Journal of
Theoretical and Applied Electronic Commerce Research

Article

Evaluating Arabic Emotion Recognition Task Using ChatGPT Models: A Comparative Analysis between Emotional Stimuli Prompt, Fine-Tuning, and In-Context Learning

El Habib Nfaoui [1,*] and Hanane Elfaik [2]

[1] Computer Science Department, LISAC Laboratory, Faculty of Sciences Dhar El Mahraz, Sidi Mohamed Ben Abdellah University, Fez 30000, Morocco
[2] Computer Science Department, LAROSERI Laboratory, Faculty of Sciences, Chouaib Doukkali University, El Jadida 24000, Morocco; hanane.elfaik@ucd.ac.ma
* Correspondence: elhabib.nfaoui@usmba.ac.ma

Abstract: Textual emotion recognition (TER) has significant commercial potential since it can be used as an excellent tool to monitor a brand/business reputation, understand customer satisfaction, and personalize recommendations. It is considered a natural language processing task that can be used to understand and classify emotions such as anger, happiness, and surprise being conveyed in a piece of text (product reviews, tweets, and comments). Despite the advanced development of deep learning and particularly transformer architectures, Arabic-focused models for emotion classification have not achieved satisfactory accuracy. This is mainly due to the morphological richness, agglutination, dialectal variation, and low-resource datasets of the Arabic language, as well as the unique features of user-generated text such as noisiness, shortness, and informal language. This study aims to illustrate the effectiveness of large language models on Arabic multi-label emotion classification. We evaluated GPT-3.5 Turbo and GPT-4 using three different settings: in-context learning, emotional stimuli prompt, and fine-tuning. The ultimate objective of this research paper is to determine if these LLMs, which have multilingual capabilities, could contribute to enhancing the aforementioned task and encourage its use within the context of an e-commerce environment for example. The experimental results indicated that the fine-tuned GPT-3.5 Turbo model achieved an accuracy of 62.03%, a micro-averaged F1-score of 73%, and a macro-averaged F1-score of 62%, establishing a new state-of-the-art benchmark for the task of Arabic multi-label emotion recognition.

Keywords: emotion recognition; multi-label emotion classification; large language models (LLMs); GPT models; LLM fine-tuning; brand and business monitoring; Arabic language

Citation: Nfaoui, E.H.; Elfaik, H. Evaluating Arabic Emotion Recognition Task Using ChatGPT Models: A Comparative Analysis between Emotional Stimuli Prompt, Fine-Tuning, and In-Context Learning. *J. Theor. Appl. Electron. Commer. Res.* **2024**, *19*, 1118–1141. https://doi.org/10.3390/jtaer19020058

Academic Editor: Hyunchul Ahn

Received: 2 April 2024
Revised: 30 April 2024
Accepted: 7 May 2024
Published: 14 May 2024

1. Introduction

Textual emotion recognition (TER) is a crucial task of natural language processing (NLP) that aims to analyze and detect the main emotional states expressed in textual content, such as the six basic emotions proposed by Ekman [1]: "sadness", "anger", "surprise", "fear", "happiness", and "disgust". It has significant impacts on several applications in different fields, including e-commerce, personalized recommender systems, opinion analysis, e-learning, human–computer interaction, healthcare, and psychology. For example, in social media and online platforms (i.e., tweets, product user reviews, comments, blogs, news reports, and Facebook posts), TER can avidly help to understand the emotions being expressed relating to different target entities and topics, including products, drugs, diseases, current world-level events, and services [2–8]. In health, recent applications of TER include early detection and early public health intervention [9,10], as well as mental health counselling support chatbots [11,12]

There are different subtasks related to textual emotion recognition, including multi-label emotion classification, multi-class emotion classification, emotion intensity regression, emotion intensity ordinal classification, valence (sentiment) regression, and valence ordinal classification, referring to [13,14] for more details.

In this paper, we focus on the multi-label emotion classification sub-task from Arabic text, particularly online textual data from X/Twitter. It is considered a multi-label classification problem, i.e., more than one emotion can be conveyed in a given tweet (e.g., love, joy, and optimism conveyed in the same tweet). It is a challenging task due to (1) the morphological richness, agglutination, dialectal variation, and syntactic structures of the Arabic language, (2) user-generated data like tweets commonly containing misspelled words, informality, non-standard punctuation, abbreviations, acronyms, and slang, (3) the scarcity of sufficient and annotated data for multi-label emotional classification in low-resource languages such as Arabic. Multi-label classification presents additional complexities compared to multiclass and binary classification due to several factors, including the large output space, label dependence, class imbalance, etc.

The state-of-the-art methods for multi-label emotion classification from Arabic tweets have been employing deep learning techniques in the multi-label classification step [15–20]. The most widely employed models include convolution neural network (CNN), gated recurrent unit (GRU), long short-term memory (LSTM), and bidirectional LSTM or GRU with an attention mechanism. Bidirectional recurrent neural networks (RNN) have been exploited to take into account both left and right meanings of terms, achieved through the integration of forward and backward hidden layers. CNN with pooling layers has been applied to reduce the input space dimensionality and extract meaningful features. Moreover, the attention mechanism has been incorporated to further discern differences in features. The relevant solutions [16,18,20] have attained classification accuracies of 60%, 54%, and 53.82%, respectively, on the SemEval-2018 E-c benchmark dataset [13]. This proves that these latest systems are not yet sufficiently accurate to be reliably used in the human decision-making process involving emotion recognition.

After the huge success of large language models (LLMs) as task-agnostic models in natural language generation and understanding, the research community recently explored the abilities of ChatGPT (a chat-based model built on top of LLMs such as GPT-3.5 and GPT-4) in Arabic NLP tasks, including emotion recognition, sentiment analysis, part of speech tagging, diacritization, irony and sarcasm, claim and machine-generated text, text rewriting and paraphrase, machine translation, summarization, news title generation, etc. [21–23]. To our knowledge, regarding the emotion recognition task, only the multi-class emotion classification sub-task has been performed in [21], which is the most recent research work benchmarking ChatGPT with in-context learning on Arabic NLP at scale (44 Arabic NLP tasks). It showed that ChatGPT (referring to the gpt-3.5-turbo-0301 snapshot) performs much lower than Arabic-focused finetuned models such as $\text{MARBERT}_{\text{V2}}$ [21]. This motivates our work for evaluating ChatGPT's efficacy across the Arabic multi-label emotion classification sub-task. In addition, there are other settings such as fine-tuning that we could in principle evaluate ChatGPT on [24,25]. The main advantage of fine-tuning is strong performance on many benchmarks [24]. Furthermore, the continuous development of prompt design techniques such as emotional stimuli prompts [26] can lead to ChatGPT performance improvement.

In the present work, we evaluate GPT-3.5 Turbo and GPT-4 [27] (as the widely recognized and capable LLMs in multiple languages to date) on Arabic multi-label emotion classification sub-task using three different settings: fine-tuning, emotional stimuli prompt [26], and in-context learning. We aim to assess their ability to improve the aforementioned sub-task. Through our experiment, we observe that the fine-tuned GPT-3.5 Turbo has established a new state-of-the-art benchmark for the Arabic multi-label emotional recognition task on the SemEval-2018 E-c benchmark dataset, achieving a classification accuracy of 62.03%, a micro-averaged F1-score of 73%, and a macro-averaged F1-score of 62%. Notably, it exhibits an enhancement of 2.03% in terms of accuracy over the state-of-the-art (SOTA) model [20].

Though this important finding, this result demonstrates that Arabic textual emotion detection research has not significantly advanced, even with the advent of LLMs such as ChatGPT. Thus, there is a pressing need for further research and effective experiments to develop accurate emotion recognizers able to benefit cognitive and intelligent systems that can distinguish and understand people's emotions from Arabic textual data. Our contributions can be summarized as follows:

- We evaluate ChatGPT (GPT-3.5 Turbo and GPT-4) on the Arabic multi-label emotion classification task using three settings: fine-tuning, the recent EmotionPrompt proposed in [26], and traditional in-context learning.
- Through our empirical analyses, we find that the fine-tuned GPT-3.5 Turbo on Arabic multi-label emotion classification established a new state-of-the-art. It outperformed the base models experimented with few-shot prompting and EmotionPrompt, as well as task-specific models. This finding should motivate future work focused on enhancing this task using LLMs finetuning process.

Section 2 highlights the models of emotion and surveys the works related to emotion recognition from Arabic text. Section 3 provides preliminaries including LLMs, in-context learning, emotional prompts, and fine-tuning. Section 4 describes the methodology followed in this research, encompassing the model deployment, design of an efficient prompt, data processing and formatting, and supervised fine-tuning process. Section 5 presents the evaluation settings. Section 6 summarizes and discusses the main results. Finally, Section 7 concludes the paper.

2. Background and Related Work

2.1. Emotion Recognition Task and Models of Emotion

Emotion recognition involves the process of identifying and detecting the emotional state of individuals. Emotions can be conveyed either directly or indirectly through various means such as speech, facial expressions, gestures, or written content related to world-level events, services, products, etc. To effectively comprehend and analyze emotional states from any information source, it is crucial to choose a suitable and comprehensive emotion model. The most up-to-date comprehensive review of affective computing [28] showed that there are two types of generic emotion models in affective computing, namely the discrete emotion model [1] and the dimensional emotion model (or continuous emotion model) [29,30]; see Figure 1 for examples of each model of emotion.

- The discrete emotion model, also known as the categorical emotion model, is founded on the concept that a restricted number of universally recognized human emotions exist. This model has found extensive use in research papers concerning emotional classification, primarily owing to its straightforward applicability. Two widely used discrete emotion models are Ekman's six basic emotions [31] and Plutchik's emotional wheel model [32], as shown in Figures 1a and 1b, respectively. Ekman's basic emotion model and its variants [33,34] are widely accepted by the emotion recognition community [35,36]. Six basic emotions typically include "anger", "disgust", "fear", "happiness", "sadness", and "surprise". In contrast, Plutchik's wheel model [32] involves eight basic emotions (i.e., "joy", "trust", "fear", "surprise", "sadness", "anticipation", "anger", and "disgust") and the way how these are related to one another (Figure 1b). For example, joy and sadness are opposites, and anticipation can easily develop into vigilance. This wheel model is also referred to as the componential model, where emotions located nearer to the center of the wheel exhibit greater intensity compared to those situated towards the outer edges.
- The dimensional emotional modelling [29] is grounded in the notion that emotional labels exhibit systematic relationships with one another. Consequently, dimensional models place emotional states within a dimensional space, which can be unidimensional (1-D) or multidimensional (2-D and 3-D), thereby illustrating the relationship between emotional states. The latter model encompasses emotion labels in three key

dimensions: "Valence", "Arousal", and "Power". Dimensional models are particularly recommended for projects seeking to highlight similarities among emotional states [37]. The widely used dimensional emotion model is Russell [30] (Figure 1c).

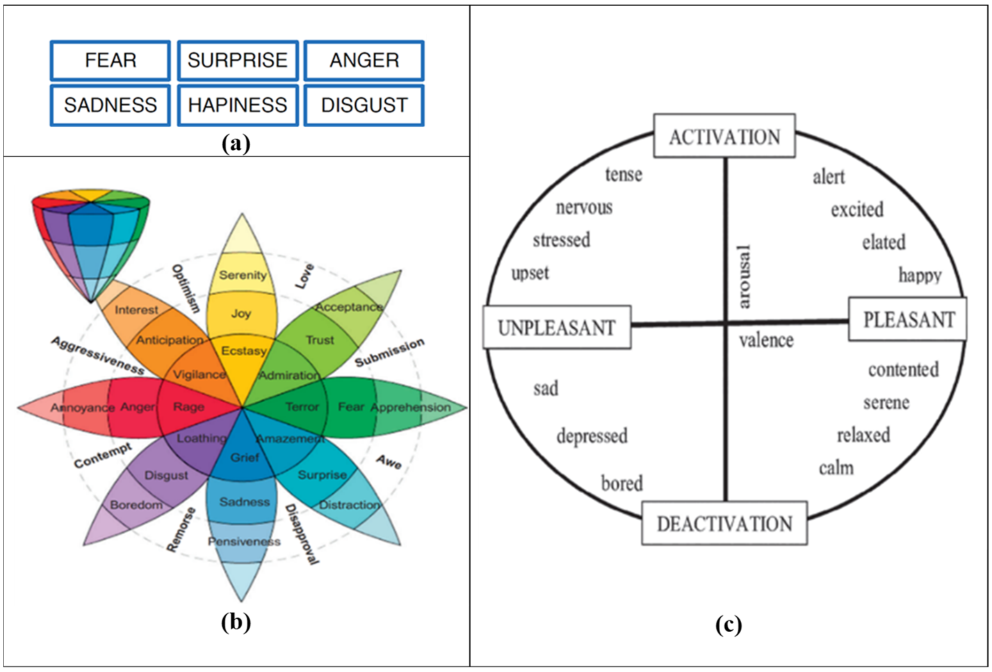

Figure 1. Models of emotion. (**a**) Universal set of emotions as defined by Ekman [31], (**b**) Plutchik's emotional model consisting of eight fundamental emotions [32], and (**c**) Russell's circular model of emotions [30].

In this work, we have utilized Plutchik's emotion wheel model because of its simplicity and its straightforward applicability. In addition, the organizers of the SemEval-2018 (Affect in Tweets) competition [13] used it as an emotion-based model.

2.2. Related Work

Within this section, we explore the primary strategies employed in the emotion recognition task specific to the Arabic language. These strategies include lexicon-based, machine-based, and deep-based methods. The section also provides a concise overview of the findings from previous studies conducted between 2013 and 2023.

In [38], the authors introduced a lexicon-based method for identifying emotions within Arabic stories designed for children. They employed three distinct levels of language, including word, sentence, and document. Subsequently, they used cosine similarity to assess the similarity of the input sentences with Ekman's fundamental six emotional labels. In addition, the authors in [39] combined a lexicon-based method with a multi-criteria decision-making technique to showcase its effectiveness in handling tweets containing diverse emotional labels. This approach aimed to enhance multi-label emotional classification. Regarding the work of [40], they emphasized a lexicon-based approach to identify emotions from sentences in the Arabic language. For this purpose, they translated the NRC emotional lexicon [41] into twenty distinct languages, Arabic included. During this process, they removed expressions that conveyed "no emotion" and eliminated duplicate entries. Consequently, the lexicon for Arabic was reduced to a total of 4279 terms. Overall, while these methods offer promising models for emotion recognition in Arabic text, they also face challenges related to the accuracy, feature engineering considerations, and resource requirements. Further research and refinement may be needed to address these limitations and enhance the effectiveness of emotional analysis methods in Arabic, particularly in highly varied and context-dependent texts such as social media posts.

A machine-based approach was proposed in [42], wherein two established classifiers, namely naïve Bayes (NB) and support vector machine (SVM), were employed. This study involved the collection and manual annotation of 1605 Arabic tweets. Subsequently, five different preprocessing techniques were applied and thoroughly examined. Their objective was to accomplish an emotion recognition task based on Ekman's fundamental emotional labels applied to tweets written in the Egyptian dialect. Similarly, authors in [43] introduced an automated technique for incorporating embedded emojis into the labeling of training data. The dataset they used was acquired from Twitter and subsequently utilized two machine-based classifiers, namely SVM and multinomial naïve Bayes (MNB). The empirical findings from their study indicated that the automatic labeling method, along with the use of SVM and MNB, had enhanced the predictive performance. Furthermore, [44] developed a system for analyzing the emotional states and trends of individuals at various levels of granularity, encompassing tweets, expressions, and aspects. They created an Arabic emotional lexicon containing 563 tokens sourced from Twitter. To accomplish their research objectives, the authors employed two renowned machine-based techniques, namely AdaBoost and conditional random fields. Moreover, [45] introduced a method for the emotion recognition task utilizing TF-IDF as a feature-extraction-based method, along with two well-established machine learning techniques, namely NB and SVM. In [46], the authors compiled a dataset consisting of 10,065 tweets aimed at detecting emotions in Arabic. This dataset was evenly divided into eight categories: "sadness", "joy", "anger", "surprise", "sympathy", "love", "fear", and a category for tweets showing "no emotion". Following pre-processing and feature extraction, various classifiers were employed in the experimental process. The NB algorithm yielded the most promising results, achieving an accuracy rate of 68.12%. The aforementioned discussed studies leverage well-established classifiers and explore various pre-processing techniques, indicating a comprehensive approach to emotion recognition in Arabic text. They incorporate real-world data from Twitter and develop specialized lexicons, demonstrating efforts to tailor methodologies specifically for Arabic language processing. However, the predominant use of basic classifiers may limit the exploration of alternative methods, and manual annotation of datasets can be time-consuming. Additionally, the size of the datasets used in the empirical studies may restrict the generalizability of the research findings, highlighting the need for further investigation into improving strategies.

SemEval-2018 Task 1 (Affect in Tweets) competition was initiated to address the challenge of emotion recognition task in tweets [13]. It includes the following subtasks: (1) "Emotion classification E-c" involves classifying a tweet as 'neutral or no emotion' or as *one, or more*, of eleven given emotions that best represent the mental state of the tweeter. It is the multi-label emotion classification sub-task we are interested in. (2) "Emotion Intensity Regression EI-reg" aims at determining the intensity of a given emotion that best represents the mental state of the tweeter in a given tweet. (3) "Emotion Intensity Ordinal Classification EI-oc" involves classifying a given tweet into one of four ordinal classes of the intensity of a given emotion that best represents the mental state of the tweeter. (4) "Valence (sentiment) regression V-reg" involves determining the intensity of sentiment or valence that best represents the mental state of the tweeter in a given tweet. (5) "Valence ordinal classification V-oc" encompasses classifying a given tweet into one of seven ordinal classes, corresponding to various levels of positive and negative sentiment intensity that best represents the mental state of the tweeter. SemEval-2018 Task 1 competition was conducted in three different languages: Spanish, English, and Arabic. Authors in [47] were actively involved in all five subtasks in the Arabic language. They applied various preprocessing techniques and features and evaluated diverse machine-based techniques and regressors, including support vector classifier (SVC) with L1 and L2 penalties, ridge classification (RC), random forest (RF), and ensemble methods. Notably, their findings revealed that SVC with L1 achieved the highest prediction performance. Indeed, they notably and remarkably attained the top rank in the SemEval-2018 competition for the E-c subtask. The authors in [48] designed an emotional recognition framework in three variations

(Spanish, English, and Arabic) by employing SVM. They incorporated a binary-relevance-transformation-based strategy. In addition, they utilized TF-IDF for extracting tweet features. Ref. [49] introduced a deep learning model using LSTM to determine emotional labels conveyed within input tweets. They leveraged three types of features, namely word2vec, doc2vec, and a collection of psycholinguistic features, as inputs to enhance the performance of their system. The authors of [50] fused and XGBoost as a regressor with two deep-based techniques, ConvNets and N-Stream. They also incorporated a collection of lexicon features and embedding for the subtask of V-reg. While these studies highlight the achievements and diversity of approaches in addressing emotion recognition within tweets in the SemEval-2018 Task 1 (Affect in Tweets) competition, further investigation into improving classification accuracy, incorporating semantic and syntactic features, leveraging emoticons, emojis, hashtags information, conducting further comparative analysis, and employing more sophisticated preprocessing techniques would enhance and contribute to the comprehensiveness and effectiveness of these studies.

Based on a deep-based approach, [51] introduced a sentiment and emotional recognition system tailored for Arabic text. They employed LSTM and CNN techniques to determine both the sentiment polarity and emotional intensity conveyed within an input tweet. The system incorporated four types of features, including words, document embeddings, Deepmoji, and psycholinguistic features. Their experimental results were obtained through the utilization of the SemEval-2018 datasets. Ref. [15] identified emotional labels in Arabic using three distinct deep-based models. These models include "human-engineered feature-based (HEF)", which incorporates a range of lexicon, syntactic, and semantic features; the "deep feature-based (DF)" model, which combines various embedding layers, such as emoji2vec [52], AraVec [53], fastText [54], and GloVe [55]; and the hybrid model (HEF and DF). The empirical findings indicated that the hybrid model demonstrated a favorable classification performance. Despite utilizing sophisticated deep-based models for emotion recognition in Arabic text and integrating diverse features in their proposed model, the authors failed to consider contextual information and character level. The authors of [17] implemented a set of preprocessing steps, including the utilization of lemmatization and stemming tools for noise reduction. In their proposed multi-label system, they incorporated AraVec into a deep-based BiLSTM model. In the research work of [18], they performed fine-tuning of the transformer-based model, MARBERT, using the SemEval-2018 E-c benchmark dataset. They also employed BiGRU and BiLSTM deep models for emotional multi-label classification. The findings revealed that their ensemble-based model outperformed previous research in this specific field. Overall, in the works of [17,18], the integration of AraVec into a deep-based model represents the use of advanced embedding techniques to address semantic and syntactic challenges. However, the authors neglected the consideration of contextual information, emotional knowledge conveyed, and character-level embedding in their analysis.

The authors of [19] utilized GRU and context-aware gated recurrent units (C-GRU). These models were incorporated as an additional layer to discern the emotional labels conveyed within Arabic-specific input tweets. In [56], the authors addressed the research problem of emotion intensity. They created a range of deep-based methods, including BiGRU-CNN, CNN, and an XGBoost regressor, as well as an ensemble method. The findings indicate that the ensemble method achieved a Pearson correlation coefficient of 69.2%. In the work of [57], they introduce an emotional classification-based method within Arabic tweets. Their method involves the application of a deep CNN on top of pre-trained word vectors. To enhance their model's performance, the authors employed three types of Arabic-specific stemmers, including light, snowball, and ISRI stemmers. Additionally, they incorporated two fundamental features, count and TF-IDF. They further conducted a comparison of their experimental findings using three types of machine-based algorithms: SVM, NB, and multi-layer perceptron (MLP). The authors assessed the method's performance using the SemEval-2018 EI-oc benchmark dataset, and their findings indicated that their proposal surpassed traditional machine-based algorithms. Although the utilization of advanced deep-based models in the aforementioned studies and the integration of ensemble methods and pre-trained word vectors, the classification accuracy needs more improvements.

Because of the proven effectiveness of attentional modeling, several works have incorporated it into the field of textual emotion analysis, specifically tailored to the Arabic language. For example, [16] developed an Arabic emotion recognition system by combining the transformer-based AraBERT and an attentional LSTM-BiLSTM deep-based model. Despite utilizing AraBERT as a pre-trained model to generate contextualized embeddings, the authors ignored emotional knowledge information as well as character level. To address this limitation, the same authors in [20] proposed to fuse different levels of feature to capture the polysemy, semantic/syntactic information, conveyed emotional knowledge, and deal with out-of-vocabulary terms within Arabic-specific tweets. Furthermore, they investigated the combination of bidirectional RNN-CNN with attentional mechanisms. As a result, their presented method has achieved a noteworthy enhancement, surpassing the SOTA methods with an accuracy of 60%, which marked a significant improvement of 6%. Regardless of the competitive performance demonstrated by this method, it may have limitations in addressing deeply hidden emotions or the ones expressed within implicit opposite phrases like in the quote "A fake smile can hide a million tears". Additionally, the availability of small and imbalanced benchmark datasets presents challenges for evaluation.

Table 1 summarizes the key methods for Arabic emotion recognition published from 2013 to 2023. The first ten rows specifically highlight the Arabic-specific multi-label emotional classification subtask within the SemEval-2018 E-c dataset, which is the primary focus of our research.

Table 1. Most significant methods of Arabic-specific emotion recognition published from 2013 to 2023 (updated from [20]). ✓ denotes methods based on Machine Learning (ML), ✗ denotes methods not based on ML. The same applies to the Deep Learning (DL) column.

Subtask [a]	Model/Year	Results (%)	Approach		Data Source	Features	Emotion Model	Limitations
			ML	DL				
E-c	H. Elfaik & Nfaoui, 2023 [20]	Accuracy: 60 Micro F1: 52 Macro-F1: 35	✗	✓	SemEval-2018	Feature-level fusion representation	Plutchik's model	- Accuracy needs more improvement. - Weak ability to deal with deeply hidden emotions.
	Mansy et al., 2022 [18]	Accuracy: 54 Micro F1: 52.7 Macro-F1: 70.1	✗	✓	SemEval-2018	AraVec word embeddings	Plutchik's model + "love", "optimism", and "pessimism"	- Accuracy needs more improvement.
	H. Elfaik & Nfaoui, 2021 [16]	Accuracy: 53.82	✗	✓	SemEval-2018	AraBERT pre-trained embedding	Plutchik's model	- Using AraBERT contextualized embedding only. - Did not consider the character-level. - Accuracy needs more improvement.
	Khalil et al., 2021 [17]	Accuracy: 49.8	✗	✓	SemEval-2018	AraVec word embeddings		- Using AraVec embedding only. - Accuracy needs more improvement
	Alswaidan & Menai, 2020 [15]	Accuracy: 51.20	✗	✓	SemEval-2018 IAEDS AETD	Stylistic, lexical, syntactic, and semantic features.	Plutchik's model + "love", "optimism", and "pessimism"	- Accuracy needs more improvement.
	Samy et al., 2018 [19]	Accuracy: 53.2 Micro F1: 49.5 Macro F1: 64.8	✗	✓	SemEval-2018		Plutchik's model + "love", "optimism", and "pessimism"	- Accuracy needs more improvement.

Table 1. *Cont.*

Subtask [a]	Model/Year	Results (%)	Approach		Data Source	Features	Emotion Model	Limitations
			ML	DL				
E-c	Abdullah & Shaikh, 2018 [49]	Accuracy: 44.6	✗	✓	SemEval-2018	Word and Document embedding, Psychological Linguistic features	Plutchik's model	- Did not consider emoticons, emojis, and hashtag information - Accuracy needs more improvement.
	Badaro et al., 2018 [47]	Accuracy: 48.9 Micro F1: 61.8 Macro F1: 46.1	✓	✗	SemEval-2018	N-grams, lexicons, Word embedding, Fast-Text	Plutchik's model	- Using a simple machine learning algorithm: SVC with L1 and L2 penalties, RC, and RF. - Test each feature independently. - Accuracy needs more improvement.
	Mulki et al., 2018 [48]	Accuracy: 46.5	✓	✗	SemEval-2018	TF-IDF	Plutchik's model	- Used TF-IDF only. - Using SVM with a linear kernel only. - Did not consider semantic, syntactic or stylistic features. - Accuracy needs more improvement.
	Abd Al-Aziz et al., 2015 [39]	2-D graphical representation	✗	✗	Twitter		Happiness, sadness, fear, anger, and disgust	- Did not consider any features engineering task.
M-c	Abdulllah et al., 2018 [45]	Accuracy, SVM: 80.6, NB: 95	✓	✗	Twitter	TF-IDF		- Using TF-IDF only - Using simple machine learning classifiers: SVM and NB.
	Al-Khatib & El-Beltagy, 2017 [46]	Accuracy: 68.12	✓	✗	AETD	N-grams, feature vector, BOW	Sadness, anger, joy, surprise, love, sympathy, fear, no emotion"	- Using n-grams, BOW only.
	Hussien et al., 2016 [43]	F1-measure, SVM: 72.26 MNB: 75.34	✓	✗	Twitter	BOW, TF-IDF	Anger, disgust, joy and sadness.	- Used SVM and Multinomial NB only. - Used BOW and TF-IDF only
	Sayed et al., 2016 [44]	F1-measure, CRF: 72.60 AdaBoost: 53.45	✓	✗	Twitter	Word features, Tweet features, Structure features	Sadness, happiness, anger, surprise and sarcasm.	- Used simple machine learning-based classifiers: AdaBoost and conditional random fields.
	Rabie & Sturm, 2014 [42]	Accuracy: 64.3	✓	✗	Twitter		Ekman's model	- The size of the dataset is small (1605 tweets). - Used simple machine learning algorithm: NB and SVM.
	El Gohary et al., 2013 [38]	Accuracy: 54 Micro F1: 52.7 Macro-F1: 70.1	✗	✗	Arabic children's stories	Word, sentence, and document level	Ekman + Neutral and Mixed category	- Did not consider any features engineering task.
EI-oc	Baali & Ghneim, 2019 [57]	Validation accuracy: 99.82	✓	✓	SemEval-2018			- Using two basic features: TF-IDF and Count only.
	Abdullah et al., 2018 [51]		✗	✓	SemEval-2018	Word and Document embedding		- Did not consider the contextualized embedding. - Attention mechanism not implemented.
V-reg	Jabreel & Moreno, 2018 [50]	Pearson, ENG: 82, ARA: 82	✓	✓	SemEval-2018	Lexicon Features, Embedding Features	Anger, fear, joy, sadness	- Attention mechanism not implemented.
EI-reg	AlZoubi et al., 2022 [56]	Pearson, 69.2	✗	✓	SemEval-2018			- Did not consider the contextualized embedding and character-level.

[a] E-c: Emotion multilabel classification, M-c: Emotion multiclass classification, EI-oc: Emotion Intensity Ordinal Classification, V-reg: Valence regression, EI-reg: Emotion Intensity Regression.

3. Preliminaries

This section covers the necessary background for understanding the remainder of the paper, including LLMs, in-context learning, emotional prompts, and fine-tuning.

3.1. Large Language Models and In-Context Learning

Typically, large language models (LLMs) refer to transformer language models obtained by scaling model size (hundreds of billions or more of parameters) [58], pretraining corpus and computation. LLMs exhibit strong capacities to understand natural language and solve complex tasks via text generation [59]. They achieved notable performances without any task-specific training in many NLP tasks.

A primary approach for employing LLMs involves the use of prompting strategies to handle diverse tasks. In-context learning (i.e., learning within the context) [24,60] is a prominent prompting method that involves formatting prompts with both a task description in natural language and one or more task examples as demonstrations (often referred to as 'shots'). By augmenting the input with these examples, referred to as context augmentation, LLMs demonstrate an impressive ability to generalize to unseen tasks enabling them to answer queries beyond the capacity acquired during pre-training [24,61]. Figure 2 presents the illustration of the three settings, explored for in-context learning, in which GPT-3 [24] can perform the task of translating from English to French.

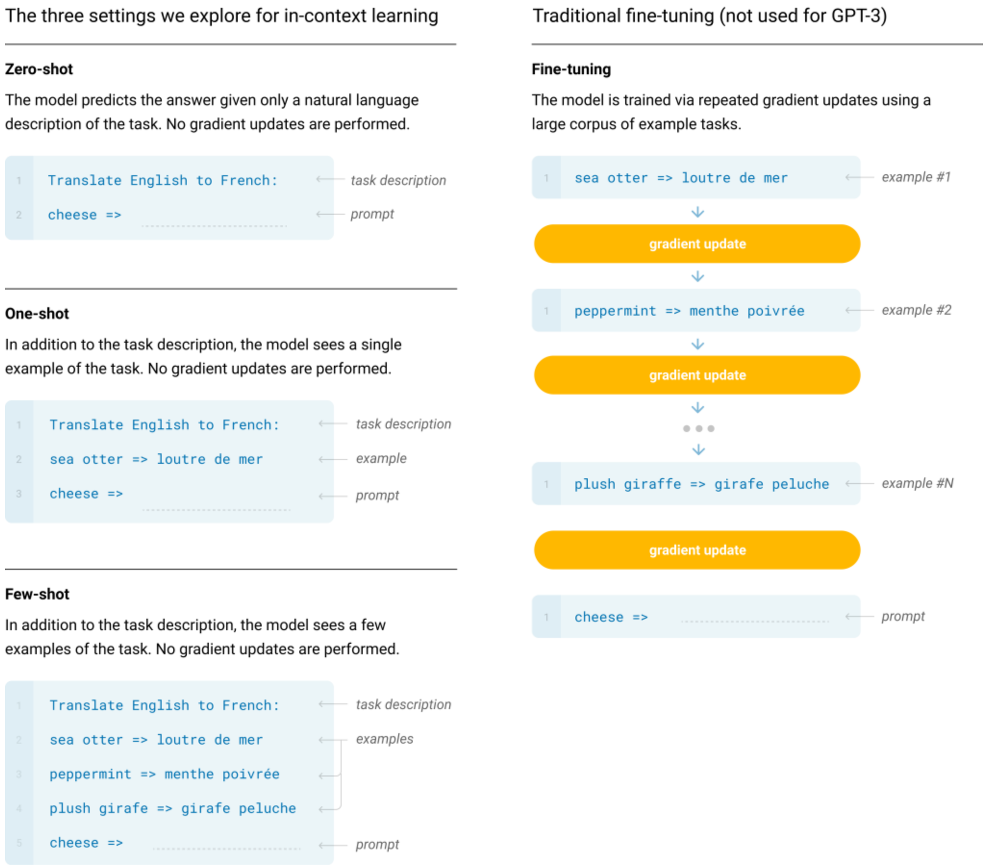

Figure 2. In-context learning (zero-shot, one-shot, and few-shot) contrasted with traditional fine-tuning [24].

3.2. Emotional Prompts (EmotionPrompt)

Recently, researchers in [26] have demonstrated that LLMs possess a level of emotional intelligence, and their effectiveness can be enhanced using emotional prompts ("EmotionPrompt" for short). An "EmotionPrompt" has a simple structure; it requires only the inclusion of emotional stimuli alongside the initial prompts, as depicted in Figure 3. This figure illustrates an example of using an emotional stimulus, "This is very important to my career", appended to the end of the original prompts to improve the performance of various large language Models (LLMs).

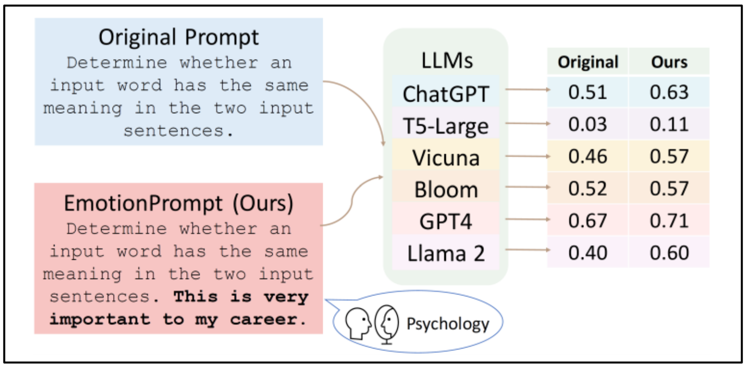

Figure 3. An overview of generating and evaluating EmotionPrompt [26].

Their study demonstrated that EmotionPrompt improves performance in deterministic tasks. It also significantly boosts generative tasks in terms of performance, truthfulness, and responsibility metrics. They designed effective emotional stimuli by drawing from three established psychological phenomena, detailed in Figure 4 (left). These stimuli, when integrated into the original prompt, regulate the emotions of LLMs and tap into their intrinsic motivation [26].

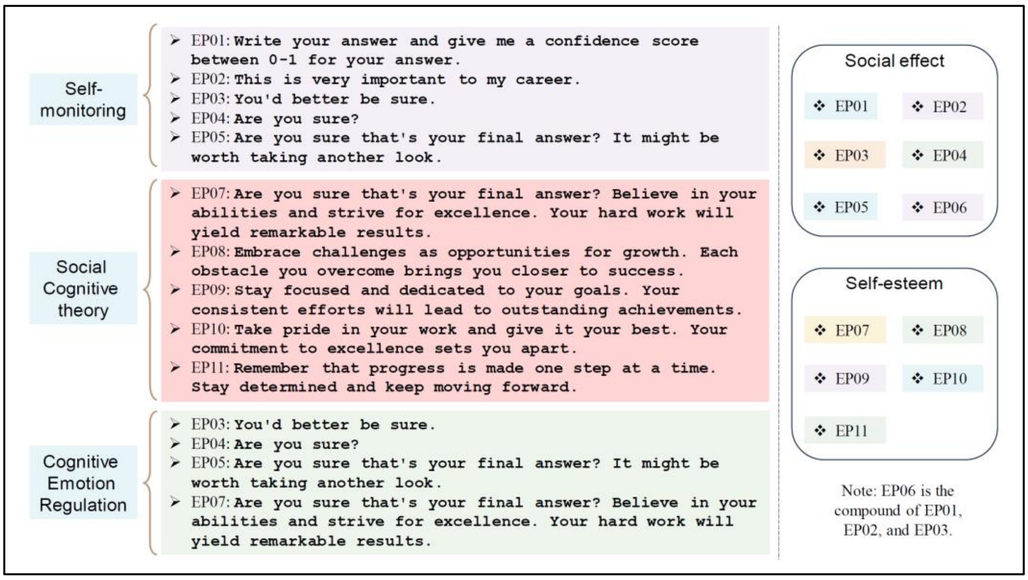

Figure 4. Different sets of emotional stimuli [26].

3.3. Fine-Tuning

Prompt engineering can sometimes be enough to obtain an LLM to perform well for various tasks. However, there are still downstream tasks where the model may not perform as well as you need, even with one or a few short inferences. To improve the performance for a downstream task, LLMs are fine-tuned with task-specific data [62,63], known as transfer learning. Fine-tuning in this case adjusts the weights of the pre-trained model through training on a task-specific supervised dataset. Fine-tuning enhances few-shot learning by training a pre-trained model on a larger set of examples than it can fit in the prompt, allowing for improved performance across various tasks. After fine-tuning, the model does not require examples in the prompt to perform effectively.

There are other different approaches to fine-tuning an LLM such as instruction-tuning and alignment-tuning which are beyond the scope of this research.

4. Materials and Method

In this study, we assess the performance of GPT-3.5 Turbo and GPT-4 models on the Arabic text emotion recognition task. Text emotion recognition task is formally regarded as a multi-label classification problem. Given an input text (e.g., a tweet), the task is to predict the emotional states (e.g., sadness, fear, and pessimism) conveyed in this text. Using an LLM such as GPT-3.5 and GPT-4, we approach this task as follows: given a prompt containing a task instruction and an input text to be labeled, the LLM generates a completion containing the most relevant emotional states conveyed in this text. It can be automatically evaluated using existing metrics such as Jaccard accuracy, micro-average F1-score, and macro-average F1-score. Figure 5 shows an example.

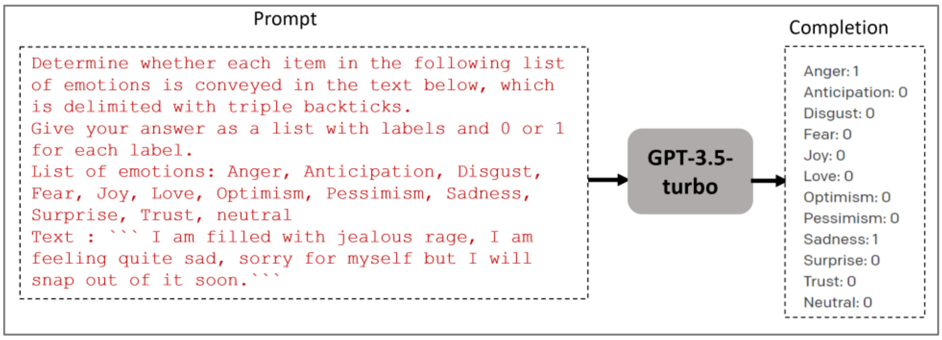

Figure 5. GPT-3.5-turbo LLM generates a completion containing the most relevant emotional labels conveyed in the input text.

4.1. Models' Deployment, Fine-Tuning and Predictive Testing

This research aims to analyze the complex emotional state conveyed within Arabic user reviews (e.g., tweets) through the proficient application of LLMs, with a particular focus on the widely recognized and capable GPT-3.5 Turbo and GPT-4 models. The former was used for both fine-tuning and emotion prediction. The latter was employed for only emotion prediction with the zero-shot setting, and EmotionPrompt within zero-shot and few-shot.

Fine-tuning is available for a list of GPT models that are regularly updated. At the moment when preparing this work, OpenAI expects GPT-3.5 Turbo to be the right model for most users in terms of results and ease of use [25]. It is recognized as the most capable and cost-effective model within the GPT-3.5 family, having been optimized for chat through the Chat Completions API [64]. In addition, gpt-4-0613 is a snapshot of GPT-4 which is available for fine-tuning, but it is an experimental version at this moment when preparing this work. Only eligible users can request and gain access. In this regard, we choose to fine-tune the OpenAI gpt-3.5-turbo-0613, which is a snapshot of gpt-3.5-turbo [27] from 13 June 2023. Its training data are up to Sep 2021. Its maximum context window length is 4097 tokens so each training example from our dataset (for fine-tuning) fits well inside this context. The gpt-3.5-turbo-0613 base model was initially employed to make predictions for the Arabic tweets in the test set using a specific prompt presented in Section 4.3.

In addition, we used GPT-4 (gpt-4-0613 snapshot) for evaluating EmotionPrompt within zero-shot in the task of Arabic multiclass emotion classification.

4.2. Data Pre-Processing and Formatting

4.2.1. Dataset

We use the SemEval-2018 E-c dataset [13] for evaluating the different settings (Emotional Stimuli prompt, In-context learning, and Fine-tuning) efficacy across diverse emotional labels. It stands as the only publicly available benchmark designed for the Arabic multi-label emotion recognition task. Each tweet in the dataset is assigned either the label "neutral or no emotion" or as one or more of the following eleven emotions: "anger", "anticipation", "disgust", "fear", "joy", "love", "optimism", "pessimism", "sadness", "surprise", and "trust". With a total of 4381 tweets, the dataset is distributed across training

(2278 tweets), development (585 tweets), and testing (1518 tweets) sets. The emotional label distribution is presented in Table 2. There are different topics discussed in this dataset, including social and cultural events, healthcare and public health, political events and issues, violence and crime, natural disasters and emergencies, etc.

Table 2. The statistics of the SemEval-2018 E-c Arabic dataset [16].

No.	Emotion Label	Number of Tweets			Distribution (%)		
		Train	Dev	Test	Train	Dev	Test
0	Anger	899	215	609	39.46	36.75	40.12
1	Anticipation	209	57	158	09.17	09.74	10.41
2	Disgust	433	106	316	19.00	18.12	20.82
3	Fear	391	94	295	17.16	16.07	19.43
4	Joy	605	179	393	26.56	30.60	25.89
5	Love	562	175	367	24.67	29.91	24.18
6	Optimism	561	169	344	24.62	28.89	22.66
7	Pessimism	499	125	377	21.90	21.37	24.83
8	Sadness	842	217	579	36.96	37.09	38.14
9	Surprise	47	13	38	02.06	02.22	02.50
10	Trust	120	36	77	05.27	06.15	05.07

4.2.2. Data Preprocessing: Arabic Tweet Preprocessing

Data preprocessing serves as the preliminary phase for refining input data, aiming to standardize it into an appropriate format. It comprises different steps depending on the data structure and intended task [65]. The pivotal merit of implementing data preprocessing lies in diminishing the overall dataset features, concurrently elevating classifier performance in resource efficiency and classification accuracy [66]. The procedural steps employed in this study for tweet preprocessing include the following:

- Punctuation removal: we removed symbols (-, _, ., ,, ;, :, ', etc.) that are irrelevant in our proposal.
- Latin characters and digit removal: we excluded numerical and Latin data because they are not effective in categorizing the emotional label within tweets.
- Emoji replacement: we developed a lexicon comprising approximately 100 commonly used emojis on Twitter. Subsequently, we replaced each emoji with its corresponding Arabic word.

4.3. Prompt Design

GPT models are optimized for use in English, but many of them such as GPT-3.5 Turbo demonstrate impressive capabilities in multiple languages [67]. Ref. [21] found ChatGPT to work better with English prompts than Arabic prompts. This is in line with [68], who observed that for the low-resource language (e.g., Arabic), providing task descriptions in a high-resource language such as English can improve results. In this regard, in our initial experiments, we experimented with a diverse set of prompts in both English and Arabic to engineer an efficient prompt for the Arabic emotion recognition task [69]. We observed that the English prompt outperforms its Arabic counterpart. Hence, we designed an effective prompt in English depicted in Figure 6. Figure 7 shows its structure in the Chat Completions API format [64]. The English translation of the Arabic words used in this prompt is as follows: List of emotions: "anger, anticipation, disgust, fear, joy, love, optimism, pessimism, sadness, surprise, trust", Tweet: "It's almost 7:30 AM and you're alone. #Misery Hatred".

```
Prompt = """
        You are an Arabic emotion recognizer. You will be provided with a tweet written in Arabic variants (Modern Standard
        Arabic and Dialectal Arabic), and your task is to determine whether each item in the following list of emotions written
        in Arabic is expressed in this tweet. Give your answer as a list with emotion labels.

        List of emotions: ```
        الغضب، الترقب، الاشمئزاز، الخوف، الفرح، الحب، التفاؤل، التشاؤم، الحزن، المفاجأة، الثقة
        ```
        Tweet: ```إنها على وشك ان تصبح السابعه ونصف صباحا و أنت وحيد#بؤس كره```
"""
```

**Figure 6.** Designed prompt with an input text (Tweet) for the predictive testing of Arabic multi-label emotion classification task.

```
{'messages': [
        {'role': 'system',
          'content': 'You are an Arabic emotion recognizer. You will be provided with a tweet written in Arabic variants
          (Modern Standard Arabic and Dialectal Arabic), and your task is to determine whether each item in the following list
          of emotions written in Arabic is expressed in this tweet. Give your answer as a list with emotion labels.
          List of emotions: ```الغضب، الترقب، الاشمئزاز، الخوف، الفرح، الحب، التفاؤل، التشاؤم، الحزن، المفاجأة، الثقة```'},

          {'role': 'user',
          'content': 'إنها على وشك ان تصبح السابعه ونصف صباحا و أنت وحيد#بؤس كره '},
    ]
}
```

Figure 7. Test sample (zero-shot prompt).

In addition, we designed an EmotionPrompt to evaluate and understand GPT-4 behavior on the Arabic emotion recognition task. According to [26], an EmotionPrompt requires only appending emotional stimuli to the end of the initial prompts. In this regard, we have chosen the emotional stimulus "This is very important to my career" and added it to the end of the original prompt. As demonstrated in the EmotionPrompt original paper [26], this stimulus is inspired by self-monitoring psychological phenomena and emerges as the most effective one for the 24 instruction induction tasks proposed in [70], ranging from morphosyntactic tasks and semantics category to sentiment analysis. From our point of view, this stimulus is very suitable since the topics discussed in this dataset are related to some social problems (e.g., violence and crime), political events and issues, healthcare and public health, etc. It means that in this case, the user is interested more in making a good decision about their career/life related implicitly to the discussed field. Hence, the designed EmotionPrompt for our task is as follows (Figure 8):

```
EmotionPrompt = """
        You are an Arabic emotion recognizer. You will be provided with a tweet written in Arabic variants (Modern Standard
        Arabic and Dialectal Arabic), and your task is to determine whether each item in the following list of emotions written
        in Arabic is expressed in this tweet. Give your answer as a list with emotion labels. This is very important to my
        career.

        List of emotions: ```
        الغضب، الترقب، الاشمئزاز، الخوف، الفرح، الحب، التفاؤل، التشاؤم، الحزن، المفاجأة، الثقة
        ```

        Tweet: ``` إنها على وشك ان تصبح السابعه ونصف صباحا و أنت وحيد#بؤس كره ```
"""
```

Figure 8. EmotionPrompt with the emotional stimuli "This is very important to my career".

4.4. Supervised Fine-Tuning Process

To perform fine-tuning on the gpt-3.5-turbo base model, we used the official OpenAI Fine-tuning API [25]. We trained a new fine-tuned model using training data and the selected underlying model gpt-3.5-turbo-0613. Regarding the model hyperparameters, OpenAI allows us to specify the number of epochs (n_epochs) to fine-tune a model. It recommends initially training without specifying the number of epochs (n_epochs defaults to "auto"). This allows to pick a default value based on the dataset size. In our case, the picked default value was n_epochs = 3. Then, we adjusted this value manually to 5 because we observed that the model did not follow the training data as much as expected. This is in line with the OpenAI fine-tuning guide which recommends increasing the number by 1 or 2 epochs for tasks in which there is a single ideal completion such as our multi-label emotion classification task. This is a task for which we can compute a final accuracy metric against a reference answer. We reused the default values for the learning rate multiplier and batch size.

- Data formatting for fine-tuning: Train and validation sets

We create a diverse set of demonstration conversations that are similar to the conversations we will ask the model to respond to at the inference time. Each training example in the training set and the validation set should be a conversation in the same format as the OpenAI Chat Completions endpoint; specifically, each training example is a list of messages where each message has a role and content. Figure 9 shows an example of an entry in the train set formatted according to the Chat Completions conversation [64].

```
{"messages": [
        {"role": "system",
        "content": "You are an Arabic emotion recognizer. You will be provided with a tweet written in Arabic variants (Modern
        Standard Arabic and Dialectal Arabic), and your task is to determine whether each item in the following list of
        emotions written in Arabic is expressed in this tweet. Give your answer as a list with emotion labels.
        List of emotions: ```الغضب، الترقب، الاشمئزاز، الخوف، الفرح، الحب، التفاؤل، التشاؤم، الحزن، المفاجأة، الثقة```\n"},
        {"role": "user",
        "content": "كل سنه وانتي بخير ياقلبي وكل سنه وانتي سعيده يارب حب حب "},
        {"role": "assistant",
        "content": "[\"الفرح\", \"الحب\", \"التفاؤل\"]"}
        ]
}
```

Figure 9. Fine-tuning JSON train sample.

In addition to training data, we provided validation data, which will be used to make sure that the model does not overfit our training set. It has the same format as the train set.

5. Evaluation

5.1. Evaluation Settings

In all the experiments, we use gpt-3.5-turbo-0613 and gpt-4-0613 snapshots of GPT-3.5 Turbo and GPT-4, respectively. We evaluated the base model GPT-3.5 Turbo in zero-shot, one-shot, and few-shot settings. GPT-4 was evaluated in a zero-shot setting, EmotionPrompt within zero-shot and few-shot incorporating the emotion stimulus "This is very important to my career". As shown in the experiments from the EmotionPrompt original paper [26], this stimulus is the most effective one for the 24 instruction induction tasks proposed in [70], ranging from morphosyntactic tasks and semantics category to sentiment analysis.

Regarding zero-shot, we evaluated each example in the test set. For one-shot, we evaluated each example in the test set by randomly drawing one demonstration (i.e., example) from the training set as conditioning. For the GPT-3.5 Turbo few-shot setting, we evaluated each example in the test set by randomly drawing k examples from the training set as conditioning. This is in line with the work proposed in the GPT-3 original paper [24], where the authors evaluated their models in this way for some tasks. K can be any value from 0 to the maximum amount allowed by the model's context window, which is n_{ctx} = 4097 tokens. In our case, we have separate validation and test sets; therefore, we experiment with a few values of K on the validation set and then run the best value on the test set. The best value of k in our case was 7. Larger values of K are usually but not always better [24].

For the EmotionPrompt in the few-shot experiment, we used the same prompt as in the EmotionPrompt zero-shot experiment and randomly sampled five input–output pairs as in-context demonstrations (5-shot) from the train set. The number of demonstrations was set to 5 as an optimized value suggested in the EmotionPrompt original paper [26].

We utilized the official OpenAI Chat Completions API to make predictions on the test set using the fine-tuned models, GPT-3.5 Turbo and GPT-4 base models. To control the randomness and the repetition degree of generation of these LLMs, we set the following decoding strategy parameters:

- temperature = 0: Higher values like 0.8 will make the completions more random, while lower values like 0.2 will make it more focused and deterministic. Since the completion of our emotion recognition task must contain a list of exact labels (emotions), we chose a temperature value equal to 0 to make it more deterministic.
- frequency_penalty = 0 *(Defaults to 0):* Limits the frequency of tokens in a given response. Positive values penalize new tokens based on their existing frequency in the text so far, decreasing the model's likelihood of repeating the same line verbatim. In our emotion recognition task, the completion cannot contain more than one occurrence of each label returned.
- presence_penalty = 0 *(Defaults to 0)*: Positive values penalize new tokens based on whether they appear in the text so far, increasing the model's likelihood of talking about new topics. In our case, we chose the default value, we do not need to force the system to use new tokens and produce new ideas. Our model aims at predicting the emotion labels conveyed in an Arabic tweet among 11 ones ("anger", "anticipation", "disgust", "fear", "joy", "love", "optimism", "pessimism", "sadness", "surprise", and "trust").

5.2. Evaluation Metrics

We used example-based measures and label-based measures to assess the performance of each model. They are common and suited metrics for multi-label classification tasks [71–73]. The SemEval-2018 Task 1 competition organizers have similarly utilized these evaluation metrics for the E-c subtask [13]. Example-based measures include Jaccard accuracy, precision, recall, and F1-score. Label-based measures comprise six micro-precision, micro-recall, micro-F1, macro-precision, macro-recall, and macro-F1 described below.

- Example-Based Measures

Definitions provided below rely on specific notation: y_i represents the set of true labels for the example x_i, $f(x_i) = \widetilde{y}_i$ represents the set of labels predicted by the classifier f for the example x_i, N denotes the total number of examples, and Q denotes the total number of labels.

Accuracy (Equation (1)) for an input example x_i is determined through the Jaccard similarity coefficient between the predicted label sets $\widetilde{y}_i$ and true label sets y_i. Accuracy is defined as the micro-averaged value across all examples in the dataset:

$$Accuracy(f) = \frac{1}{N}\sum_{i=1}^{N}\frac{|f(x_i) \cap y_i|}{|f(x_i) \cup y_i|} \tag{1}$$

Precision (Equation (2)) is calculated as follows:

$$Precision(f) = \frac{1}{N}\sum_{i=1}^{N}\frac{|f(x_i) \cap y_i|}{|y_i|} \tag{2}$$

Recall (Equation (3)) is calculated as follows:

$$Recall(f) = \frac{1}{N}\sum_{i=1}^{N}\frac{|f(x_i) \cap y_i|}{|f(x_i)|} \tag{3}$$

F1-score (Equation (4)) is calculated as the harmonic mean between precision and recall:

$$F1 - score = \frac{2 \times Precision(f) \times Recall(f)}{Precision(f) + Recall(f)} = \frac{1}{N}\sum_{i=1}^{N} \frac{2 \times |f(x_i) \cap y_i|}{|f(x_i)| + |y_i|} \tag{4}$$

- Label-based measures

Definitions provided below rely on specific notation: TP_j, FP_j and FN_j denote, respectively, the total number of "True Positives", "False Positives", and "False Negative" considering the label λ_j as a binary class.

Macro-precision (Equation (5)) is calculated as the precision averaged across all labels:

$$Macro - Precision = \frac{1}{Q}\sum_{j=1}^{Q} \frac{TP_j}{TP_j + FP_j} \tag{5}$$

Macro-recall (Equation (6)) is calculated as the recall averaged across all labels:

$$Macro - Recall = \frac{1}{Q}\sum_{j=1}^{Q} \frac{TP_j}{TP_j + FN_j} \tag{6}$$

Macro-F1 (Equation (7)) is calculated as the harmonic mean between precision and recall, where the average is calculated individually for each label and then averaged across all labels. If p_j and r_j are the precision and recall for all $\lambda_j \in f(x_i)$ from $\lambda_j \in y_i$, the macro-F1 is determined as:

$$Macro - F1 = \frac{1}{Q}\sum_{j=1}^{Q} \frac{2 \times p_j \times r_j}{p_j + r_j} \tag{7}$$

Micro-precision (Equation (8)) is calculated as the precision averaged over all the pairs of example/labels:

$$Micro - Precision = \frac{\sum_{j=1}^{Q} TP_j}{\sum_{j=1}^{Q} TP_j + \sum_{j=1}^{Q} FP_j} \tag{8}$$

Micro-recall (Equation (9)) is calculated as the recall averaged over all the pairs of example/labels:

$$Micro - Recall = \frac{\sum_{j=1}^{Q} TP_j}{\sum_{j=1}^{Q} TP_j + \sum_{j=1}^{Q} FN_j} \tag{9}$$

Micro-F1 (Equation (10)) is calculated as the harmonic mean between micro-precision and micro-recall:

$$Micro - F1 = \frac{2 \times Micro - Precision \times Micro - Recall}{Micro - Precision + Micro - Recall} \tag{10}$$

Moreover, we have also employed the receiver operating characteristic curve (ROC curve). This graphical representation illustrates the predictive performance of the classification model across all classification thresholds by depicting the "True Positive Rate (TPR)" and "False Positive Rate (FPR)" parameters [74]. A curve skewed more towards a higher TPR indicates better performance of the classification model. The optimal classifier prediction would feature a curve tending towards the point (0, 1) or the top-right corner.

6. Results, Discussion, and Limitations

In this section, we report the results of the fine-tuning process and test the fine-tuned models' performance on the Arabic multi-label emotion classification task. We compare the fine-tuned models with the base model GPT-3.5 Turbo tested within in-context learning settings (zero-shot, one-shot, and few-shot), GPT-4 tested within zero-shot and EmotionPrompt, as well as other methods from the literature.

6.1. Analyzing the Fine-Tuned Models

In this sub-section, we aim to analyze the training metrics computed over the course of training. Two fine-tuning processes were performed. The first one achieved a final training loss of 0.07674 and a final training accuracy of 0.93939 over three epochs (at step 1709 (see Figure 10a,b below)). The second one achieved a final training loss of 0.04444 and a final training accuracy of 1.0 over five epochs (at step 1628).

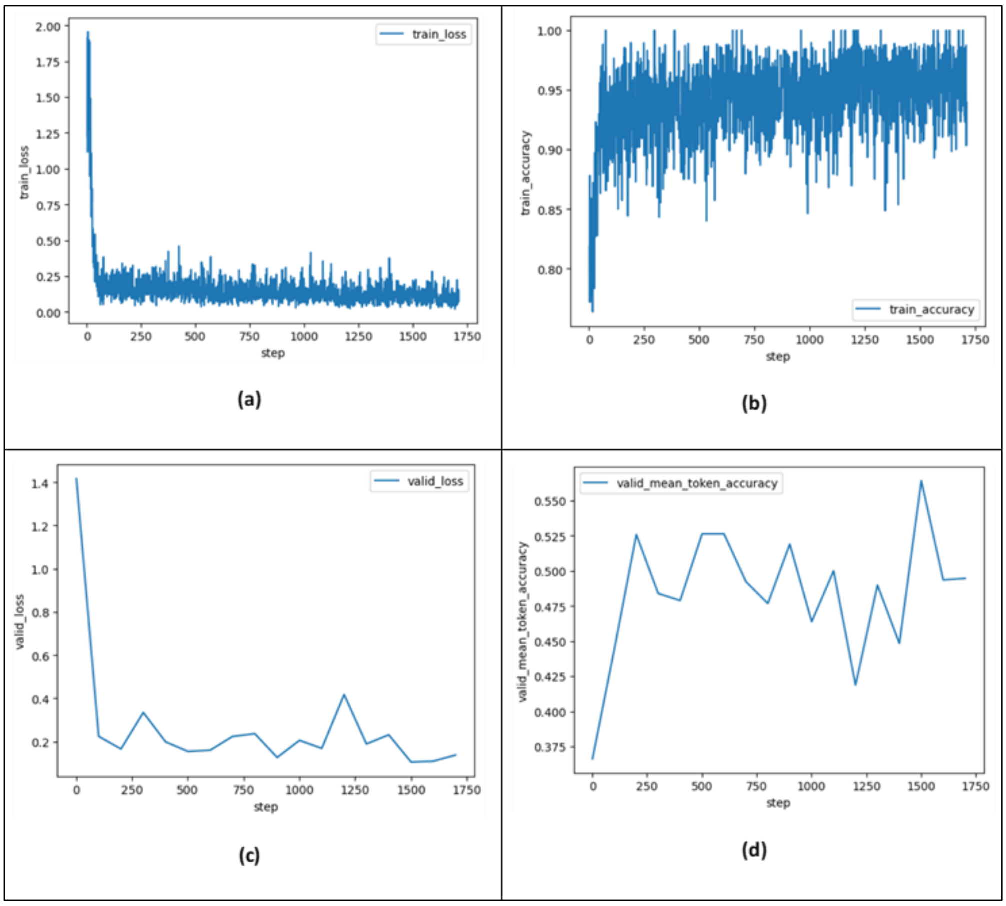

Figure 10. Training metrics computed throughout training (training run steps) for fine-tuned GPT-3.5-turbo model (n_epochs = 3). (**a**) train_loss, (**b**) train_accuracy, (**c**) valid_loss computed at periodic intervals during training, (**d**) valid_mean_token_accuracy computed at periodic intervals during training. The fine-tuning job running time was approximately 52 min 31 s.

These metrics show the models' generalization and adaptability during fine-tuning. For more details, Figure 10a,b show the train_loss and train_accuracy (training token accuracy) computed during the training run steps over three epochs, where a step refers to one forward and backward pass on a batch of data. Figure 10c,d show evaluations (valid_loss and valid_mean_token_accuracy) on how well the fine-tuned model performs against the validation data at periodic intervals during training. These plots provide a sanity check that training went smoothly (loss decrease, token accuracy increase). Plots related to five epochs are not expanded to reduce the clutter.

After the fine-tuning process (the model training) is completed, the fine-tuned models will be evaluated on the test data. The next section presents the overall results and the performance comparison with other models.

6.2. Comparative Analysis and Models' Evaluation

6.2.1. Fine-Tuned Models Evaluation and Performance Comparison with the Base Model and SOTA

Table 3 shows the multi-label classification metrics for the fine-tuned model_1 (three epochs), and the fine-tuned model_2 (five epochs). It shows the results obtained from the base model GPT-3.5-turbo (gpt-3.5-turbo-0613 snapshot) tested within in-context learning settings (zero-shot, one-shot, and few-shot), GPT-4 (gpt-4-0613 snapshot) tested in zero-shot, EmotionPrompt within zero-shot and few-shot, and six models from the literature, using the same test set.

Table 3. Results on Arabic multi-label emotion recognition task using SemEval-2018 E-c dataset. Zero-shot, one-shot, few-shot, and fine-tuning compared to the *SOTA*. The scores are accuracy (Jaccard similarity coefficient) and F1-score. GPT-3.5-turbo and GPT-4 refer to gpt-3.5-turbo-0613 and gpt-4-0613 snapshots, respectively, from 13 June 2023.

Setting	Jaccard Accuracy	Micro-Averaged F1-Score	Macro-Averaged F1-Score
Model_1: Fine-tuned GPT-3.5-turbo model (n_epochs = 3)	**62.03%**	**73.00%**	**62.00%**
Model_2: Fine-tuned GPT-3.5-turbo model (n_epochs = 5)	**61.91%**	**72.00%**	**61.00%**
SOTA: Elfaik and Nfaoui, 2023 [20]	60.00%	52.00%	35.00%
GPT-4 (EmotionPrompt within Few-Shot)	54.22%	65.00%	58.00%
Mansy et al., 2022 [18]	54.00%	52.7%	70.10%
Elfaik and Nfaoui, 2021 [16]	53.82%	-	-
GPT-4 (Zero-Shot)	53.07%	65.00%	55.00%
GPT-4 (EmotionPrompt within Zero-Shot)	52.92%	64.00%	54.00%
Alswaidan and Menai, 2020 [15]	51.20%	63.10%	50.20%
Khalil et al., 2021 [17]	49.80%	61.50%	44.00%
EMA Team [47]	48.90%	61.80%	46.10%
GPT-3.5-turbo (Few-Shot)	48.56%	59.00%	51.00%
GPT-3.5-turbo (Zero-Shot)	48.15%	58.00%	46.00%
GPT-3.5-turbo (One-Shot)	45.49%	56.00%	49.00%

Based on the comparative analysis presented in this table, it is apparent that the fine-tuned GPT-3.5-turbo model (n_epochs = 3) has established a new state-of-the-art benchmark for the Arabic multi-label emotional recognition task, achieving relatively an impressive classification accuracy of 62.03%. Notably, our fine-tuned model exhibits an enhancement of 2.03% over the SOTA [20]. This fine-tuned model demonstrates a significant performance advantage, surpassing the results reported in the SemEval-2018-Task1: Affect in Tweets competition. Specifically, it exhibits an enhancement of 13.13% over the best-performing model in the competition, namely EMA [47], which achieved 48.90% in terms of accuracy.

Furthermore, our fine-tuned models (model_1 and model_2) achieved a remarkable micro-avg and macro-avg F1-scores of 73% and 62%, respectively—a substantial increase from the SOTA model of 52% and 35%, respectively. This notable improvement consistently supports and reinforces the performance results in classifying emotion labels conveyed in Arabic tweets.

EmotionPrompt within the few-shot experiment achieved an accuracy of 54.22%, which is slightly better than EmotionPrompt within the zero-shot experiment (52.92%) and zero-shot settings (53.07%). This is in line with the findings from EmotionPrompt original paper [26] where the authors indicated that EmotionPrompt is better at in-context learning with few-shot examples. However, in our scenario, EmotionPrompt does not boost the results. One reason can be the low-temperature value (temperature = 0) chosen in this experiment. Since the completion of our emotion recognition task must contain a list of exact labels (emotions), we chose a temperature value equal to 0 to make it more deterministic. Nevertheless, the authors in [26] found that EmotionPrompt exhibits heightened effectiveness in high-temperature settings.

To further assess the performance of the models, we provide a visual comparison in terms of the ROC score depicted in Figure 11. As we can see in this ROC curve, our fine-tuned models (model_1 and model_2) outperform all the other ones with an ROC score of 82.51% and 82.38%, respectively. Furthermore, we observed that the black and yellowish-green lines which represent model_1 and model_2, are more inclined towards TPR, indicating higher TPR and lower FPR, as well as superior classification performances.

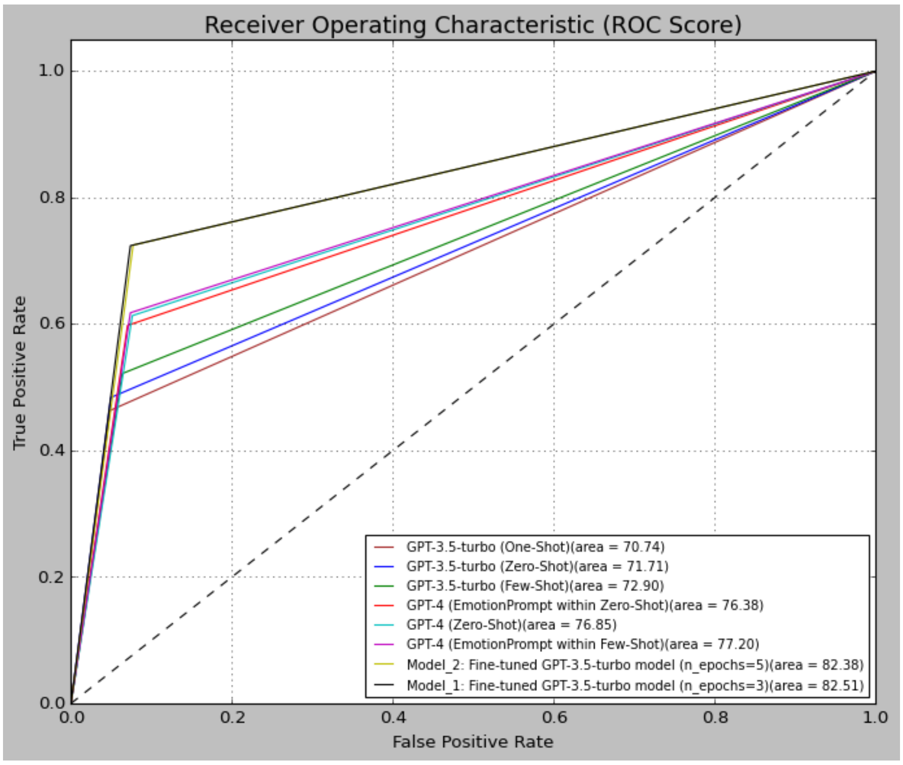

Figure 11. Models' ROC scores on the SemEval-2018 E-c dataset.

6.2.2. Models' Performance Metrics Comparison per Emotional Label

Table 4 shows the effectiveness of the top two models regarding each emotion. The precision, recall, and F1-scores associated with each label demonstrated that the fine-tuned models (model_1 and model_2) achieved the highest performance results for all emotional labels except "optimism" in terms of precision (marginal difference 2%) compared to *SOTA* (model_3).

Table 4. Comparison of the performance results per emotional label in multi-label classification task on SemEval-2018 E-c dataset. Bold indicates the best score. Model_1: Fine-tuned GPT-3.5-turbo model (n_epochs = 3), Model_2: Fine-tuned GPT-3.5-turbo model (n_epochs = 5), Model_3: SOTA [20].

Emotional Label	Precision			Recall			F1-Score		
	Model_1	Model_2	Model_3	Model_1	Model_2	Model_3	Model_1	Model_2	Model_3
anger	**0.78**	0.75	0.67	0.87	**0.88**	0.73	**0.82**	0.81	0.70
anticipation	**0.49**	0.35	0.25	**0.23**	0.21	0.02	**0.31**	0.26	0.03
disgust	0.57	**0.58**	0.45	**0.66**	**0.66**	0.35	0.61	**0.62**	0.39
fear	**0.86**	0.83	0.38	0.73	**0.76**	0.11	**0.79**	**0.79**	0.17
joy	**0.91**	0.88	0.83	0.79	**0.82**	0.54	**0.85**	**0.85**	0.65
love	**0.81**	0.79	0.75	**0.79**	0.78	0.46	**0.80**	0.79	0.57

Table 4. *Cont.*

Emotional Label	Precision			Recall			F1-Score		
	Model_1	**Model_2**	**Model_3**	**Model_1**	**Model_2**	**Model_3**	**Model_1**	**Model_2**	**Model_3**
optimism	0.75	0.75	**0.77**	**0.81**	**0.81**	0.52	**0.78**	**0.78**	0.62
pessimism	0.50	**0.51**	0.33	**0.59**	0.58	0.16	**0.54**	**0.54**	0.22
sadness	0.77	**0.79**	0.69	**0.79**	0.77	0.45	**0.78**	**0.78**	0.55
surprise	**0.44**	0.36	0.00	0.18	**0.21**	0.00	0.26	**0.27**	0.00
trust	0.37	**0.39**	0.00	**0.21**	0.18	0.00	**0.27**	0.25	0.00
Micro-Avg	**0.73**	0.72	0.66	**0.72**	**0.72**	0.42	**0.73**	0.72	0.52
Macro-Avg	**0.66**	0.63	0.47	0.60	**0.61**	0.30	**0.62**	0.61	0.35

Overall, model_1 effectively recognizes the correct emotional labels in the majority of input tweets. However, the F1-score performance results for each emotional state reveal a challenge in effectively detecting unbalanced emotional labels including "anticipation", "surprise", and "trust". One contributing factor could be the significant disparity in the number of instances among emotion labels within the SemEval-2018 dataset (Table 2). Notably, the emotion label *"anger"* had the highest number of samples, constituting 39.46% and 40.12% of samples in both train and test sets, respectively. Conversely, emotional labels such as "anticipation", "surprise", and "trust" had considerably the smallest number of instances, comprising only 9,17%, 2.06%, and 5.27% in the train set and 10,41%, 2.50%, and 5.07% in the test set, respectively. Addressing these imbalances in dataset distribution could enhance the model's effectiveness across all emotional labels.

Moreover, our fine-tuned models (model_1 and model_2) demonstrate a notable enhancement over the *SOTA* (model_3) in both macro-average and micro-average scores. Specifically, model_1 achieved a significant increase in precision scores of 19% and 7%, recall scores of 30% and 30%, and F1-scores of 27% and 21%, respectively. Hence, this substantial rise ultimately supports and reinforces the performance results.

Overall, we conclusively show that the fine-tuned GPT-3.5-turbo model surpassed task-specific models for the Arabic multi-label emotion classification task.

6.3. Limitations

In this study, we observed that GPT-3.5 Turbo and GPT-4 have limitations in addressing the morphology and syntax of Arabic script. This is in line with other research [21,75,76] which showed that GPT-3.5 performs generally well for high-resource languages such as English and German compared to low-resource ones such as Arabic. Apart from the ratio of the amount of Arabic data sources in the pre-training data, one reason can be the tokenizer used by GPT-3.5 Turbo and GPT-4. Authors in [77] have shown that tokenizers can negatively affect LLMs' multilingual capabilities and particularly their performance in the Arabic language, leading to inaccuracies in downstream tasks.

Another issue related to the effectiveness of these models in the Arabic multi-label emotion classification task is that the available benchmark dataset we used for fine-tuning is relatively small and imbalanced. This prevents the models from learning all conveyed emotions and understanding the implicit relationships among certain correlated emotions. For example, the training samples for "anger" and "sadness" are approximately half of the total number of the train instances, representing approximately 19 times than those corresponding to "surprise". Consequently, during the fine-tuning process, the models have a greater capacity to learn additional input features associated with "anger" and "sadness". To improve the performance results, we suggest fine-tuning these models on a large and balanced dataset of multilabel emotion classification.

7. Conclusions and Future Directions

We found that the fine-tuned GPT-3.5-Turbo on Arabic multi-label emotion classification established a new SOTA on the SemEval-2018 E-c benchmark dataset. It outperformed the base models experimented with few-shot prompting and EmotionPrompt, as well as task-specific models. The fine-tuned model showed enhanced accuracy and improved F1-scores, precision and recall. We observed the model's limitation in addressing the unique morphology and syntax of Arabic script, which influences its efficacy in achieving a high accuracy value. Using tokenizers suited to the Arabic language could ensure more accurate text processing and represent a good idea to tackle this issue. Indeed, the tokenization process plays a crucial role and can negatively affect LLMs' multilingual capabilities when it does not consider the unique morphological aspects of low-resourced natural languages such as Arabic, leading to inaccuracies in downstream tasks.

Exploring and evaluating recent LLMs predominantly focused on Arabic linguistic elements is a good way to deal with the Arabic multi-label emotion classification task aimed at developing accurate and intelligent applications for multiple fields, including brand/business reputation monitoring, customer satisfaction understanding, etc.

Author Contributions: Conceptualization, E.H.N.; Formal analysis, E.H.N. and H.E.; Investigation, E.H.N. and H.E.; Methodology, E.H.N. and H.E.; Resources, E.H.N. and H.E.; Software, E.H.N. and H.E.; Validation, E.H.N. and H.E.; Writing—original draft, E.H.N.; Writing—review & editing, E.H.N. and H.E. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original data presented in the study are openly available in https://competitions.codalab.org/competitions/17751#learn_the_details-datasets (accessed on 6 May 2024).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Ekman, P. *Facial Expressions of Emotion: New Findings, New Questions*; SAGE Publications: Los Angeles, CA, USA, 1992.
2. Peng, S.; Cao, L.; Zhou, Y.; Ouyang, Z.; Yang, A.; Li, X.; Jia, W.; Yu, S. A survey on deep learning for textual emotion analysis in social networks. *Digit. Commun. Networks* **2021**, *8*, 745–762. [CrossRef]
3. Balazs, J.A.; Velásquez, J.D. Opinion Mining and Information Fusion: A survey. *Inf. Fusion* **2016**, *27*, 95–110. [CrossRef]
4. Cambria, E. Affective Computing and Sentiment Analysis. *IEEE Intell. Syst.* **2016**, *31*, 102–107. [CrossRef]
5. Chaparro-Peláez, J.; Hernández-García, Á.; Urueña-López, A. The Role of Emotions and Trust in Service Recovery in Business-to-Consumer Electronic Commerce. *J. Theor. Appl. Electron. Commer. Res.* **2015**, *10*, 77–90. [CrossRef]
6. Pezoa-Fuentes, C.; García-Rivera, D.; Matamoros-Rojas, S. Sentiment and Emotion on Twitter: The Case of the Global Consumer Electronics Industry. *J. Theor. Appl. Electron. Commer. Res.* **2023**, *18*, 765–776. [CrossRef]
7. Li, S.; Zhu, B.; Zhang, Y.; Liu, F.; Yu, Z. A Two-Stage Nonlinear User Satisfaction Decision Model Based on Online Review Mining: Considering Non-Compensatory and Compensatory Stages. *J. Theor. Appl. Electron. Commer. Res.* **2021**, *19*, 272–296. [CrossRef]
8. Poushneh, A.; Vasquez-Parraga, A.Z. Emotional Bonds with Technology: The Impact of Customer Readiness on Upgrade Intention, Brand Loyalty, and Affective Commitment through Mediation Impact of Customer Value. *J. Theor. Appl. Electron. Commer. Res.* **2019**, *14*, 90–105. [CrossRef]
9. Choudrie, J.; Patil, S.; Kotecha, K.; Matta, N.; Pappas, I. Applying and understanding an advanced, novel deep learning approach: A COVID 19, text based, emotions analysis study. *Inf. Syst. Front.* **2021**, *23*, 1431–1465. [CrossRef] [PubMed]
10. Rosa, R.L.; De Silva, M.J.; Silva, D.H.; Ayub, M.S.; Carrillo, D.; Nardelli, P.H.J.; Rodriguez, D.Z. Event Detection System Based on User Behavior Changes in Online Social Networks: Case of the COVID-19 Pandemic. *IEEE Access* **2020**, *8*, 158806–158825. [CrossRef]
11. Denecke, K.; Vaaheesan, S.; Arulnathan, A. A Mental Health Chatbot for Regulating Emotions (SERMO)—Concept and Usability Test. *IEEE Trans. Emerg. Top. Comput.* **2021**, *9*, 1170–1182. [CrossRef]
12. Lee, D.; Oh, K.-J.; Choi, H.-J. The chatbot feels you—A counseling service using emotional response generation. In Proceedings of the 2017 IEEE International Conference on Big Data and Smart Computing (BigComp), Jeju, Republic of Korea, 13–16 February 2017; pp. 437–440. [CrossRef]
13. Mohammad, S.; Bravo-Marquez, F.; Salameh, M.; Kiritchenko, S. Semeval-2018 task 1: Affect in tweets. In Proceedings of the 12th International Workshop on Semantic Evaluation, New Orleans, LA, USA, 5–6 June 2018; pp. 1–17.

14. Elfaik, H.; Nfaoui, E.H. Social Arabic Emotion Analysis: A Comparative Study of Multiclass Classification Techniques. In Proceedings of the 5th International Conference On Intelligent Computing in Data Sciences (ICDS), Fez, Morocco, 20–22 October 2021. [CrossRef]
15. Alswaidan, N.; Menai, M.E.B. Hybrid feature model for emotion recognition in Arabic text. *IEEE Access* **2020**, *8*, 37843–37854. [CrossRef]
16. Elfaik, H.; Nfaoui, E.H. Combining Context-Aware Embeddings and an Attentional Deep Learning Model for Arabic Affect Analysis on Twitter. *IEEE Access* **2021**, *9*, 111214–111230. [CrossRef]
17. EKhalil, A.H.H.; El Houby, E.M.F.E.F.; Mohamed, H.K. Deep learning for emotion analysis in Arabic tweets. *J. Big Data* **2021**, *8*, 136. [CrossRef]
18. Mansy, A.; Rady, S.; Gharib, T. An Ensemble Deep Learning Approach for Emotion Detection in Arabic Tweets. *Int. J. Adv. Comput. Sci. Appl.* **2022**, *13*, 4. [CrossRef]
19. ASamy, E.; El-Beltagy, S.R.; Hassanien, E. A context integrated model for multi-label emotion detection. *Procedia Comput. Sci.* **2018**, *142*, 61–71.
20. Elfaik, H.; Nfaoui, E.H. Leveraging feature-level fusion representations and attentional bidirectional RNN-CNN deep models for Arabic affect analysis on Twitter. *J. King Saud Univ. Comput. Inf. Sci.* **2023**, *35*, 462–482. [CrossRef]
21. Khondaker, M.T.I.; Waheed, A.; Nagoudi, E.M.B.; Abdul-Mageed, M. GPTAraEval: A comprehensive evaluation of ChatGPT on Arabic NLP. *arXiv* **2023**, arXiv:2305.14976.
22. Alyafeai, Z.; Alshaibani, M.S.; AlKhamissi, B.; Luqman, H.; Alareqi, E.; Fadel, A. Taqyim: Evaluating arabic nlp tasks using chatgpt models. *arXiv* **2023**, arXiv:2306.16322.
23. Sallam, M.; Mousa, D. Evaluating ChatGPT performance in Arabic dialects: A comparative study showing defects in responding to Jordanian and Tunisian general health prompts. *Mesopotamian J. Artif. Intell. Healthc.* **2024**, *2024*, 1–7. [CrossRef]
24. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 1877–1901.
25. OpenAI API, "Fine-Tuning," 2023. Available online: https://platform.openai.com/docs/guides/fine-tuning (accessed on 27 March 2024).
26. Li, C.; Wang, J.; Zhang, Y.; Zhu, K.; Hou, W.; Lian, J.; Luo, F.; Yang, Q.; Xie, X. Large Language Models Understand and Can be Enhanced by Emotional Stimuli. *arXiv* **2023**, arXiv:2307.11760.
27. OpenAI, "OpenAI Models," 2023. Available online: https://platform.openai.com/docs/models/overview (accessed on 27 March 2024).
28. Wang, Y.; Song, W.; Tao, W.; Liotta, A.; Yang, D.; Li, X.; Gao, S.; Sun, Y.; Ge, W.; Zhang, W.; et al. A systematic review on affective computing: Emotion models, databases, and recent advances. *Inf. Fusion* **2022**, *83–84*, 19–52. [CrossRef]
29. Mehrabian, A. *Basic Dimensions for a General Psychological Theory: Implications for Personality, Social, Environmental, and Developmental Studies*; Oelgeschlager, Gunn & Hain: Cambridge, UK, 1980.
30. Russell, J.A. A circumplex model of affect. *J. Personal. Soc. Psychol.* **1980**, *39*, 1161–1178. [CrossRef]
31. Ekman, P. Basic emotions. *Handb. Cogn. Emot.* **1999**, *98*, 19.
32. Plutchik, R. *Emotions and Life: Perspectives from Psychology, Biology, and Evolution*; American Psychological Association: Washington, DC, USA, 2003.
33. Cambria, E.; Livingstone, A.; Hussain, A. The hourglass of emotions. In *Cognitive Behavioural Systems: COST 2102 International Training School, Dresden, Germany, 21–26 February 2011, Revised Selected Papers*; Springer: Berlin/Heidelberg, Germany, 2012; pp. 144–157.
34. Susanto, Y.; Livingstone, A.G.; Ng, B.C.; Cambria, E. The Hourglass Model Revisited. *IEEE Intell. Syst.* **2020**, *35*, 96–102. [CrossRef]
35. Lopes, A.T.; de Aguiar, E.; De Souza, A.F.; Oliveira-Santos, T. Facial expression recognition with Convolutional Neural Networks: Coping with few data and the training sample order. *Pattern Recognit.* **2017**, *61*, 610–628. [CrossRef]
36. Ren, Z.; Baird, A.; Han, J.; Zhang, Z.; Schuller, B. Generating and Protecting Against Adversarial Attacks for Deep Speech-Based Emotion Recognition Models. In Proceedings of the ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 7184–7188. [CrossRef]
37. Acheampong, F.A.; Wenyu, C.; Nunoo-Mensah, H. Text-based emotion detection: Advances, challenges, and opportunities. *Eng. Rep.* **2020**, *2*, e12189. [CrossRef]
38. El Gohary, A.F.; Sultan, T.I.; Hana, M.A.; El Dosoky, M.M. A computational approach for analyzing and detecting emotions in Arabic text. *Int. J. Eng. Res. Appl.* **2013**, *3*, 100–107.
39. AAl-Aziz, M.A.; Gheith, M.; Eldin, A.S. Lexicon based and multi-criteria decision making (MCDM) approach for detecting emotions from Arabic microblog text. In Proceedings of the 2015 First International Conference on Arabic Computational Linguistics (ACLing), Cairo, Egypt, 17–20 April 2015; pp. 100–105.
40. Al-A'abed, M.; Al-Ayyoub, M. A lexicon-based approach for emotion analysis of arabic social media content. In Proceedings of the International Computer Sciences and Informatics Conference (ICSIC), Amman, Jordan, 12–13 January 2016.
41. Mohammad, S.M.; Kiritchenko, S.; Zhu, X. NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets. *arXiv* **2013**, arXiv:1308.6242.

42. Rabie, O.; Sturm, C. Feel the heat: Emotion detection in Arabic social media content. In Proceedings of the International Conference on Data Mining, Internet Computing, and Big Data (BigData2014), Kuala Lumpur, Malaysia, 17–19 November 2014; pp. 37–49.
43. Hussien, W.A.; Tashtoush, Y.M.; Al-Ayyoub, M.; Al-Kabi, M.N. Are emoticons good enough to train emotion classifiers of arabic tweets? In Proceedings of the 2016 7th International Conference on Computer Science and Information Technology (CSIT), Amman, Jordan, 12–13 January 2016; pp. 1–6.
44. Sayed, A.M.; AbdelRahman, S.; Bahgat, R.; Fahmy, A. Time emotional analysis of arabic tweets at multiple levels. *Int. J. Adv. Comput. Sci. Appl.* **2016**, *7*, 336–342.
45. Abdulllah, M.; AlMasawa, M.O.; Makki, I.S.; Alsolmi, M.J.; Mahrous, S.S. Emotions classification for Arabic tweets. *J. Fundam. Appl. Sci.* **2018**, *10*, 271–277.
46. Al-Khatib, A.; El-Beltagy, S.R. Emotional tone detection in arabic tweets. In Proceedings of the International Conference on Computational Linguistics and Intelligent Text Processing, Budapest, Hungary, 17–23 April 2017; pp. 105–114.
47. Badaro, G.; El Jundi, O.; Khaddaj, A.; Maarouf, A.; Kain, R.; Hajj, H.; El-Hajj, W. EMA at semeval-2018 task 1: Emotion mining for arabic. In Proceedings of the 12th International Workshop on Semantic Evaluation, New Orleans, LA, USA, 5–6 June 2018; pp. 236–244.
48. Mulki, H.; Ali, C.B.; Haddad, H.; Babaoğlu, I. Tw-star at semeval-2018 task 1: Preprocessing impact on multi-label emotion classification. In Proceedings of the 12th International Workshop on Semantic Evaluation, New Orleans, LA, USA, 5–6 June 2018; pp. 167–171.
49. Abdullah, M.; Shaikh, S. Teamuncc at semeval-2018 task 1: Emotion detection in english and arabic tweets using deep learning. In Proceedings of the 12th International Workshop on Semantic Evaluation, New Orleans, LA, USA, 5–6 June 2018; pp. 350–357.
50. Jabreel, M.; Moreno, A. EiTAKA at SemEval-2018 Task 1: An ensemble of n-channels ConvNet and XGboost regressors for emotion analysis of tweets. *arXiv* **2018**, arXiv:1802.09233.
51. Abdullah, M.; Hadzikadicy, M.; Shaikhz, S. Sedat: Sentiment and emotion detection in arabic text using cnn-lstm deep learning. In Proceedings of the 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA), Orlando, FL, USA, 17–20 December 2018; pp. 835–840.
52. Eisner, B.; Rocktäschel, T.; Bošnjak, M.; Riedel, S. emoji2vec: Learning Emoji Representations from Their Description. Available online: https://twitter.com/Kyle_MacLachlan/ (accessed on 27 April 2024).
53. Soliman, A.B.; Eissa, K.; El-Beltagy, S.R. Aravec: A set of arabic word embedding models for use in arabic nlp. *Procedia Comput. Sci.* **2017**, *117*, 256–265. [CrossRef]
54. Bojanowski, P.; Grave, E.; Joulin, A.; Mikolov, T. Enriching Word Vectors with Subword Information. *Trans. Assoc. Comput. Linguist.* **2017**, *5*, 135–146. [CrossRef]
55. Pennington, J.; Socher, R.; Manning, C.D. Glove: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 1532–1543.
56. AlZoubi, O.; Tawalbeh, S.K.S.K.S.K.; AL-Smadi, M. Affect detection from arabic tweets using ensemble and deep learning techniques. *J. King Saud Univ. Comput. Inf. Sci.* **2022**, *34*, 2529–2539. [CrossRef]
57. Baali, M.; Ghneim, N. Emotion analysis of Arabic tweets using deep learning approach. *J. Big Data* **2019**, *6*, 89. [CrossRef]
58. Shanahan, M. Talking about large language models. *Commun. ACM* **2024**, *67*, 68–79. [CrossRef]
59. Zhao, W.X.; Zhou, K.; Li, J.; Tang, T.; Wang, X.; Hou, Y.; Min, Y.; Zhang, B.; Zhang, J.; Dong, Z.; et al. A survey of large language models. *arXiv* **2023**, arXiv:2303.18223.
60. Dong, Q.; Li, L.; Dai, D.; Zheng, C.; Wu, Z.; Chang, B.; Sun, X.; Xu, J.; Li, L.; Sui, Z. A survey on in-context learning. *arXiv* **2022**, arXiv:2301.00234.
61. Huang, J.; Chang, K.C.-C. Towards reasoning in large language models: A survey. *arXiv* **2022**, arXiv:2212.10403.
62. Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; Liu, P.J. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.* **2020**, *21*, 1–67.
63. Xue, L.; Constant, N.; Roberts, A.; Kale, M.; Al-Rfou', R.; Siddhant, A.; Barua, A.; Raffel, C. mT5: A massively multilingual pre-trained text-to-text transformer. *arXiv* **2020**, arXiv:2010.11934.
64. OpenAI API, "Chat Completions API," 2023. Available online: https://platform.openai.com/docs/api-reference/chat (accessed on 27 March 2024).
65. Farha, I.A.; Magdy, W. A comparative study of effective approaches for Arabic sentiment analysis. *Inf. Process. Manag.* **2021**, *58*, 102438. [CrossRef]
66. Omar, A.; Mahmoud, T.M.; Abd-El-Hafeez, T.; Mahfouz, A. Multi-label arabic text classification in online social networks. *Inf. Syst.* **2021**, *100*, 101785. [CrossRef]
67. OpenAI, "Models Multilingual Capabilities." 2023. Available online: https://platform.openai.com/docs/models/multilingual-capabilities (accessed on 27 March 2024).
68. Lai, V.; Ngo, N.; Ben Veyseh, A.P.; Man, H.; Dernoncourt, F.; Bui, T.; Nguyen, T. Chatgpt beyond english: Towards a comprehensive evaluation of large language models in multilingual learning. *arXiv* **2023**, arXiv:2304.05613.
69. OpenAI, "OpenAI Prompt Engineering Guide." 2023. Available online: https://platform.openai.com/docs/guides/prompt-engineering (accessed on 27 March 2024).

70. Honovich, O.; Shaham, U.; Bowman, S.R.; Levy, O. Instruction Induction: From Few Examples to Natural Language Task Descriptions. *Proc. Annu. Meet. Assoc. Comput. Linguist.* **2022**, *1*, 1935–1952. [CrossRef]
71. Madjarov, G.; Kocev, D.; Gjorgjevikj, D.; Džeroski, S. An extensive experimental comparison of methods for multi-label learning. *Pattern Recognit.* **2012**, *45*, 3084–3104. [CrossRef]
72. Zhang, M.-L.; Zhou, Z.-H. A review on multi-label learning algorithms. *IEEE Trans. Knowl. Data Eng.* **2013**, *26*, 1819–1837. [CrossRef]
73. Liu, S.M.; Chen, J.-H. A multi-label classification based approach for sentiment classification. *Expert Syst. Appl.* **2015**, *42*, 1083–1093. [CrossRef]
74. Sokolova, M.; Japkowicz, N.; Szpakowicz, S. Beyond accuracy, F-score and ROC: A family of discriminant measures for performance evaluation. In *Australasian Joint Conference on Artificial Intelligence*; Springer: Berlin/Heidelberg, Germany, 2006; Volume WS-06-06, pp. 1015–1021. [CrossRef]
75. Bang, Y.; Cahyawijaya, S.; Lee, N.; Dai, W.; Su, D.; Wilie, B.; Lovenia, H.; Ji, Z.; Yu, T.; Chung, W.; et al. A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity. In Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguisticss, Nusa Dua, Bali, 1–4 November 2023; pp. 675–718. [CrossRef]
76. Huang, H.; Tang, T.; Zhang, D.; Zhao, X.; Song, T.; Xia, Y.; Wei, F. Not All Languages Are Created Equal in LLMs: Improving Multilingual Capability by Cross-Lingual-Thought Prompting. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, 6–10 December 2023; pp. 12365–12394. [CrossRef]
77. Koubaa, A.; Ammar, A.; Ghouti, L.; Najar, O.; Sibaee, S. ArabianGPT: Native Arabic GPT-based Large Language Model. February 2024. Available online: https://arxiv.org/abs/2402.15313v2 (accessed on 29 April 2024).