cyberbullying through predefined dictionaries of offensive keywords and phrases. The study in (Bashir and Bouguessa, 2021), investigate questions of how to protect Arabic text from cyberbullying based on specific swear words that are posted on Twitter. they collect a data from twitter with some cyberbullying words, using an application programming interface (API) and use several learning approaches to produce the results. Several learning approaches are investigated, the results according to their experiments suggest using models like LSTM (Long Short-Term Memory) with accuracy of 72% compared to other traditional classifiers.

The authors in (Mouheb *et al.*, 2019) produce a scheme for detection cyberbullying in Arabic comments using machine learning based on Naïve Bayes (NB) algorithm for classification, by collecting dataset from Twitter and YouTube, the

accuracy was 0.959. The study suggests to perform the classification on the comments in the dataset that is collected not only bullying keywords but using different algorithms of data mining and not only Naïve Bayes.

While the keyword-based approaches are simple and efficient, their lack of contextual understanding can cause them to miss subtle forms of cyberbullying. Moreover, they require huge manual efforts for constructing the dictionaries and refining them continually with updated terms in the context.

## 2.2 Machine Learning Models

Computational algorithms designed with the purpose of learning from a specific data set are known as machine learning algorithms (Nasteski, 2017). These algorithms are typically used in two common types, the first one is supervised learning (classification) that used to identify a certain target or group of targets which called Labels that exist in the dataset. The other is Unsupervised Learning (clustering) used to identify and group together data samples that are similar to one another, here the Labels are not exist (Liu, 2011).

Utilizing supervised learning algorithms, models learn to classify comments as cyberbullying or non-cyberbullying based on features extracted from text content. The authors in paper (Alsubait and Alfageh, 2021) make comparison in performance of three machine learning models by using Multinomial Naïve Bayes(MNB) algorithm, Complement Naïve Bays( CNB) algorithm and Linear Regression (LR) algorithm, with two feature extraction methods: Count Vectorizer and Tfidf Vectorizer. The authors used a labeled dataset that contains 15,000 of YouTube comments in Arabic publicly available, the comments are labeled as positive for bullying and negative for non-bullying comments. The results indicate that when employing the Count Vectorizer for feature extraction, the Logistic Regression model outperforms both the Multinomial and Complement Naïve Bayes models. However, when using the Tfidf Vectorizer for feature extraction, the Complement Naïve Bayes model surpasses the other two models.

On the other hand, paper (AlHarbi *et al.*, 2019) explored automated cyberbullying detection using sentiment analysis and lexicon approaches. They implemented the experiment in Java, utilizing datasets gathered from Twitter API, Microsoft Flow, and YouTube comments. These diverse sources were combined into a single file exceeding 100,000 tweets and comments. The data was classified into "bullying" and "non-

bullying" categories. The odd number of reviewers ensured a definitive classification based on majority agreement. Following data cleaning and pre-processing, they employed three lexicon generation methods: Pointwise Mutual Information (PMI), Chi-square, and Entropy. The results revealed that the PMI approach produces the best performance in cyberbullying detection in comparison with both Chi-square and Entropy. This suggests its potential for further development and integration into automated cyberbullying detection systems.

Early studies, such as the work presented in (Alakrot, Murray and Nikolov, 2018b) produced a dataset of 15,050 Arabic YouTube comments. The authors evaluated the impact of various features and pre-processing techniques on the performance of an SVM classifier for identifying offensive comments. The results show that data pre-processing with stemming significantly improves offensive language detection in casual Arabic social media text. N-gram features also enhance the classifier's performance. Combining stemming and N-grams, however, proves counterproductive, reducing both precision and recall. These results suggest that stemming and N-grams offer valuable tools for detecting offensive language in Arabic online communication, but their optimal combination requires further investigation.

In (Alakrot, Murray and Nikolov, 2018a), the authors designed a dataset in Arabic for offensive word from YouTube comments to be used for detection in machine learning modeling, the dataset is labeled and taking into account the diversity of offensive languages and Arabic dialect in the Arab world. On the other hand the authors in paper (Muneer and Fati, 2020) attempted to compile a dataset of 37,373 tweets, and used seven machine learning classifiers: Support vector Machine, Logistic Regression, Random Forest, Light Gradient Boosting Machine, Stochastic Gradient Descent, AdaBoostm and Naïve Bayes. they are evaluated using accuracy, precision, recall and F1 score for performance. The results showed that Logistic Regression achieved the best of 0.928 F1 score compared with other classifiers and Stochastic Gradient Descent achieved 0.968 precision and Support Vector Machine achieved 1.00 recall score. However, machine learning algorithms require large amounts of labeled data and may struggle with the nuances of language.

### **2.3 Deep Learning Models:**

Researchers from various backgrounds are more motivated to conduct research in deep learning, enhanced and hybrid deep learning algorithms have been developed, applied, and published in reputable journals, exhibiting exceptional results when compared with traditional techniques. CNN, Long Short-Term Memory (LSTM), BERT and other Deep Learning approaches have produced novel records of success in resolving a wide range of challenging in computer science issues (Teoh *et al.*, 2021),(Yang *et al.*, 2020). These three techniques are also getting more attention than others in NLP classification of texts analysis due to their extremely accurate results, Text classification is the process of classifying text documents into fixed number of predefined classes. The application of text classification includes spam filtering, email routing, sentiment analysis, language identification (Zhang *et al.*, 2020).

Leveraging advanced neural network architectures, deep learning models achieve good performance in cyberbullying detection. Paper (Iwendi *et al.*, 2023) presents a novel LSTM-based approach for detecting cyberbullying in social commentary. Compared to traditional LSTMs, this article utilizes doubled input gates, output gates, and forget gates, leading to improved accuracy. However, it incurs higher computational complexity and cost. They address this trade-off by analyzing the effectiveness of various deep learning models for insult detection. For empirical analysis, four models were evaluated: RNN, LSTM, GRU, and BLSTM. After applying text cleaning, tokenization, stemming, and lemmatization, the data was fed into the models for prediction. As the results, BLSTM achieved superior accuracy of 82.18% and F1-measure of 88% scores compared to the other models, demonstrating its effectiveness for cyberbullying detection. Future direction to integrate deep learning approach with automatic detection and object identification mechanisms. These models can learn complex patterns and relationships within textual data, enabling more accurate classification. Paper (Alzaqebah *et al.*, 2023) produced an automated framework for detection cyberbullying imbalanced datasets contain short Arabic text in different dialects, three datasets available from other research 15,049 record dataset 4000 record and 2000 record collected from YouTube and Twitter. Support Vector Machine and deep learning algorithms like (LSTM) and (Bi-LSTM) are used for evaluation process, they used accuracy, specificity, recall, sensitivity and mean squared error to improve the performance. Bi-LSTM outperforms the classification of cyberbullying detection method

Convolutional Neural Network (CNN) was used by the authors in paper (Rachid, Azza and Ben Ghezala, 2020). They explored the effectiveness of various deep learning models for classifying cyberbullying comments in Arabic news channels. They tested 34 models, that are using CNN, LSTM (Long Short Term Memory), GRU (Gated Recurrent Unit) and combination of them, on both balanced and unbalanced versions of a deleted comments dataset, with several key findings: Combined CNN/LSTM/GRU models with pre-trained Arabic embedding (AraVec and Fasttext) achieved the best performance, reaching an impressive F1-score of 84% on the balanced dataset. Based on a single text feature, the simulation results CNN obtained for Tweet sentiment analysis, which includes 377,616 geotagged tweets, has the maximum accuracy of 66.0%.

On the other hand, a combination of text and count of neighboring location categories features yields the maximum accuracy of 78.0%. The accuracy has been raised to 74% and 78% by employing a pre-trained 6B GloVe model with CNN. For the same dataset, it is noticed that the CNN employing the pre-trained 27B GloVe model achieved the best accuracy, at 83.9% and 94%, respectively(Lim, Ho and Ting, 2020). The ensemble model, a hybrid CNN and LSTM version that was also successfully used by (Heikal, Torki and El-Makky, 2018) for the Arabic Twitter dataset classification, achieves an F1-score of 64.46%, which is higher than the F1-score of 53.6% for the state-of-the-art deep learning model. This highlights the potential of combining different neural network architectures and utilizing domain-specific embedding for accurate cyberbullying detection in Arabic text.

A comparison of ensemble CNN-SVM and BERT Models (Saini *et al.*, 2023) ensemble CNN-SVM model combined two techniques, Convolutional Neural Network (CNN) that Extracted relevant features from text data and Support Vector Machine (SVM) that Classified text as cyberbullying based on the extracted features. It Achieved an accuracy of 96.88% in identifying cyberbullying. The second model was BERT-based model, it Leverages the pre-trained BERT (Bidirectional Encoder Representations from Transformers) language model, BERT excels at understanding the context and nuances of language, making it well-suited for detecting subtle forms of cyberbullying. That demonstrated superior performance with an accuracy of 97.34% compared to the Ensemble CNN-SVM model. That means BERT outperforms the Ensemble CNN-SVM model. For future research the authors suggest exploring ways to improve detection

accuracy for low-resource languages and develop methods for mitigating real-world cyberbullying based on these models. In other research the authors (Teng and Varathan, 2023) compared two approaches for automatic cyberbullying detection, Conventional Machine Learning (CML) that employed a variety of features, including textual, sentiment, emotional, word embedding, and novel "toxicity" features also used the latest psycholinguistic tools and achieved an F-measure of 64.8% with Logistic Regression the other approach was Transfer Learning that used as Fine-tuned pre-trained language models (DistilBert, DistilRoBerta, Electra-small) for faster training and better performance, Fine-tuned DistilBert achieved the highest F-measure of 72.42%, surpassing CML. As the results, Transfer Learning outperformed CML with higher accuracy, less effort (no feature engineering needed) and faster in the training computation. The authors suggest for future work to investigate ways to improve accuracy for low-resource languages and diverse types of cyberbullying. In this thesis, we adopted using Bert Transformer model, that has become a popular choice for natural language processing tasks. Its ability to understand context, its efficiency, and its lower data requirements make it a strong candidate for many applications. These advantages are particularly valuable when dealing with large datasets or when time is a constraint.

BERT is bidirectional encoder representations from transformers, was developed in late 2018 at Google AI research centers. It outperforms its predecessors in terms of learning speed and outcomes. Additionally, after receiving pre-training in an unsupervised way, it generates its own linguistic "representation." In order to fast and with little data, it can be trained continuously (in a supervised manner). Its advantage over competitors such as OpenAI's generative pre-trained transformer (GPT) and embedding from language models (ELMo) [12], is that it is bi-directional; unlike OpenAI GPT, it is not required to look only backwards or concatenate the independently driven "front" and "back" views, as is the case with ELMo(Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, 2018) as shown in figure 3. There are numerous benefits to using pre-trained models. Initially, they significantly reduce the amount of computing power needed for training, additionally, you could use this by utilizing the power of advanced models without having to spend time training one from scratch. A large library of pre-trained models, each capable of handling a wide range of tasks, is available on platforms like Transformers. Using your own dataset to fine-tune the pre-trained model to your particular task is how the magic is performed. The

enormous potential of pre-trained models is unlocked by this fine-tuning technique (HuggingFace, no date).

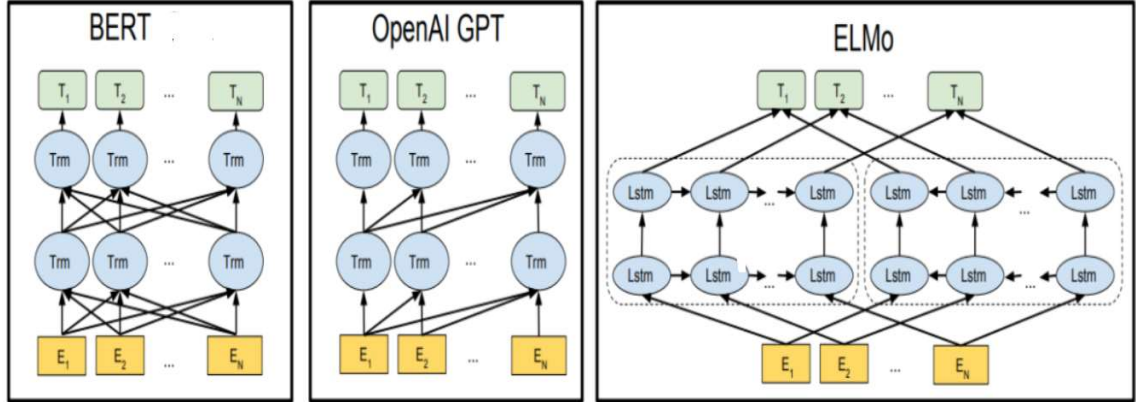


Figure 3. Differences in the architecture of pre-training models(El Moubtahij, Abdelali and Tazi, 2022).

A set of methods that are able at explaining a wide variety of Natural Language tasks, including Automatic Translation(Och and Ney, 2004), Text Summarization(Rush and Weston, 2015), and Text Generation(Yu *et al.*, 2017), is the goal of the Natural Language Processing (NLP) domain. The technique of successfully extracting meaning involves all those tasks. It goes without saying that most current NLP issues would be resolved if a method understood the basic semantics of texts. A major issue limiting the general Natural Language Processing (NLP) is the single-task training approach. Collecting information and developing a customized model for solving a specific problem is successful. On the other hand, it forces us to apply the concept to a different area in addition to having to solve problems every time a new one emerges. To save time, it could be better to use a general multitask solver instead of this one.

In the past, recurrent neural networks (RNNs) dominated the field of natural language processing (NLP), especially when it involved supervised learning tasks like regression and classification. The Long Short Term Memory (LSTM)(Sepp Hochreiter, 1997) and Gated Recurrent Unit (GRU)(Kyunghyun ChoDzmitry BahdanauFethi Bougares Holger Schwenk, 2014) architectures, which addressed the widely recognized decreasing gradient issue that struck previous RNNs, accelerated the technology to importance. This problem emerged because long-term dependencies in sentences are

hard to learn since information deteriorates during training as it passes back through the network. RNNs can handle longer sequences since LSTMs and GRUs introduced gates that selectively retained relevant data. RNNs have some limitations even if they showed incredible adaptability. Training speed and scalability were restricted by their inability to parallelize on many processors due to their recurrent structure (Greff *et al.*, 2017). Furthermore, long words may still have the vanishing gradient problem despite the improvements of LSTM and GRU.

## **2.4 Transformer**

Presently, the Transformer—a cutting-edge development was presented by Vaswani et al. in 2017—which eliminated RNNs as the dominant NLP technology (Vaswani Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, 2017). In contrast to its predecessors, the Transformer only uses an attention mechanism instead of the recurrent structure. This approach solves the vanishing gradient problem and makes it possible for the model to analyze even the longest sentences efficiently by allowing it to focus on relevant portions of the input sequence at each step. Additionally, the Transformer's architecture can be parallelized through structure, which makes training much faster than RNNs, as shown in figure 4.



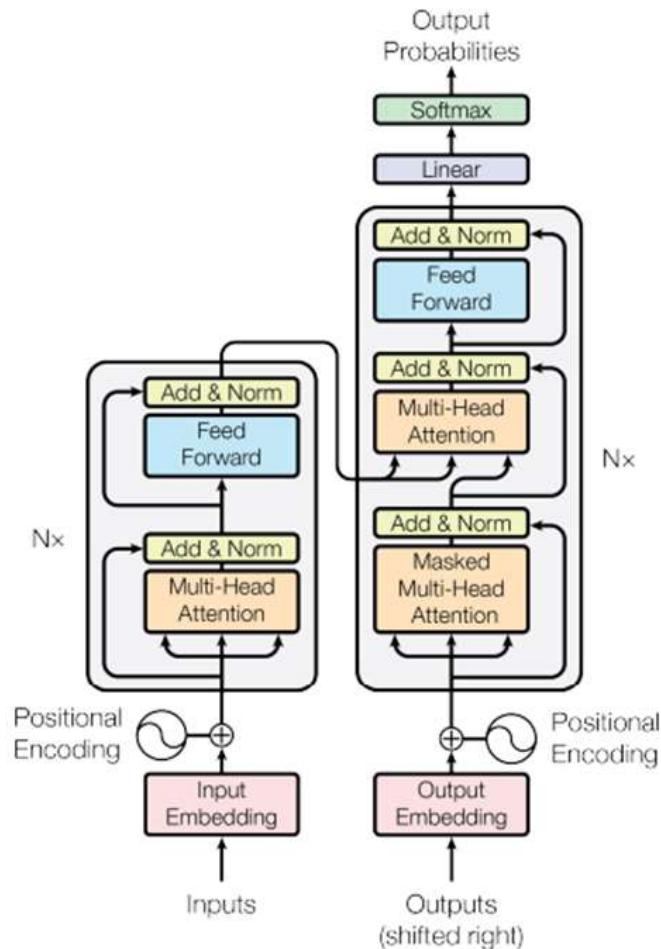


Figure 4. The architecture of the Transformer model (Vaswani Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, 2017).

On the other hand, in the auto-regressive models, the next results are generated using the previous outputs. It is basically a unidirectional network since it can only access the left context of the token that is being evaluated. This feature, combined with the Transformer architecture, is being used by several models for developing accurate Language Model languages like GPT and GPT-2. These models employ a two-stage learning approach. It uses unsupervised language modeling on large text corpora during the first pre-training stage to build an extensive understanding of language structure and semantics. Then, supervised fine-tuning improves the model's performance for the intended use by focusing its expertise on particular downstream tasks (Radford and Salimans, 2018), GPT-2 model seeks, without additional training, to duplicate GPT's performance on a range of tasks, it aims to achieve this, in contrast to GPT, by extracting more subtle information from a larger amount of text data, this needs a model

with more parameters and greater complexity. Without more training, it performs competitively on certain tasks like reading comprehension, but not well enough in other areas, like summarization (Radford *et al.*, 2019).

Transformer networks currently perform reasoning tasks in natural language processing, they have outperformed humans on several GLUE (General Language Understanding Evaluation) benchmarks. They efficiently handle several tasks at once and exceed RNNs in speed and accuracy. The authors in (Alshamsi, 2023) presented a transformer-based approach for detecting offensive speech in Arabic tweets, which is validated using a mixture of four benchmark Twitter Arabic datasets annotated for hate speech detection task, including the (OSACT5 2022) workshop shared task dataset . This model employs versions of the CAMELBERT model. Transformer-based model achieved 87.15% accuracy in detecting offensive speech in Arabic tweets. Model achieved 83.6% F1 score in detecting offensive speech. On the other hand, paper (Paula *et al.*, 2023) discussed hate speech detection in Arabic languages using transformer models and ensemble approaches. They tested six transformer models and combined them using two different ensemble approaches. As a result, combining models through majority voting led to the best performance on the training data, achieving an F1-score of 0.60 and an accuracy of 86%. Although the results are motivating, hate speech detection in Arabic text still has potential for improvement. Despite progress of Transformer , low-power device operation and human-level understanding of texts remain challenging tasks(Gillioz *et al.*, 2020).

## 2.5 BERT models

Text classification is the main process in machine learning for natural language processing to classify the text into one or more categories according to their content, text classification needed in spam filtering ,sentiment analysis and email routing applications (Vijayan, Bindu and Parameswaran, 2017). It has achieved high performance but still need more improvement especially when using the Arabic language.

A new language representation model that stands for Bidirectional Encoder Representations from Transformers. Unlike previous models that process text in a single direction, BERT captures deep bidirectional representations by simultaneously

considering both left and right context within every layer. This innovation is achieved through pre-training on vast amounts of unlabeled text, enabling BERT to develop a comprehensive understanding of language patterns and relationships (Devlin *et al.*, 2019).

BERT was introduced in 2018, the beginning of contextual information concept, and uses a transformer structure with many encoded layers. It has demonstrated improvements on a variety of natural language processing tasks, including classification, inference, and semantic comprehension. Transformer structures analyze words efficiently, quickly and simultaneously, allowing for improved word context learning because they can understand context concurrently from both sides. The encoder and decoder layers are the two main parts of the transformer. Using transformer components, the Google AI developer worked on the feed-forward neural network (FFNN) and self-attention encoder layers in 2018. The result was a proposal for BERT, or bidirectional encoder representations from transformers. BERT is designed to pre-train deep bidirectional representations from an unlabeled text by concurrently conditioning the left and right context in all levels (Devlin *et al.*, 2019).

Two stages should be involved in BERT training. The model gains language and context understanding during the first step of pre-training, and problem-solving skills during the second phase of fine-tuning, show figure 5. Neural machine translation, question answering, sentiment analysis, summarization, and other problems are among the numerous problems that BERT can solve. Furthermore, BERT produced cutting-edge results in over 11 NLP tasks.

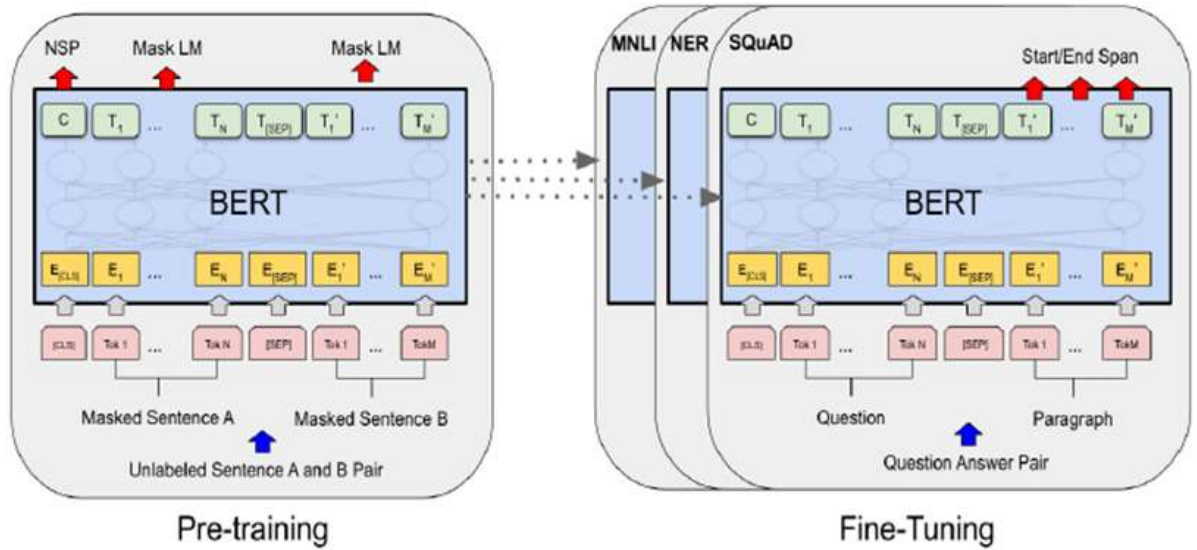


Figure 5. Pre-training and Fine-tuning stages for BERT (Devlin *et al.*, 2019).

Several pre-trained BERT models are available to support the Arabic languages. Devlin and his team, the original developer of BERT model, in paper (Devlin *et al.*, 2019) developed a single model Multilingual BERT, supporting over 100 languages, including Arabic. A lot of researches used the Multilingual BERT model (Alammary, 2022). Paper (Alshalan and Al-Khalifa, 2020) worked to investigate some neural network models to detect hate tweet speech in Arabic, a publicly available dataset of 9,316 labeled tweets categorized as hateful, abusive, and normal. This dataset is used to evaluate and compare four different sentiment analysis models: CNN, GRU, CNN-GRU hybrid, and Multilingual BERT model. The results demonstrated the effectiveness of these models in accurately detecting hateful and abusive language, CNN model outperformed the other models, achieving an impressive F1-score of 0.79 and an AUROC (Area Under the Receiver Operating Characteristic Curve) of 0.89, AUROC is a popular metric used in machine learning to assess the performance of binary classification models. This suggests that CNNs may be particularly well-suited for capturing the linguistic patterns and features relevant to hate speech and abuse detection. Despite the strong performance of BERT model in other natural language processing tasks, it did not outperform the baseline models. This could be due to its training on a fundamentally different genre of data (Wikipedia) compared to the informal and often aggressive language found in tweets. This highlights the importance of domain-specific training data for achieving optimal performance in hate speech detection tasks.

On the other hand, the authors in paper (Pelicon *et al.*, 2021) explored the effectiveness of cross-lingual training for creating offensive speech detection models across multiple languages. It showed that models based on language models trained on smaller language sets outperformed general multilingual models like mBERT for most languages tested (Croatian, Slovenian, English, German, Arabic). When target language data is limited, intermediate training with other languages (except German in this study) improved model performance considerably. This benefit diminished with large target language data. Qualitative analysis suggests vocabulary coverage may not be the sole factor influencing performance. Models struggle with instances requiring context or socio-political knowledge. For future research, broader analysis on diverse languages and NLP tasks, exploring language models beyond BERT like XLM (Extensible Language Model) which are designed to handle multiple languages simultaneously, are desired approach. Paper (Alsafari, Sadaoui and Mouhoub, 2020) Developed a high-quality Arabic corpus of hate and offensive speech, annotated with three tasks: clean/hate, clean/hate/offensive, and clean/offensive/protected hate characteristics, conducted an extensive analysis of various feature selection and classification methods for hate and offensive speech detection (2, 3, and 6 class tasks) also the paper demonstrated the effectiveness of deep learning models (CNN+mBERT) over traditional machine learning models (SVM) for all classification tasks, with higher accuracy (up to 87.05%). The results were that word embedding, especially contextual embedding, was efficient with deep learning models not with machine learning models. The model's performance in the 2-class task aligns with existing Arabic hate and offensive language detection models, demonstrating its effectiveness in this domain. While results for the 3-class task fall within the previously reported range, it's important to note the unique challenges of multi-class classification and the impact of limited data for each sub-category. Also, the model outperforms existing English hate speech detection models in the 2-class task. This suggests potential for Arabic-specific models to achieve higher accuracy in this context. For example, compared to a Bert-based CNN achieving 82.9% F-score for English offensive language, our model reached a similar score. Multilingual BERT model offers broad coverage but may not be as specialized for each language. Antoun et al. created a dedicated Arabic model called AraBERT [45] trained on 24 GB of text, this provides a deeper understanding of Arabic nuances. The authors in paper [46] leveraged 1 billion tweets to train their Arabic BERT model

MARBERT, this leverages the richness of social media data for a potentially more conversational and informal understanding of the language.

## 2.6 AraBERT Model

AraBERT, a pre-trained language model specifically designed for Arabic. It possesses a deep understanding of the Arabic language and can capture subtle linguistic nuances often missed by generic models. The fact that AraBERT was trained on a large amount of Arabic text indicates that it has an excellent understanding of the language and is capable of high-accuracy language translation, classification of text, and sentiment analysis, which are all potential benefits of applying it. AraBERT may be fine-tuned for specific applications with relatively little extra training data since it has already been pre-trained on a large dataset. This reduces the time and resources needed to create and deploy natural language processing models. AraBERT is a useful tool for tasks involving Arabic text because it was created especially for Arabic language processing. Since AraBERT is pre-trained, it can be quickly and simply included into current pipelines for natural language processing, allowing it to be used for a wide range of purposes. Transfer learning—where knowledge from one task can be applied to another related activity—is made possible by the ability to fine-tune pre-trained models like AraBERT. This reduces the amount of training data needed while also enhancing performance (Antoun, Baly and Hajj, 2020) (Faraj, 2022) (Nada *et al.*, 2020).

Several studies were introduced using AraBERT, we quote them in the following, in 2023 The study (Sulaiman Aftan, 2023) examines the use of Arabic tweet analysis to forecast customer sentiment and feedback for Saudi Arabian telecom companies. It is inspired by the way the human brain processes information and makes use of Deep Learning methods, specifically the AraBERT model. Due to its various language features and limited resources, Arabic text analysis is more difficult than English text analysis that was the main challenge. So AraBERT, a model trained especially for Arabic language processing, is used in this study. CNN and RNN, two well-known Deep Learning models, are used to compare AraBERT's performance. The results indicated that AraBERT predicts customer sentiment with the highest accuracy, particularly for the Mobily and Saudi Telecom Company datasets. Apart from others, in 2022, the strategy used in (El Moubtahij, Abdelali and Tazi, 2022) was based on the AraBERT language model. The first step involved the model being input into a

pre-processed text from the ARev database. Farasa segmentation is then used to develop version 1 of the AraBERT model. Furthermore, the ARev dataset—which has over 40,000 comments and reviews—is the foundation of our evaluation. they achieved an accuracy value of 92.5% using the AraBERT model's well-tuned parameters, which is a very competitive result. the future efforts will focus on resolving the Arabic text segmentation issue and improving the Farasa segmentation version.

On the other hand, the authors in (Salomon, Palé & Kechaou, Zied & Wali, 2022) Emphasize the need for automatic detection of toxic contents, a Tunisian dataset are developed and proposed a model based on AraBERT .the experimental results show that the AraBERT model performed well and achieved an F1 score of 0.99. the authors suggest for future work to generalize the AraBERT model to other Arabic speaking countries and dialects.

the research (El-Alami, Ouatik El Alaoui and En Nahnahi, 2022) tackles the challenging task of Multilingual Offensive Language Detection by leveraging the power of transfer learning and fine-tuning. they propose a novel approach featuring the highly successful Bidirectional Encoder Representations from Transformers (BERT) model, known for its exceptional ability to capture both the meaning and context of text. The system operates in three stages: Preprocessing for Cleaning and preparing the text data for analysis, Text Representation for Employing BERT models to convert text into meaningful numerical representations, capturing the intricacies of language, Classification for Categorizing texts as either offensive or non-offensive. To address the multilingual nature of online communication, the authors explore two distinct techniques, joint-multilingual by Developing a single classification system capable of handling multiple languages simultaneously. Translation-based by Translating all texts into a common language before performing classification. The experiments utilize a bilingual dataset derived from the Semi-Supervised Offensive Language Identification Dataset. The results are promising, revealing that the translation-based method combined with Arabic BERT (AraBERT) achieves an impressive F1-score of over 93% and an accuracy exceeding 91%. That demonstrates the effectiveness of combining BERT with transfer learning and fine-tuning for accurately detecting offensive language across languages. The translation-based approach, particularly successful with AraBERT, opens exciting possibilities for developing robust multilingual offensive language detection models.

Paper (Wadhawan, 2021) tackles the challenge of detecting sarcasm and sentiment in Arabic tweets (EACL WANLP-2021 Shared Task 2). The stages for Their approach: Data Preprocessing, ArSarcasm-v2 dataset is modified by altering text to optimize for model training. Transformer Models that Multiple variants of AraELECTRA and AraBERT, pre-trained on Arabic text, are experimented with. Final Model Selection, the best-performing model variant is chosen for final Sarcasm and Sentiment Detection submissions. Their approach achieved promising 7th and 4th ranks on the private leaderboard for Sarcasm and Sentiment Detection respectively. Exploring other relevant features and building ensemble models combining word-vector and transformer-based embedding for further improvement that was future work suggestion.

Paper (Aldjanabi *et al.*, 2021) tackles the automated detection of hate speech and abusive content in Arabic tweets. Their Innovations Were Multitask learning (MTL) model that Employed two pre-trained Arabic language models (AraBERT and MarBERT) to handle both Modern Standard Arabic (MSA) and dialectal Arabic (DA). Also, Task-specific training that Adapted the model to specific datasets for offensive and hate speech detection. As a result, MTL model outperforms single-task models in classification accuracy across three out of four datasets.

Another research in (Bashmal and Alzeer, 2021) tackled the challenging task of detecting sarcasm in Arabic tweets. The authors explored innovative approaches leveraging the power of BERT models: Fine-tuning AraBERTv02, this established model was tailored specifically for the sarcasm detection task by fine-tuning its internal parameters. Capturing Tweet Essence with Sentence-BERT, trained through contrastive learning, this specialized model extracted concise yet comprehensive representations of each tweet, capturing its core meaning. Finally combining Brains AraBERT + Sentence-BERT, inspired by how humans interpret sarcasm both literally and figuratively, they combined the strengths of AraBERT with the semantic richness of Sentence-BERT embeddings. As a results, the model achieved F1-score of 59.89%, this impressive achievement validates the effectiveness of our combined model in uncovering the complexities of sarcasm in Arabic tweets.

BERT's groundbreaking impact on natural language processing is undeniable. Yet, much of the research has focused on applying it to the well-resourced English language.



Recognizing this gap, the analysis in paper (Alammary, 2022) delved into the cutting-edge applications of BERT for Arabic text classification. The analyses Identify BERT models employed for Arabic text classification, compared their performance against each other and assessed their effectiveness relative to the original English BERT models. Comprehensively analyzing 48 articles, it uncovers several key findings: Nine diverse BERT models cater to Arabic text classification needs. Two multilingual models include Arabic, one caters to both Arabic and English, while six are Arabic-specific. High performance, comparable to English BERT models, is prevalent across the reviewed studies. Standout performers, in descending order, are MARBERT, QARiB, ARBERT, AraBERT, and ArabicBERT.

There are remaining research gaps in cyberbullying detection using AraBERT, limited focus on Facebook comments, most existing studies utilize Twitter data, while Facebook comments present unique challenges due to their informal nature and diverse contexts, also lack of domain-specific adaptation, AraBERT may benefit from fine-tuning on Facebook comment data to improve its understanding of specific linguistic nuances and cyberbullying expressions in this context. This thesis proposes to address these research gaps by investigating the effectiveness of AraBERT for detecting cyberbullying in Facebook comments. Table 1, comparing the different methods for Arabic cyberbullying detection.

Table 1. Cyberbullying Detection Methods in Arabic Text

Method	Approach	Limitations	Advantages
Keyword-based	Matching predefined keywords or phrases	Limited to explicit bullying language, struggles with sarcasm and context	Simple to implement, computationally efficient
Machine Learning Models	Classifying text based on features	Requires manual feature engineering, sensitive to data imbalance	Can capture patterns beyond keywords, more adaptable
Deep Learning Models	Learning contextual relationships between words	Computationally expensive, requires large datasets	Excellent performance on various NLP tasks, including sentiment analysis
BERT Models	Pre-trained language model for various NLP tasks	Requires fine-tuning for specific tasks, may struggle with domain-specific language	Strong performance on many NLP tasks, including text classification

AraBERT Model	Arabic-specific pre-trained language model	May struggle with low-resource languages or dialects	Tailored for Arabic, leverages pre-trained knowledge on Arabic language nuances
---------------	--	--	---

### Chapter 3: Research Methodology

The framework of our thesis to effectively detect cyberbullying comments in Arabic, a substantial Arabic-language dataset is essential. While there are numerous datasets available online related to bullying, many of them are in English and primarily focus on content from platforms like Twitter and YouTube. Therefore, we intend to build a dataset that contains Arabic comments from Facebook and label part of them manually, more details are described in section 3.1 and preprocessing steps are applied on the dataset. After that we use the Bert Transformer, figure 6 summarize the framework and methodology of our research.

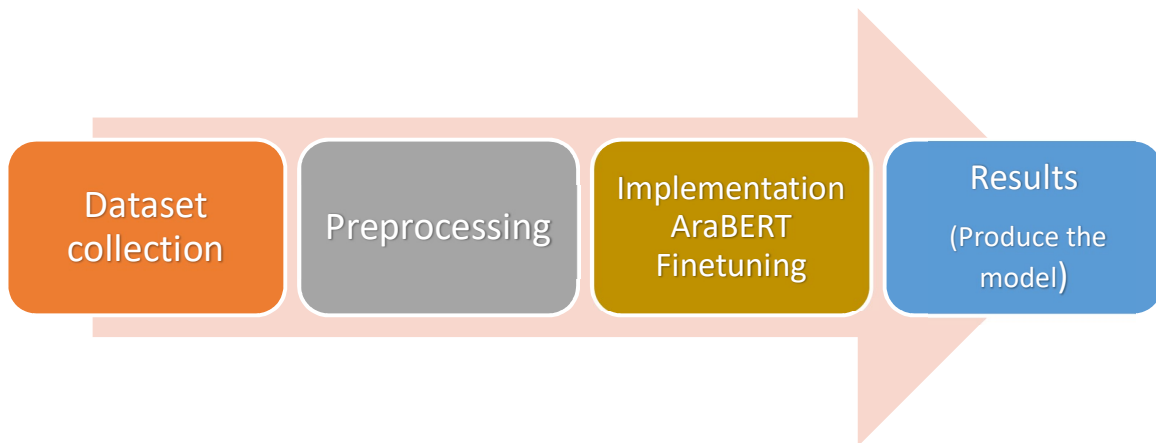


Figure 6. Methodology.

#### 3.1 Dataset Collection

With 2.93 billion active users, Facebook has grown to become an essential resource for businesses and other applications, However, gathering Facebook data manually can require a lot of time and resources. Errors and mistakes are common with manual web scraping, particularly in large-scale data collection techniques (Sedat Dogan, 2024). No code with Facebook scrapers, users can get a lot of data from several Facebook pages. Even though, a lot of these Facebook scrapers provide free schemes, some of their additional expensive features are available for a paid subscription. However other method using Facebook APIs that needs authorized and controlled access. In this thesis

we use Apify tool because it support Python and JavaScript libraries and gives free usage for 7 days using google account, also it allows to extract data from Facebook pages using URLs that can be downloaded as JSON, CSV and Excel file to be ready for using in applications and researches(*Facebook Posts Scraper*, 2024). Facebook Pages Scraper has been developed to be simple to use, even for those who have never taken an interest in gathering data from web pages. The maximum size of the resulted dataset allowed is 5000 rows, and the main features of the dataset when collecting comments of the post are: date, Facebook Id, Facebook URL, feedback Id, id, likes Count, post Title, profile Id, profile Name, profile Picture.

For our model implementation, we use the addresses of Arabic pages that have millions of followers like Aljazeera channel, Roya kitchen, MTV Lebanon, Ramallah News and other public pages in different subjects and years, then we choose the posts that have more than 1k comments and contains bully and nun bully comments and built a dataset of 40,000 comments. The next step, we prepare a labeled dataset of 20,000 rows to train the model, the labeling process was in collaboration with psychology specialists from Al Quds Open University to determine which the comment is bully or not, the comments that contains bullying annotated as 1 and the others does not contain bullying annotated as 0, the percentage of bullying comments and not bullying comments shown in Figure 7. Figure 8 shows a sample of our dataset.

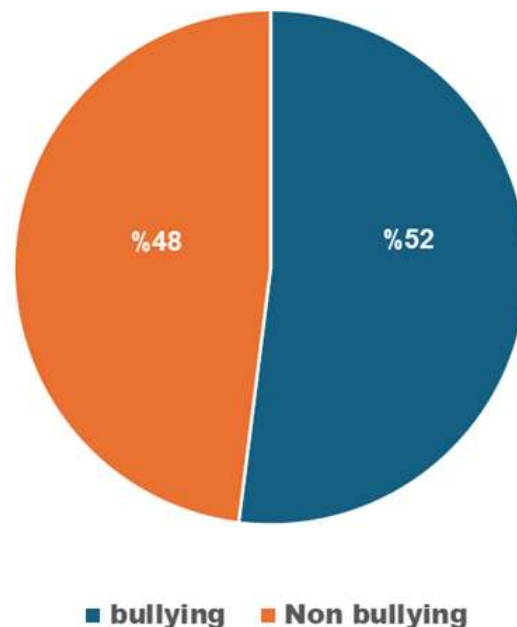


Figure 7. The percentage of bully and not bully comments in the dataset.

	text	label
3143	الله لا يبارك ولا فيه ولا فيها	1
19546	يلعن حرمكم شو انكم قرون	1
6556	تسلم الأبيادي ابدعني يا شيف	0
2448	الا لعنة الله عليك وعلى من هو على شاكلتك	1
7087	جسسيل	0

Figure 8. A sample of the collected dataset.

### 3.2 Dataset Preprocessing

Any effective NLP application must prioritize eliminating noise and minimizing data sparsity. Working with Arabic dialects in the context of user-generated online content presents significant challenges. The diversity of Arabic dialects, coupled with their widespread use in online platforms, can lead to data sparsity and hinder the performance of natural language processing models. This is due to the difficulty in training models on limited data from various dialects and the potential for models to struggle with generalizing to unseen dialectal variations.(Althobaiti, 2022)(Althobaiti, 2020). Furthermore, content on the internet usually looks noisy, containing a lot of tags, a lot of spaces, characters that are repeated, and Arabizi, a language that some people transliterate from Arabic to Latin letters and numbers while writing online. During the preprocessing stage, we used some procedures to delete HTML tags and other symbols, deleting URLs, usernames, deleting hashtags and mentions to prepare our dataset.

The study of natural language processing NLP focuses on making machines able to understand and analyze human language. To do this, text data must be represented in a form that is understandable by computers. Tokens, vectors, and embeddings are all significant here. Tokenization is effectively dividing text into smaller units called tokens, this is the first stage. Depending on the chosen method, these tokens can be single words, punctuation, or even smaller units, tokens are then converted into numerical representation called vectors. A deep learning model, typically a transformer model, is trained on these vectors to provide meaning to these tokens. This makes it

possible for the model to understand word meanings and their relationships. The aim of this approach is to enabling NLP models to learn the semantics and meaning of various words as well as their context inside a sentence or text. For example, the term "King" and "Man" may have similar vectors, also do the words "Queen" and "Woman" Additionally, these vectors have certain characteristics that make them useful for Language model training. Words are not represented by a single number in NLP models. They frequently represent a single word with more than 1,000 integers, it is known as a word vector. One way to think about a word vector is as a point in a multidimensional space, with each dimension representing a different feature or aspect of the word(COLYER, 2016) . For instance, a word vector for the term "queen" might have low values for the dimensions that represent "masculinity" and high values for the dimensions that represent "femininity" and "royalty". The vector for "queen" is created by combining all these dimensions, that gives our model the ability to understand the meaning of the word and its relationships with other words.

Word tokenization has different forms, character, word and subword (combine the benefit of using character and word forms). The most basic form of tokenization is character tokenization, which treats each character as a unique token. Character-level tokenization interprets the entire string as a set of individual characters and excludes any text structure. The main disadvantage of this strategy is that it requires a lot of computer power, memory, and data to learn words character by character, even though it can be helpful for handling misspellings and rare words. Character tokenization is therefore not commonly applied in practical situations. An alternative form character tokenization is word tokenization, by treating text as individual words, word tokenization makes training simpler, but it can result in a big vocabulary for uncommon words(Lewis Tunstall, Leandro von Werra, 2022). As a result, a neural network has more parameters to learn. Some approaches handle this by limiting the vocabulary to the most widely used words and designating unknown words with a common "unknown" token. However, this method loses some potentially important information in the process. The best of both worlds is provided by subword tokenization that combines character and word tokenization benefits, it preserves common phrases while breaking uncommon words down into smaller but meaningful pieces (such as prefixes or suffixes). This method prevents the model from becoming extremely overloaded by

an excessively big vocabulary, while yet enabling it to handle complex words and potential errors (such as misspellings), figure 9 show tokenization forms(Ashraf, 2024).



Figure 9. Tokenization forms(Ashraf, 2024).

By generating a set of unique tokens from the training data (corpus), tokenization helps build up the vocabulary. Although a large corpus is useful, high-quality data is crucial. The model's performance might be greatly affected by even small errors in the training set. For effective NLP model training, a big corpus of clear, correct text is essential. Even though we now know that tokens are just numbers that represent words, these original vectors are meaningless. These vectors are converted into a more meaningful space through the embedding process. The model improves these vectors throughout training in order to capture the semantic relationships between words. Imagine that in this multi-dimensional space, words that have similar meanings are closer to one another. As a result, the model can automatically discover word correlations without human assistance. By effectively giving the numerical representations life, embedding makes it possible for the model to understand the nuances of language(Metzger, 2022). Pre-trained on an extensive corpus of Arabic text, AraBERT captures important semantic details. As compared to training a model from scratch, this reduces the training

time, and it has been trained only for Arabic, which is essential for handling dialect variances and understanding the nuances of the language.

### 3.3 AraBERT Model Selection and Fine-tuning.

With the advancement of transformer-based models in NLP, the pre-training of language-specific models, such as BERT, has shown to be very effective in achieving high performance on various NLP tasks. By pre-training BERT specifically for the Arabic language, researchers aim to overcome the challenges faced by Arabic NLP and achieve similar success to what BERT has achieved for the English language. The newly developed model, AraBERT, has shown to achieve state-of-the-art results on various Arabic NLP tasks, such as sentiment analysis, named entity recognition, and question answering, compared to multilingual BERT models and other state-of-the-art approaches (WissamAntoun, 2022). The architecture of our model is shown in figure 10.

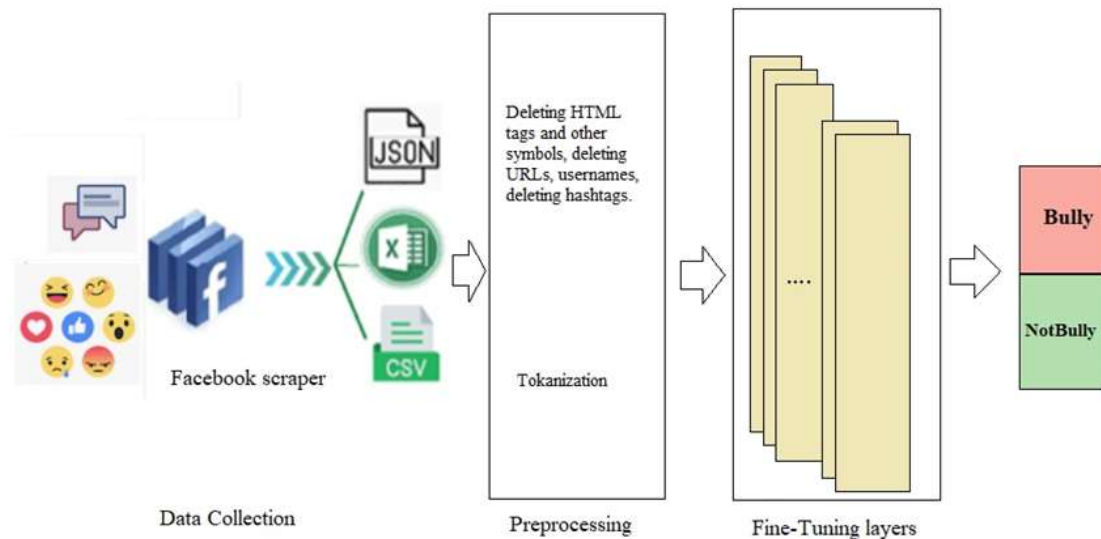


Figure 10. The proposed architecture using AraBERT.

The main differences between AraBERT versions according to Vocabulary and Training Data: v0.1, Less data and a smaller vocabulary (maybe worse performance). v1, highlights pre-segmented text, which could enhance performance for tasks that required for segmentation. v2 (base & large), is the most recent variation, with improved vocabulary and data and considered to be the most performant, as shown in table 2. Ara-BERTv2 emerged as the most accurate model and also excelled in other

natural language processing (NLP) tasks that corresponds to the findings of (Inoue *et al.*, 2021)(Al-Ghamdi, Al-Khalifa and Al-Salman, 2023). AraBERTv2 employs its own pre-processing techniques that are likely to include handling punctuation, numbers, and building a rich vocabulary, eliminating the need for Farasa segmentation.

Table 2. AraBERT model versions differences.

Model	Size (MB/Params)	Pre-Segmentation	DataSet (Sentences/Size/nWords)
AraBERTv0.2-base	543MB / 136M	No	200M / 77GB / 8.6B
AraBERTv0.2-large	1.38G / 371M	NO	200M / 77GB / 8.6B
AraBERTv2-base	543MB / 136M	YES	200M / 77GB / 8.6B
AraBERTv2-large	1.38G / 371M	YES	200M / 77GB / 8.6B
AraBERTv0.1-base	543MB / 136M	No	77M / 23GB / 2.7B
AraBERTv1-base	543MB / 136M	YES	77M / 23GB / 2.7B

For AraBERT model integration there are two approaches: Fine-tuning and feature extraction, according to the literature reviews, fine-tuning approach outperforms feature extraction by using the pre-training knowledge of AraBERT for our task in detecting cyberbullying also AraBERT's deep learning architecture enables it to identify the subtle characteristics of language used in cyberbullying by capturing complex semantic relationships within the text. After the preparation of the dataset, including cleaning, preprocessing and split the dataset into training and testing sets, we choose the AraBERT version based on the data size and task complexity. Fine-tuning process contains the following steps:

- Loading the pre-trained AraBERT model for fine-tuning, utilize libraries like Transformers provided by the AraBERT developers to load the chosen AraBERT version, like AraBERTv2-base. A new layer appropriate for our binary classification task (cyberbullying vs. not cyberbullying) should be implemented in place of the model's final classification layer, this layer is designed for the original task trained on like masked language modeling.
- In the pre-trained model, we freeze the weights of the earlier layers, it is a common technique used in transfer learning. It involves preventing the weights of the initial layers from being updated during the fine-tuning process and helps focus on features more specific for the new task, prevent overfitting especially when dealing with limited datasets\_and



reduce computational load. During training, the weights in these frozen layers are not changing, the only weights that are changed are those in the recently placed classification layer and maybe a few last layers of the pre-trained model(Pirinççi, 2022).

- Using our labeled cyberbullying dataset, we train the model by changing the number of epochs (training iterations), batch size, and learning rate. The model will be trained to map the text features (AraBERT embedding) to the relevant cyberbullying labels by changing the weights in the newly added layers as well as some of the upper layers of the pre-trained model.
- Tracking the validation set's performance to avoid overfitting and determine the perfect model configuration. Split the dataset into training, validation, and test sets. The validation set is used to monitor performance during training and identify the optimal stopping point, track metrics like accuracy, precision, recall, and F1-score on the validation set after each epoch. Early stopping is a common technique where training is halted if validation performance plateaus or starts to decline, preventing overfitting to the training data.
- Evaluate the Test Set, test our optimized model on the held-out test set once we've adjusted its hyperparameters based on validation set performance. Test set performance offers a more accurate approximation of the model's ability to generalize to unseen data.

The advantages of Fine-tuning, leverage using of AraBERT's pre-trained knowledge to Arabic language understanding and AraBERT is modified for the specific task of detecting cyberbullying in Arabic. Fine-tuning is often better than using a separate classifier for feature extraction, especially with existing limited data.

### **3.3.1 AraBERT Model Setup**

In the implementation process, Google Colaboratory which is known as Colab is used. It is a free Jupyter Notebook environment that contains many pre-installed libraries like Keras, Tensorflow, Open CV and others. Colab is a free cloud-based platform that facilitate the coding experience by import dataset and training models(*Welcome To*

Colab, 2024). In the first step for implementing AraBERT model, we start by installing two libraries from HuggingFace that will make our job easier, transformers and datasets, these libraries are essential for working NLP tasks, **transformers** is a popular Python library, provides tools for utilizing pre-trained language models to understand and manipulate text data, which contains powerful algorithms trained on large size of data(*The AI community building the future.*, 2024). Other one, **datasets** libraries that needs to work and access relevant datasets for training and evaluation the NLP models. We also import the relevant libraries that are commonly used in NLP tasks like numpy which provide powerful tools for numerical computation, panda which used for organizing and analyzing data and torch that is a popular framework for deep learning used in building and training neural network models including text classification. BertForSequenceClassification class is imported, it is a pre-trained BERT model designed for sequence classification tasks also BertTokenizerFast class that helps tokenize text data to numerical representation suitable for the pre-trained BERT model specially for processing large datasets.

In our model, we initialize bert-base-arabertv2 model, then we prepare our data for training the AraBERT model, after preparing the data, we spilt ratio (80%/20%) for training and testing the model.

### 3.3.2 Training and Evaluation

The training process starts by definition of a function that will help us monitor the training progress and evaluate results on the validation dataset. We will focus on the following metrics:

- Precision: it measures the proportion of correctly identified cyberbullying comments (true positives) among all the comments the model predicted as cyberbullying (true positives + false positives). Measures the model's ability to avoid false positives, a high precision indicates that when the model predicts an instance as positive, it is likely to be correct.
- Recall: this measures the proportion of actual cyberbullying comments the model identified correctly (true positives) to the total number of actual positive comments (true positives + false negatives), a high recall indicates that the model is able to capture most of the cyberbullying cases.

- F1: the average of the harmonics of recall and precision gives the F1 score, which offers an accurate evaluation of the performance of the model.
- Accuracy: The total proportion of correctly classified comments ((both true positives and true negatives)) to the total number of comments, is known as accuracy.

HuggingFace wraps up the default transformer fine-tuning approach in the Trainer object, and we can customize it by passing training arguments such as learning rate, number of epochs, batch size etc. We set logging steps to 20, so that we can frequently evaluate how the model performs on the validation set throughout the training. Initially, the most recent stable version of PyTorch is downloaded and installed from the PyPI (Python Package Index) repository, and it is included to the libraries that are available in our current Colab runtime. We can import and utilize PyTorch functionalities in our Python code within the Colab notebook as soon as the installation is finished, tensors (multidimensional data structures) can be worked with, deep learning operations can be carried out, neural network architectures can be defined, and deep learning models can be built and trained using PyTorch. Leveraging GPU Acceleration for faster training in Colab by change the runtime type to GPU, this will allow PyTorch to utilize GPUs for faster computation. From the Hugging Face Transformer library, we import two components, TrainingArguments and Trainer, these components are important for training and evaluation machine learning models especially NLP tasks. TrainingArguments class is needed to define and configure training hyperparameters for the model, these are including Learning rate, number of epochs, Batch size and Optimizer. Trainer class functions as a high-level wrapper for executing several deep learning operations as well as for training and evaluating the model.

The hyperparameters for training our AraBERT model, learning rate arguments is  $2e-5$ , it is common point for starting fine-tuning, however, it controls how the model weights are updating during training process. Epochs value sets as 3, it is the number of time the training dataset passed through the model in training process. Batch size is 16, it indicates number of examples processed together in training step, increase this value will decrease the time of training process, because of hardware limitations, batch size 16 is reasonable in this model. For not using mixed precision training, fp16 was

disabled and using GPU for potentially faster training without effects on accuracy also logging steps sets to 20 to provide regular logging updates (Moonat, 2022) (GitHub, 2022). Evaluation strategy arguments is 'steps', this means the evaluation process of the model will be after a fixed number of training steps without relying on epochs.

## Chapter 4: Experiments and Results

In this chapter, the experimentation process and the manual annotation results are presented in detail for detecting Arabic cyberbullying using AraBERTv2 model. we investigate the effectiveness of apply freezing and fine-tuning different number of layers in the model. The proposed models' performance is evaluated using several kinds of measures, including F-measure, accuracy, recall, and precision. We use a 20K balanced dataset of Arabic cyberbullying comments that are collected from Facebook pages, 10246 of them labeled as 1 (bullying comments).

### 4.1 Experimentation process

The core elements of text analysis are BERT layers, such as those likely employed in Ara-BERTv2. They start by converting words into numbers, and then they use a really intelligent technique known as self-attention to examine the relationships between each word and other words in the phrase. Imagine being able to understand a sentence and how each word links to the others to make an overall thought. BERT layers achieve exactly that, and by stacking up layers, they can progressively improve this understanding to extract the text's main ideas. Because of this, BERT and its variations are effective tools for a range of NLP applications (Rogers, Kovaleva and Rumshisky, 2020). BERT model contains 12 layers; it seems that the most information on linear word order could be found in the lower layers. According to (Lin, Tan and Frank, 2019), there has been a decrease in the understanding of linear word order in layer 4 of BERT-base. Several researches with different tasks agreed that the syntactic information are captured in the middle layers of BERT, which refer to the grammatical structure and how words function in the sentence (Hewitt and Manning, 2019) and the final layers are tasks specific (Liu *et al.*, 2019). Figure 11 indicate how the different layer transfer the learning knowledge for various tasks of NLP, the columns in the figure refer to the probing tasks.

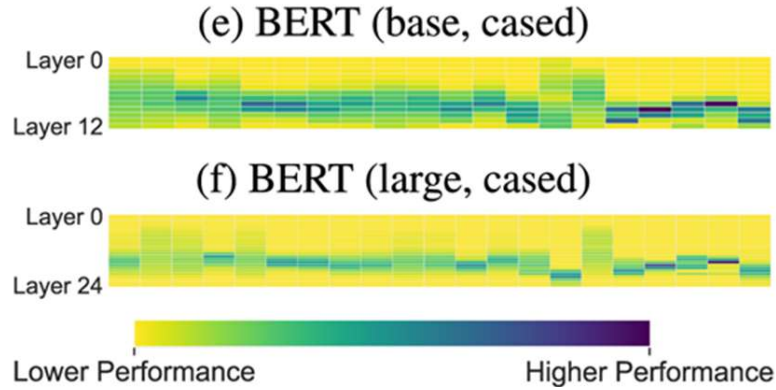


Figure 11. BERT Layers Transferability by Task(Liu *et al.*, 2019).

Despite the BERT models are well- defined and relatively complex structure, several research studies have examined how BERT models can be used to improve model performance and whether fine-tuning all pre-trained layers is required for efficient NLP tasks (Rogers, Kovaleva and Rumshisky, 2020). Fine-tuning more layer in the model lead to learn a combined representation of information from both deep and output layers of the BERT model, potentially capturing aspects missed by focusing on a single layer, also using information from all BERT layers with assigned weights, this allows the model to prioritize specific layers based on their relevance to the task (You *et al.*, 2020). To explore our model, we evaluated different techniques that fine-tuned different number of layers in AraBERTv2 model and freeze the others. Freezing layers in a pre-trained model like AraBERT is a common technique, it reduces the number of training parameters and reduces training time(Al-Ghamdi, Al-Khalifa and Al-Salman, 2023).

In our experiments, the same training parameters setting were applied in all runs, batch size =16, epoch = 3 and learning rate =  $2e-5$ . We compare in Table 3 between freezing strategies used in the model according to computational load and training time, tuning more layers need more for using system resources like RAM, GPU and Hard Disk. Unfreeze layers' strategy need more computational load, so if we increase the size of the dataset for training our model may lead to overload. Because of using the same batch size in all runs, system RAM consumptions are the same. We applied for some experiments:

- Freeze all layers but the final layer, this is a common method where we train the final classification head only and freeze all the AraBERT pre-trained layers

(encoder layers). This makes use of the pre-trained features and modifies the model according to the specific classification task.

- Freeze specific layers: we may choose to freeze a certain number of the initial encoder layers while training the remaining layers and the classification head, here we freeze 9, 7 and 5 layers respectively. This allows for some fine-tuning of the pre-trained features based on our data.
- Unfreezing layers: in this method we fine-tuning all layers.

Table 3. Computational load and training time.

Strategy	Training time (minute)	System RAM (12.7 GB)	GPU RAM (15 GB)	System Hard Disk (78.2 GB)
Freeze all layer but final layer	24.5	3.6	7.5	35.2
Freeze 9 layers	25.54	3.6	7.9	35.7
Freeze 7 layers	26.42	3.3	8.2	36.3
Freeze 5 layers	27.40	3.6	8.5	36.7
Unfreeze layers	30.22	3.6	9.3	38.0

## 4.2 Results and error analysis

We experimented with AraBERT fine-tuning to detect cyberbullying in Arabic. During the fine-tuning stage, the model layout required freezing or unfreezing of layers. We used the testing accuracy to evaluate the performance. Results in table 4 indicated that 91.9% testing accuracy was achieved by unfreezing all layers. This significant enhancement implies that the model is more capable to capture the complexity of Arabic cyberbullying language by unfreezing layers, figure 12 show the evaluation matrix of our model results.

Table 4. Freezing strategies used in the model.

strategy	Accuracy	F1	Precision	Recall
Freeze all layer but final layer	81.7	83.5	86.7	80.5
Freeze 9 layers	83.0	84.8	87.0	82.8
Freeze 7 layers	85.3	86.6	90.8	82.8
Freeze 5 layers	87.7	88.7	93.8	84.5
Unfreeze layers	91.9	92.8	94.7	90.9

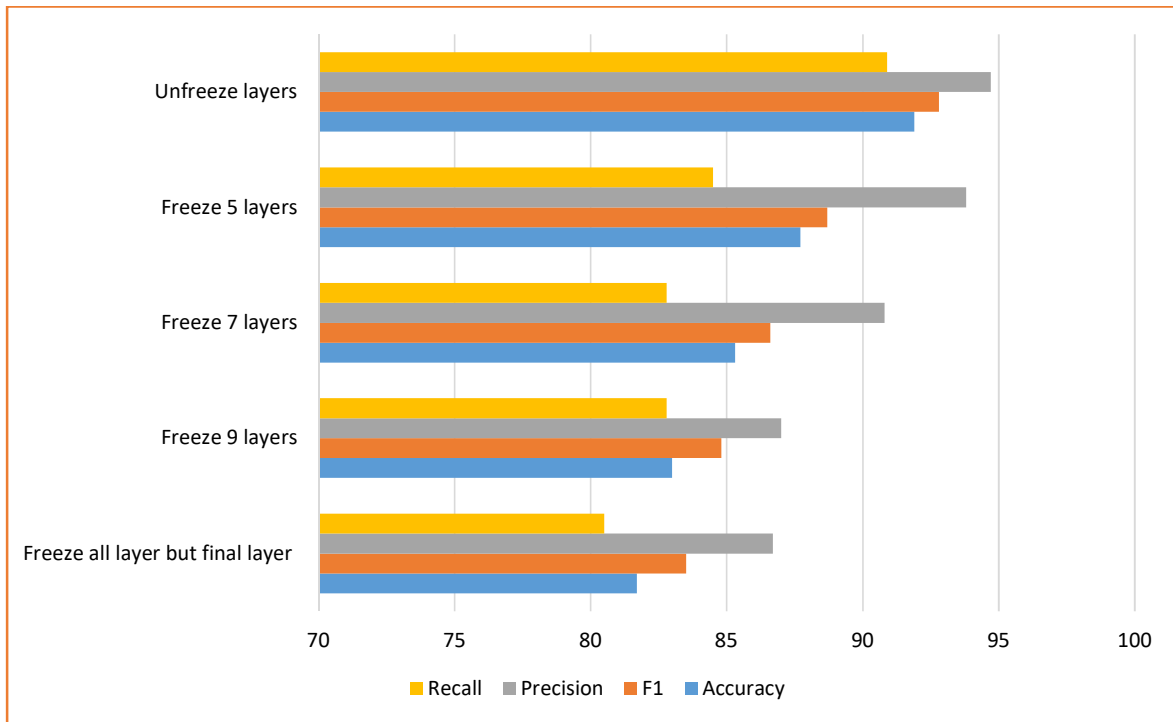


Figure12. The evaluation matrix of our model results.

For error analysis, we utilized confusion matrices and sample-level analysis in order to obtain a deeper knowledge of the behavior of the model. Confusion Matrix is an effective way to visualize error analysis, when evaluating a model's performance in classification tasks, a confusion matrix is a powerful visual tool. It clearly demonstrates the number of samples from each class that were correctly (true positives and true negatives) and incorrectly (false positives and false negatives) classified. For construction of the confusion matrix, we generate a matrix where the true and predicted classes are represented by rows and columns, respectively. Provide sample counts to each cell in the testing set according to the true and expected labels. This matrix helps in highlighting areas in need of development. High false negatives (FN), for instance, suggest that the model missed an excessive amount of real cyberbullying comments(Bhandari, 2024). for our AraBERT fine-tuning experiment on cyberbullying detection, the "Cyberbullying" and "Not Cyberbullying" classes would be represented by rows and columns in the confusion matrix. we can learn more about the following by examining the distribution of values (TP, TN, FP, FN), how well the model recognizes real cyberbullying comments (TP and FN) and the frequency with which non-bullying comments are misclassified as cyberbullying by the model (FP). Figure 13 show the confusion matrix for the model experiments.



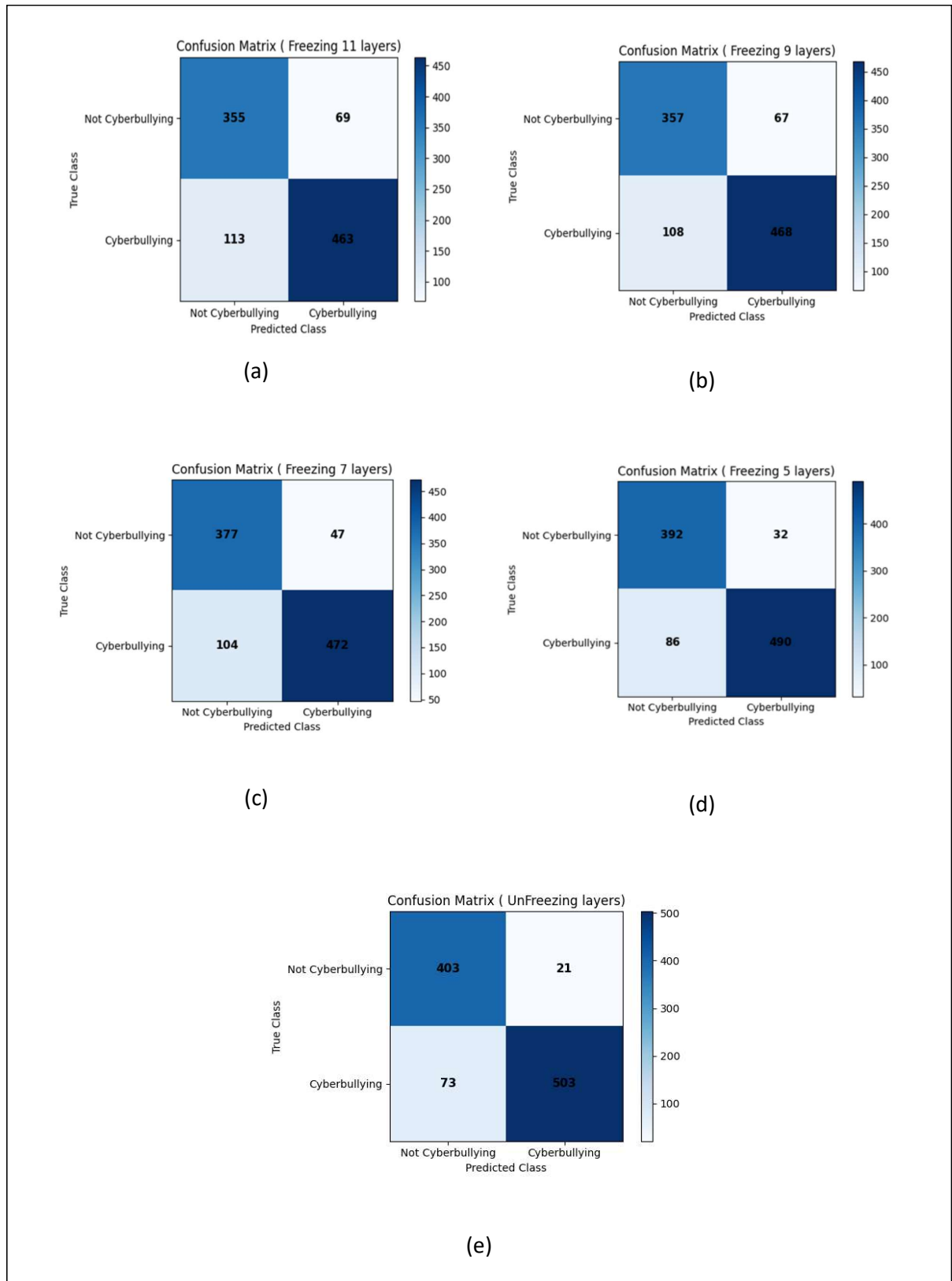


Figure13. Confusion matrix for the model experiments.

Error analysis showed that unfreezing layers improved the ability of the model to identify specific types of cyberbullying comments, like sarcasm or indirect language. Furthermore, accurate identification of even minor bullying is prioritized in the real-world setting of cyberbullying detection. In view of the significant improvement in accuracy and the focus on efficiency, unfreezing layers is appropriate for this purpose.

Remarkably, the tests showed that the model's detection of cyberbullying comments using sarcasm or indirect language improved by unfreezing layers. The frozen model often misclassified these more complex types of bullying, indicating that unfreezing enables the model to accurately represent the nuances of Arabic cyberbullying speech. This is consistent with the noted increase in testing accuracy, since the model is probably able to learn domain-specific features of Arabic cyberbullying text that are not fully represented by pre-trained AraBERT because of the unfreezing layers.

### **4.3 Discussion**

The results of our investigation showed that the best accuracy for Arabic cyberbullying detection came from fine-tuning the full AraBERTv2 model. Below is an in-depth analysis of the relevant triggers:

- **Learning and Model Capacity Complex patterns**, Since AraBERT v2 is a pre-trained model, the early layers of the model extract the general language features from large size of Arabic corpus. These layers serve as a basis for higher layers by acting as a feature extractor in the model. Also AraBERT v2 is not Task-Specific learning, that means, the following layers are not trained for cyberbullying detection in mind. By fine-tuning each layer, we enable them to learn features that differentiate cyberbullying text from non-cyberbullying text, allowing the layers to adapt to the specific task.

This is consistent with the research by(Howard and Ruder, 2018), who indicate that tasks requiring the learning of complex patterns, such as text classification or sentiment analysis, can benefit greatly by fine-tuning all layers of a pre-trained model. In our situation, understanding subtle linguistic clues that go beyond basic language features is probably necessary to detect cyberbullying. By fine-tuning, the model is able to learn these complex patterns to the best of its abilities.

- In related to feature extraction and adaption, the first layers of the model capture the basic of general language feature such as word embedding and syntactic structure. These features are essential for understanding the overall meaning of the text. the next layers generally concentrate on either higher-level features or word relationships. These layers can be made more suitable for the specific task of detecting cyberbullying by fine-tuning them. They can be trained to recognize features of cyberbullying, such as word choice, sentiment, and sentence structure patterns.

## Chapter 5: Conclusion and Future work

This thesis explored how well the AraBERT v2 pre-trained model identified Arabic cyberbullying text. A variety of AraBERT v2 model layers were frozen in our examinations, but the remaining layers continued trainable, in an effort to assess various fine-tuning approaches. Our results show that the best accuracy for Arabic cyberbullying detection was achieved by fine-tuning the total AraBERT v2 model (unfreezing all layers), freezing some layers may limit the model's ability to learn the complex patterns required for best cyberbullying detection. This suggests that the model's full capacity is essential for learning the nuances of cyberbullying language in Arabic, freezing layers might limit the model's ability to adapt to specific patterns and features that are crucial for accurate detection. These results add to the significant growing of literature on the use of deep learning techniques and AraBERT for Arabic natural language processing applications, especially in the area of Arabic cyberbullying detection, also in this thesis we built an Arabic dataset that contain Arabic cyberbullying text to increase Arabic resources to be available for other application.

There are some limitations to this research that present possibilities for future works study, the findings may vary depending on the quantity and quality of the Arabic cyberbullying dataset applied, an interesting approach for future research is to examine how bigger and more diverse datasets affect AraBERT v2 and other models' efficacy in detecting cyberbullying. A potential next step could be integrating the fine-tuned AraBERTv2 model into practical applications for online content monitoring or educational platforms. Practical implementation would need evaluating the model's performance in such settings and addressing potential challenges like scalability and simplicity.

The potential impact of this research is the development of applications that employ deep learning and AraBERT for Arabic text analysis. In addition, the results can produce automated tools that can effectively help in addressing and detection Arabic cyberbullying comments in Facebook and other online platforms, that will constitute a safer online environment for Arabic speaking communities

## References

- Al-Ghamdi, S., Al-Khalifa, H. and Al-Salman, A. (2023) ‘Fine-Tuning BERT-Based Pre-Trained Models for Arabic Dependency Parsing’, *Applied Sciences (Switzerland)*, 13(7). Available at: <https://doi.org/10.3390/app13074225>.
- Alakrot, A., Murray, L. and Nikolov, N.S. (2018a) ‘Dataset Construction for the Detection of Anti-Social Behaviour in Online Communication in Arabic’, *Procedia Computer Science*, 142, pp. 174–181. Available at: <https://doi.org/10.1016/j.procs.2018.10.473>.
- Alakrot, A., Murray, L. and Nikolov, N.S. (2018b) ‘Towards Accurate Detection of Offensive Language in Online Communication in Arabic’, *Procedia Computer Science*, 142, pp. 315–320. Available at: <https://doi.org/10.1016/j.procs.2018.10.491>.
- Alammary, A.S. (2022) ‘BERT Models for Arabic Text Classification: A Systematic Review’, *Applied Sciences (Switzerland)*, 12(11). Available at: <https://doi.org/10.3390/app12115720>.
- Aldjanabi, W. *et al.* (2021) ‘Arabic offensive and hate speech detection using a cross-corpora multi-task learning model’, *Informatics*, 8(4), pp. 1–13. Available at: <https://doi.org/10.3390/informatics8040069>.
- AlHarbi, B.Y. *et al.* (2019) ‘Automatic cyber bullying detection in Arabic social media’, *International Journal of Engineering Research and Technology*, 12(12), pp. 2330–2335.
- Alsafari, S., Sadaoui, S. and Mouhoub, M. (2020) ‘Hate and offensive speech detection on Arabic social media’, *Online Social Networks and Media*, 19(September), p. 100096. Available at: <https://doi.org/10.1016/j.osnem.2020.100096>.
- Alshalan, R. and Al-Khalifa, H. (2020) ‘A deep learning approach for automatic hate speech detection in the saudi twittersphere’, *Applied Sciences (Switzerland)*, 10(23), pp. 1–16. Available at: <https://doi.org/10.3390/app10238614>.
- Alshamsi, S.A.-D.A.E.B.A.A. (2023) ‘Transformer-based Arabic Offensive Speech Detection’, in *2023 International Conference on Emerging Smart Computing and Informatics (ESCI)*. Pune, India.
- Alsubait, T. and Alfageh, D. (2021) ‘Comparison of Machine Learning Techniques for Cyberbullying Detection on YouTube Arabic Comments’, *International Journal of Computer Science & Network Security*, 21(1), pp. 1–5. Available at: <https://doi.org/10.22937/IJCSNS.2021.21.1.1>.
- Althobaiti, M.J. (2020) ‘Automatic Arabic Dialect Identification Systems for Written Texts: A Survey’, *International Journal of Computational Linguistics (IJCL)*, 11(3).
- Althobaiti, M.J. (2022) ‘Creation of annotated country-level dialectal Arabic resources: An unsupervised approach’, *Natural Language Engineering*. 2021/08/09, 28(5), pp. 607–648. Available at: <https://doi.org/DOI: 10.1017/S135132492100019X>.
- Alzaqebah, M. *et al.* (2023) ‘Cyberbullying detection framework for short and imbalanced Arabic datasets’, *Journal of King Saud University - Computer and Information Sciences*, 35(8), p. 101652. Available at: <https://doi.org/10.1016/j.jksuci.2023.101652>.

Antoun, W., Baly, F. and Hajj, H. (2020) ‘AraBERT: Transformer-based Model for Arabic Language Understanding’, in *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools*. Marseille, pp. 9–15.

Ashraf, A. (2024) *Tokenization in NLP : All you need to know, medium*. Available at: <https://medium.com/@abdallahashraf90x/tokenization-in-nlp-all-you-need-to-know-45c00cfa2df7>.

Bashir, E. and Bouguessa, M. (2021) ‘Data Mining for Cyberbullying and Harassment Detection in Arabic Texts’, *International Journal of Information Technology and Computer Science*, 13(5), pp. 41–50. Available at: <https://doi.org/10.5815/ijitcs.2021.05.04>.

Bashmal, L. and Alzeer, D.H. (2021) ‘ArSarcasm Shared Task: An Ensemble BERT Model for Sarcasm Detection in Arabic Tweets’, *WANLP 2021 - 6th Arabic Natural Language Processing Workshop, Proceedings of the Workshop*, pp. 323–328.

Bhandari, A. (2024) *Understanding & Interpreting Confusion Matrix in Machine Learning, analytics vidhya*. Available at: <https://www.analyticsvidhya.com/blog/2020/04/confusion-matrix-machine-learning/>.

Boudad, N. *et al.* (2018) ‘Sentiment analysis in Arabic: A review of the literature’, *Ain Shams Engineering Journal*, 9(4), pp. 2479–2490. Available at: <https://doi.org/10.1016/j.asej.2017.04.007>.

Boyd, D.M. and Ellison, N.B. (2007) ‘Social network sites: Definition, history, and scholarship’, *Journal of Computer-Mediated Communication*, 13(1), pp. 210–230. Available at: <https://doi.org/10.1111/j.1083-6101.2007.00393.x>.

COLYER, A. (2016) *The amazing power of word vectors, KDnuggets*. Available at: <https://www.kdnuggets.com/2016/05/amazing-power-word-vectors.html>.

Copeland, B. (2024) *artificial intelligence, Encyclopedia Britannica*. Available at: <https://www.britannica.com/technology/artificial-intelligence>.

Cortis, K. and Handschuh, S. (2015) ‘Analysis of cyberbullying tweets in trending world events’, *ACM International Conference Proceeding Series*, 21-22-Octo. Available at: <https://doi.org/10.1145/2809563.2809605>.

Devlin, J. *et al.* (2019) ‘BERT: Pre-training of deep bidirectional transformers for language understanding’, *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 1(Mlm), pp. 4171–4186.

Dreamgrow (2024) *The 15 Biggest Social Media Sites and Apps in 2024, dreamgrow*. Available at: <https://www.dreamgrow.com/top-15-most-popular-social-networking-sites/>.

El-Alami, F. zahra, Ouatik El Alaoui, S. and En Nahnahi, N. (2022) ‘A multilingual offensive language detection method based on transfer learning from transformer fine-tuning model’, *Journal of King Saud University - Computer and Information Sciences*, 34(8), pp. 6048–6056. Available at: <https://doi.org/10.1016/j.jksuci.2021.07.013>.

*Facebook Posts Scraper (2024) APIFY*. Available at: <https://apify.com/apify/facebook-posts-scraper>.

- Faraj, D. (2022) ‘SarcasmDet at Sarcasm Detection Task 2021 in Arabic using AraBERT Pretrained Model SarcasmDet at Sarcasm Detection Task 2021 in Arabic using AraBERT Pretrained Model and Technology’, in *Proceedings of the Sixth Arabic Natural Language Processing Workshop*. Kyiv, Ukraine, pp. 345–350.
- Gillioz, A. *et al.* (2020) ‘Overview of the Transformer-based Models for NLP Tasks’, *Proceedings of the 2020 Federated Conference on Computer Science and Information Systems, FedCSIS 2020*, 21, pp. 179–183. Available at: <https://doi.org/10.15439/2020F20>.
- GitHub (2022) *vuaphapthuat410 / transformer\_ner*. Available at: [https://github.com/vuaphapthuat410/transformer\\_ner](https://github.com/vuaphapthuat410/transformer_ner).
- Greff, K. *et al.* (2017) ‘LSTM: A Search Space Odyssey’, in *IEEE Transactions on Neural Networks and Learning Systems*. Available at: <https://doi.org/10.1109/TNNLS.2016.2582924>.
- Heikal, M., Torki, M. and El-Makky, N. (2018) ‘Sentiment Analysis of Arabic Tweets using Deep Learning’, *Procedia Computer Science*, pp. 114–122. Available at: <https://doi.org/10.1016/j.procs.2018.10.466>.
- Hewitt, J. and Manning, C.D. (2019) ‘A structural probe for finding syntax in word representations’, *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 1, pp. 4129–4138.
- Howard, J. and Ruder, S. (2018) ‘Universal language model fine-tuning for text classification’, *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, 1, pp. 328–339. Available at: <https://doi.org/10.18653/v1/p18-1031>.
- HuggingFace (no date) *Fine-tune a pretrained model, HuggingFace*. Available at: <https://huggingface.co/docs/transformers/en/training>.
- Inoue, G. *et al.* (2021) ‘The Interplay of Variant, Size, and Task Type in Arabic Pre-trained Language Models’, *WANLP 2021 - 6th Arabic Natural Language Processing Workshop, Proceedings of the Workshop*, pp. 92–104.
- Iwendi, C. *et al.* (2023) ‘Cyberbullying detection solutions based on deep learning architectures’, *Multimedia Systems*, 29(3), pp. 1839–1852. Available at: <https://doi.org/10.1007/s00530-020-00701-5>.
- Kyunghyun Cho Dzmitry Bahdanau Fethi Bougares Holger Schwenk, Y.B. (2014) ‘Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation’, *arXiv:1406.1078* [Preprint]. Available at: <https://arxiv.org/abs/1406.1078>.
- Lewis Tunstall, Leandro von Werra, T.W. (2022) *Natural Language Processing with Transformers, Revised Edition*. Revised Ed. O’Reilly Media, Inc.
- Lim, W.L., Ho, C.C. and Ting, C.Y. (2020) ‘Tweet sentiment analysis using deep learning with nearby locations as features’, *Lecture Notes in Electrical Engineering*, 603(January), pp. 291–299. Available at: [https://doi.org/10.1007/978-981-15-0058-9\\_28](https://doi.org/10.1007/978-981-15-0058-9_28).
- Lin, Y., Tan, Y.C. and Frank, R. (2019) ‘Open Sesame: Getting inside BERT’s

Linguistic Knowledge’, in *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Florence, Italy: Association for Computational Linguistics, pp. 241–253. Available at: <https://doi.org/10.18653/v1/w19-4825>.

Liu, B. (2011) *Web Data Mining*. Second Edi. Springer.

Liu, N.F. *et al.* (2019) ‘Linguistic knowledge and transferability of contextual representations’, *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 1, pp. 1073–1094. Available at: <https://doi.org/10.18653/v1/n19-1112>.

Mahmud, T. *et al.* (2023) ‘Cyberbullying detection for low-resource languages and dialects: Review of the state of the art’, *Information Processing and Management*, 60(5), pp. 1–52. Available at: <https://doi.org/10.1016/j.ipm.2023.103454>.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, L.Z. (2018) ‘Deep Contextualized Word Representations’, in. New Orleans, Louisiana: Association for Computational Linguistics, pp. 2227–2237.

Meta (2023) *Prevent bullying and harassment*, Meta. Available at: <https://about.meta.com/actions/safety/topics/bullying-harassment>.

Metzger, S. (2022) *A Beginner’s Guide to Tokens, Vectors, and Embeddings in NLP*. Available at: <https://medium.com/@saschametzger/what-are-tokens-vectors-and-embeddings-how-do-you-create-them-e2a3e698e037>.

Mitsu, R. and Dawood, E. (2022) ‘Cyberbullying: An Overview’, *Indonesian Journal of Global Health Research*, 4(1), pp. 195–202. Available at: <https://doi.org/10.37287/ijghr.v4i1.927>.

Moonat, D. (2022) *HuggingFace Transformer Model Using Amazon Sagemaker, Analytics Vidhya*. Available at: <https://www.analyticsvidhya.com/blog/2022/01/huggingface-transformer-model-using-amazon-sagemaker/>.

El Moubtahij, H., Abdelali, H. and Tazi, E.B. (2022) ‘AraBERT transformer model for Arabic comments and reviews analysis’, *IAES International Journal of Artificial Intelligence*, 11(1), pp. 379–387. Available at: <https://doi.org/10.11591/ijai.v11.i1.pp379-387>.

Mouheb, D. *et al.* (2019) ‘Detection of arabic cyberbullying on social networks using machine learning’, *Proceedings of IEEE/ACS International Conference on Computer Systems and Applications, AICCSA*, 2019-Novem, pp. 1–5. Available at: <https://doi.org/10.1109/AICCSA47632.2019.9035276>.

Muneer, A. and Fati, S.M. (2020) ‘A comparative analysis of machine learning techniques for cyberbullying detection on twitter’, *Future Internet*, 12(11), pp. 1–21. Available at: <https://doi.org/10.3390/fi12110187>.

Nada, A.M.A. *et al.* (2020) ‘Arabic Text Summarization Using AraBERT Model Using Extractive Text Summarization Approach’, *International Journal of Academic Information Systems Research (IJAIRS)*, 4(8), pp. 6–9.

Nasteski, V. (2017) ‘An overview of the supervised machine learning methods’,



- Horizons.B*, 4(December 2017), pp. 51–62. Available at: <https://doi.org/10.20544/horizons.b.04.1.17.p05>.
- Och, F.J. and Ney, H. (2004) ‘The Alignment Template Approach to Statistical Machine Translation’, *Association for Computational Linguistics*, 30(November 2002).
- Patchin, J. W. & Hinduja, S. (2023) *Summary of Our Cyberbullying Research (2007-2021)*, Cyberbullying Research center. Available at: <https://cyberbullying.org/summary-of-our-cyberbullying-research>.
- Patchin, J.W. and Hinduja, S. (2006) ‘Bullies Move Beyond the Schoolyard: A Preliminary Look at Cyberbullying’, *Youth Violence and Juvenile Justice*, 4(2), pp. 148–169. Available at: <https://doi.org/10.1177/1541204006286288>.
- Paula, A. *et al.* (2023) *Transformers and Ensemble methods: A solution for Hate Speech Detection in Arabic languages*, *arXiv.2303.09823*. [arXiv.2303.09823](https://arxiv.org/abs/2303.09823). Available at: <https://arxiv.org/abs/2303.09823>.
- Pelicon, A. *et al.* (2021) ‘Investigating cross-lingual training for offensive language detection’, *PeerJ Computer Science*, 7, pp. 2–39. Available at: <https://doi.org/10.7717/peerj-cs.559>.
- Pirinççi, M. (2022) *How to know which layers should i freeze in deep learning architectures for transfer learning?*, *ResearchGate*. Available at: [https://www.researchgate.net/post/How\\_to\\_know\\_which\\_layers\\_should\\_i\\_freeze\\_in\\_deep\\_learning\\_architectures\\_for\\_transfer\\_learning](https://www.researchgate.net/post/How_to_know_which_layers_should_i_freeze_in_deep_learning_architectures_for_transfer_learning).
- Rachid, B.A., Azza, H. and Ben Ghezala, H.H. (2020) ‘Classification of Cyberbullying Text in Arabic’, *Proceedings of the International Joint Conference on Neural Networks [Preprint]*. Available at: <https://doi.org/10.1109/IJCNN48605.2020.9206643>.
- Radford, A. *et al.* (2019) *Language Models are Unsupervised Multitask Learners*, *semanticscholar*. Available at: <https://www.semanticscholar.org/paper/Language-Models-are-Unsupervised-Multitask-Learners-Radford-Wu/9405cc0d6169988371b2755e573cc28650d14dfe>.
- Radford, A. and Salimans, T. (2018) *Improving Language Understanding by Generative Pre-Training*, *open ai*. Available at: [https://cdn.openai.com/research-covers/language-unsupervised/language\\_understanding\\_paper.pdf](https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf).
- Rogers, A., Kovaleva, O. and Rumshisky, A. (2020) ‘A primer in bertology: What we know about how bert works’, *Transactions of the Association for Computational Linguistics*, 8, pp. 842–866. Available at: [https://doi.org/10.1162/tacl\\_a\\_00349](https://doi.org/10.1162/tacl_a_00349).
- Rush, A.M. and Weston, J. (2015) ‘A Neural Attention Model for Abstractive Sentence Summarization’, in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal, pp. 379–389.
- Saini, H. *et al.* (2023) ‘Enhancing cyberbullying detection: a comparative study of ensemble CNN–SVM and BERT models’, *Social Network Analysis and Mining*, 14(1), p. 1. Available at: <https://doi.org/10.1007/s13278-023-01158-w>.
- Salomon, Palé & Kechaou, Zied & Wali, A. (2022) ‘Arabic hate speech detection system based on AraBERT’, in. *IEEE 21st International Conference on Cognitive Informatics & Cognitive Computing (ICCI\*CC)*, pp. 208–213. Available at: [https://doi.org/10.1109/ICCI\\*CC57084.2022.10101577](https://doi.org/10.1109/ICCI*CC57084.2022.10101577).

Sedat Dogan, G.K. (2024) *Facebook Scraper: How to Scrape Facebook in 2024, AIMultiple*. Available at: <https://research.aimultiple.com/facebook-scraping/>.

Sepp Hochreiter, J.S. (1997) 'Long Short-Term Memory', *Neural Comput* 1997, 9(8), pp. 1735–1780.

Statistics, P.C.B. of (2018) -2017-2019. Available at: [https://www.pcbs.gov.ps/Portals/\\_Rainbow/Documents/المؤشرات الرئيسية لتكنولوجيا المعلومات والاتصالات في فلسطين، 2017-2019](https://www.pcbs.gov.ps/Portals/_Rainbow/Documents/المؤشرات الرئيسية لتكنولوجيا المعلومات والاتصالات في فلسطين، 2017-2019). Available at: [https://www.pcbs.gov.ps/Portals/\\_Rainbow/Documents/المؤشرات الرئيسية لتكنولوجيا المعلومات والاتصالات حسب سنوات مختارة.html](https://www.pcbs.gov.ps/Portals/_Rainbow/Documents/المؤشرات الرئيسية لتكنولوجيا المعلومات والاتصالات حسب سنوات مختارة.html).

Sulaiman Aftan, H.S. (2023) 'Using the AraBERT Model for Customer Satisfaction Classification of Telecom Sectors in Saudi Arabia', *Brain Sciences*, 13(6), p. 147.

Teng, T.H. and Varathan, K.D. (2023) 'Cyberbullying Detection in Social Networks: A Comparison Between Machine Learning and Transfer Learning Approaches', *IEEE Access*, 11(June), pp. 55533–55560. Available at: <https://doi.org/10.1109/ACCESS.2023.3275130>.

Teoh, K.H. *et al.* (2021) 'Face Recognition and Identification using Deep Learning Approach', *Journal of Physics: Conference Series*, 1755(1). Available at: <https://doi.org/10.1088/1742-6596/1755/1/012006>.

*The AI community building the future.* (2024) *Hugging Face*. Available at: <https://huggingface.co/models?p=1&sort=trending>.

unicef (2023) *Cyberbullying: What is it and how to stop it, unicef*. Available at: <https://www.unicef.org/end-violence/how-to-stop-cyberbullying>.

Vaswani Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, I.P. (2017) 'Attention Is All You Need', in *Conference on Neural Information Processing Systems (NIPS 2017)*. Long Beach, CA, USA. Available at: [https://proceedings.neurips.cc/paper\\_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf).

Vijayan, V.K., Bindu, K.R. and Parameswaran, L. (2017) 'A comprehensive study of text classification algorithms', *2017 International Conference on Advances in Computing, Communications and Informatics, ICACCI 2017*, 2017-Janua, pp. 1109–1113. Available at: <https://doi.org/10.1109/ICACCI.2017.8125990>.

Wadhawan, A. (2021) 'AraBERT and Farasa Segmentation Based Approach for Sarcasm and Sentiment Detection in Arabic Tweets', *WANLP 2021 - 6th Arabic Natural Language Processing Workshop, Proceedings of the Workshop*, 1, pp. 395–400.

*Welcome To Colab* (2024) *google*. Available at: <https://colab.research.google.com/#scrollTo=Wf5KrEb6vrkR>.

WissamAntoun (2022) *aub-mind / arabert, github*. Available at: <https://github.com/aub-mind/arabert>.

Yang, J. *et al.* (2020) 'Using deep learning to detect defects in manufacturing: A comprehensive survey and current challenges', *Materials*, 13(24), pp. 1–23. Available at: <https://doi.org/10.3390/ma13245755>.

You, Y. *et al.* (2020) 'Large Batch Optimization for Deep Learning: Training Bert in

76 Minutes’, *8th International Conference on Learning Representations, ICLR 2020* [Preprint].

Yu, L. *et al.* (2017) ‘SeqGAN: Sequence Generative Adversarial Nets with Policy Gradient’, in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)*. San Francisco California USA: AAAI Press, pp. 2852–2858.

Zhang, A. *et al.* (2020) ‘MII: A novel text classification model combining deep active learning with BERT’, *Computers, Materials and Continua*, 63(3), pp. 1499–1514. Available at: <https://doi.org/10.32604/CMC.2020.09962>.

## المخلص

التممر الإلكتروني إحدى القضايا الرئيسية المرتبطة بمنصات التواصل الاجتماعي مثل فيسبوك. ومن المثير للقلق بشكل خاص أنه، على عكس التمر التقليدي، يمكن أن يكون له عواقب عاطفية خطيرة ويتبع الضحايا في كل وقت، وبالتالي فإن الاهتمام بالتممر الإلكتروني أخذ في الازدياد على مواقع التواصل الاجتماعي مثل فيسبوك، الذي يعد أكبر منصات التواصل انتشاراً بين المجتمع العربي بمختلف فئاته.

حققت نماذج اللغة المدربة مسبقاً نجاحاً كبيراً في مجموعة متنوعة من مهام معالجة اللغة الطبيعية. تم إجراء بعض الأبحاث لتحسين نماذج BERT القائمة على الكلمات في الكشف عن التمر الإلكتروني وإنشاء العديد من النماذج العربية المدربة مسبقاً للحصول على نتائج مقبولة، بالرغم من ذلك، هناك اهتمام محدود للكشف عن التمر الإلكتروني باللغة العربية مقارنة باللغة الانجليزية بسبب قلة الموارد المتاحة باللغة العربية.

يهدف هذا البحث إلى دراسة فعالية استخدام AraBERT ، وهو نموذج لغوي عربي مدرب مسبقاً، في الكشف عن التعليقات الساخرة باللغة العربية. قمنا بإنشاء مجموعة بيانات متوازنة من 20,000 تعليق على فيسبوك باللغة العربية تم تصنيفها يدوياً على أنها تتمر أو غير تتمر. استخدمنا تقنيات Fine-tuning لتدريب AraBERTv2 مع مهمة الكشف عن التمر الإلكتروني. من خلال التجربة مع تقنية Freezing layer وإلغاء تجميد طبقات مختلفة من النموذج، استكشفنا التوازن بين الاستفادة من المعرفة المدربة مسبقاً AraBERT وتدريب النموذج مع المهمة المحددة. تُظهر نتائجنا أن ضبط جميع الطبقات في AraBERTv2 حقق أعلى أداء، وحقق نتائج أفضل حيث كانت الدقة 91.9% وكان مقياس F1 ما نسبته 93.0%.