Arabic Emotion Cause Extraction Using Deep Learning

Yasmin Shaaban*[1], Hoda K. Mohamed *[2], Walaa Medhat **[3]

[*] *Department of Computer and Systems Engineering, Faculty of Engineering, Ain Shams University, Elsarayat Abbasseya, CAIRO, EGYPT*

[1] g18092931@eng.asu.edu.eg

[2] hoda.korashy@eng.asu.edu.eg

[**] *Department of Information technology and computer science/ FCAI, Nile University/ Benha University, GIZA/BENHA, EGYPT*

3 wmedhat@nu.edu.eg

Abstract: ***Emotion cause extraction is a challenging task nowadays. Causes behind emotions are extracted from textual data. Emotion cause extraction has many applications such as extracting causes from reviews that are extracted from social networks and recommender websites where users give their feedback. The resources in this field are limited. There are some corpora built for western languages like English and far east languages like Chinese. Arabic language resources in this field are very limited. This paper introduces emotion cause detection in Arabic Language. A dialectal Arabic annotated corpus is built for the purpose of emotion cause extraction. The data collected from many resources. Sequence labelling techniques are applied with IOB2 scheme using BiLSTM-CRF algorithm and BERT-CRF algorithm. BERT-CRF outperforms BiLSTM-CRF in both span-level and token-level measure evaluation. BERT-CRF achieves a 0.29 F1 score in case of span-level measure evaluation and a 0.84 F1 score in case of token-level measure evaluation.***

Keywords: ***Natural language processing, Emotion analysis, Emotion cause extraction, Sequence labelling, Deep learning***

1 INTRODUCTION

Emotion cause extraction (ECE) is a main task for emotion analysis. Emotional analysis represents a primary part of affective computing. “Affect” means emotion and “computing” means calculating or measuring [1]. Affective computing results in the design of systems. These systems process, recognize, interpret, and simulate human affect. These systems allow us to analyze the human-machine interactions [1]. Business organizations benefit a lot from analyzing emotions of various textual data. It helps them measure the degree of their customers’ satisfaction by analyzing their comments or feedback about the products they provide. Emotion analysis also provides a way for opinion mining for various organizations. ECE is different from other emotion analysis tasks such as emotion recognition. ECE not only focuses on emotion expression, but also cares about the emotion stimuli [2]. In certain cases, the cause behind an emotion is more important than the emotion itself [2]. For example, on the social recommendation websites, there are many evaluations and feedback from users. The service providers, for example, restaurants or hotels care more about why customers like or dislike their service rather than the emotion included in the comments [2].

In this paper, the focus is on the ECE task which its target is to extract the causes of emotions within textual data. Many approaches have been used in the ECE task, rule-based approach, common-sense-based approach, learning-based approach, and hybrid approach. The rule-based approach is an approach where rules are constructed depending on linguistic rules [3,4]. In common-sense-based approach, emotion cognition lexicon is used that contains emotion stimulations and their corresponding reflection words [3]. Learning-based approach is an approach where the ECE task is tackled using learning-based techniques, traditional machine learning [4] or deep learning techniques [7-10]. In hybrid approach, previously mentioned approaches are combined such as rule-based or common-sense-based or both with learning-based traditional machine learning approach [7,8]. In this paper, we will address learning-based techniques.

Learning-based ECE has been implemented for clause-level as a clause classification task and for span-level as sequence labelling (SL) task [5] [6] or start/end position identification task [6] or span-detection task [7]. Clause-level ECE is to extract the cause as a clause where the document is tokenized into clauses depending on punctuation marks such as comma, question mark, exclamation mark, and period. The problem is implemented as a classification problem where the clause is classified a as cause clause or non-cause clause. The main defect of this method is that the clause is not the accurate unit to be extracted but the cause can be part of a clause, or it can be just a word. Span-level has been recently used in ECE task because it is more accurate and extracts the main part of the cause.

Arabic language is one of the most spoken languages nowadays. It is the official language in 22 countries. It is spoken by hundreds of millions of people. It is very important for the 1.5 billion Muslims around the world who use it in their daily rituals and acts of worship [8]. It is the Internet's fourth most used language [9]. Arabic is classified into three main types: Classical Arabic (CA), Modern Standard Arabic (MSA), and Dialectal Arabic (DA). CA is a form of Arabic language in the Qur'an (Islam's Holy Book). The media and education use MSA as their primary language. DA is used in daily life communication and informal exchanges. DA is mostly divided into six main groups: Egyptian (EGY), Levantine (LEV), Gulf (GLF), Iraqi (IRQ), Maghrebi (MAGH) and others containing the remaining dialect [8].

In this paper, we have constructed a Dialectal Arabic dataset for the purpose of the ECE task, tackling the problem using learning-based techniques for span-level cause extraction. We have looked for the emotion causes when the emotion is conveyed explicitly in the text or implicitly. We have applied sequence labelling techniques, deep learning approach with two algorithms the first one is Bidirectional Long Short-Term Memory with Conditional Random Field (BiLSTM-CRF) and self-attention layers, and the other is Bidirectional Encoder Representation Transformers (BERT)[10] with CRF layer. BERT-CRF outperforms BiLSTM-CRF in both span-level and token-level evaluation measures.

The organization of paper is as follows; section 2 presents the literature review. Section 3 tackles the methodology. Section 4 addresses the emotion cause extraction approaches. Section 5 shows the deep learning models. Results and discussion are presented in section 6. The paper is concluded in section 7.

2 LITERATURE REVIEW

A. Emotion Cause Extraction Corpora

An English corpus of 532 sentences [5] containing emotion causes specifically for 22 emotions was constructed. 22 sentences provided in [11] were used. 510 sentences were manually collected sentences from the online ABBYY Lingvo dictionary with emotion tokens and emotion causes explicitly mentioned. 118 emotion tokens were found to be effective. One cause-containing sentence at least per emotion token was extracted. The annotation task was divided into some subtasks, defining the emotion experiencer specified by emotion token, and then extracting the phrase describing the emotion cause. The linguistic relation was then defined between emotion and its cause so that the cause was classified as positive, negative, or neutral and tokens that affected the cause polarity were extracted.

An emotion cause annotated corpus [12] consisting of the annotations for 1,333 Chinese Weibo text documents was constructed. The NLPCC13 corpus was selected as the primary resource for annotation. NLPCC13 corpus was annotated up to two basic emotion categories for each sentence and Weibo. This dataset included the seven primary emotion categories of fear, happiness, disgust, anger, surprise, and sadness. The main corpus contained the emotion annotations for 10,000 Weibo text documents. First, the Weibo text documents with explicit emotion cause were selected for annotation. Second, according to the psychological assessment of the relationship between "physiological arousal" and "expressive actions", both emotion expression and emotion cause were annotated. Based on the part-of-speech labelling of emotion causes, there were two basic types of causes: noun/noun phrase and verb/verb phrase.

An English dataset annotated with both emotion expressions and emotion causes [13] using FrameNet's emotions-directed frame was built. They utilized the Oxford Dictionary and thesaurus.com to annotate emotions in corpus. The two sources did not always agree thus emotion annotation was performed manually. They used two more sources the NRC emotion lexicon and the WordNet affect lexicon. Each lexical unit was assigned the emotion that received the most votes. They selected the sentences which contain emotion cause and then annotated with the emotion class that corresponds to it.

Chinese annotated dataset consisting of 2,105 articles [2] was released. The raw corpus was NEWS SINA2 that contained 20,000 articles. It followed the W3C Emotion Markup Language scheme. Keyword matching was used to extract 15,687 emotion keywords from the raw corpus which is based on 10,259 Chinese primary emotion keywords list [14]. After removing irrelevant instances (emotion keywords) there were still 2,105 instances remaining. The emotion categories and causes were manually annotated in the W3C Emotion Markup Language (EML) format by two annotators.

English news headlines corpus consisting of 5000 headlines [15] was constructed. It was collected from multiple resources the news publishers, social media from Twitter and Reddit. All news sources available as Really Simple Syndication (RSS) feed were from the Media Bias Chart. Using crowdsourcing, causes and emotions that correspond to them, matching experiences of emotion, associated emotion targets, cues that help in extracting causes and the perception of the headline's emotion were annotated. Proposed annotation technique is a multiphase one in which instances with emotional content are identified and finer-grained characteristics are marked.

B. Emotion Cause Extraction Tasks

Oberländer and Klinger [16] proposed an integrated framework which enabled them to evaluate the two approaches used to tackle the ECE problem, the span-level approach, and the clause-level approach. They compared the token sequence labelling (span-level) and clause classification (clause-level). They implemented models inspired by state-of-the-art approaches and evaluated them on four English datasets from different domains. Token sequence labelling achieved better results than clause classification in three out of four datasets.

Li et al. formulated the ECE problem using the span-level approach as sequence labelling and start/end position identification tasks. They also addressed the problem using the clause-level approach as a clause classification task. They applied their experiments on two datasets, the English (ENG) dataset [13] and the Chinese (CHI) dataset [17]. They showed better results on the English dataset than the Chinese dataset.

C. Emotion Cause Extraction Methods and Models

Ghazi et al. [13] constructed a CRF model. CRF is a sequential learning model which they used to detect the emotion causes spans in emotion-bearing sentences. They evaluated the model on their constructed English dataset. Their model significantly achieved a 0.78 F1 score for token-level measure evaluation and a 0.63 F1 score for span-level measure evaluation.

Bostan et al. [15] extracted emotion cues, experiencers, targets, and causes from their English constructed dataset. They built a bidirectional long short-term memory network with a CRF layer (BiLSTM-CRF). They used Embeddings from Language Model (ELMo) embeddings [18] as input and an IOB alphabet as output. They achieved a 0.14 F1 score in case of span-level measure evaluation.

Token Sequence Labeling (SL), Independent Clause Classification (ICC), and Joint Clause Classification (JCC) are the three models that Oberlander and Klinger [14] used to develop emotion stimulus detection (JCC). The architecture of the SL model consisted of a bidirectional LSTM with an attention layer, and a CRF output layer. They employed GloVe embeddings. The ICC design was similar to that of the SL. The difference between them is the final layer, which was a single SoftMax that produced a single label. The purpose of training was to minimize cross-entropy loss. The ICC model had no access to clauses other than the one for which it makes predictions. As word-level encoders, the JCC model architecture had multiple LSTM modules, one for each sentence. At the word level, the LSTM encoded the tokens in a clause into a single representation. The following layer was a clause-level encoder based on two bidirectional LSTMs, in which the representations of clauses were learned and updated by integrating the relations between different clauses. After obtaining clause-level representations, they were transmitted to the output CRF layer at the clause level. The objective of training was to minimize the loss of negative log-likelihood over all sentences. In the span-level examination, the SL model had the highest F1 score of 0.71 on the Emotion-Stimulus dataset.

Li et al. [6] applied their experiments using BERT with SoftMax layer and BERT with different networks which were CRF network, Pointer network and Gated Recurrent Unit (GRU) network. Pointer Network performed the best on the English dataset [13] and the second-best on the Chinese dataset [17]. Pointer was worse than CRF on the Chinese dataset. They achieved a 0.9 F1 score on the English dataset using BERT-Pointer and 0.57 using BERT-CRF on the Chinese dataset using span-level evaluation measure.

3 EMOTION CAUSE EXTRACTION APPROACHES

Many approaches have been used in solving ECE problem. The approaches are rule-based approach, common-sense-based approach, learning-based approach, and hybrid approach as shown in Figure 1.

A. Rule-based Approach

Rule-based approach is an approach where rules are constructed based on linguistic rules. Many rule-based systems for emotion cause extraction have been constructed depending on the linguistic rules. Some Chinese rule-based systems [19] were developed based on some observed linguistic cues that are grouped into a number of groups. These grouped linguistic cues are generalized to identify some linguistic relations between the cues, causes, experiencers, and emotion keywords. English rule-based system [5] is also built based on the analysis of the used corpus where some linguistic relations between emotion and its cause are identified.

B. Common-sense-based Approach

Common-sense-based approach is an approach where emotion cognition lexicon is used which contains emotion stimulations and their corresponding reflection words. The identification of the emotion cause events in context could be performed by looking for a plausible set of nouns which are associated with a specific emotion keyword and assumed to be its cause. ECE methodology is described based on the interplay between relevant linguistic patterns and a repository of common-sense knowledge of emotion keywords and emotion causes couples [3].

C. Learning-based Approach

Learning-based approach learns from the data itself to try to overcome the constraints of rule-based approach and common-sense-based approach. Learning-based approach constructs a predication model to determine the relation between the input text and the corresponding output emotion cause without the need to find an explicit relation to the emotion cause in the input text. The learning-based approach is divided into traditional machine learning and deep learning. Traditional machine learning models that have been used are multi-kernel Support Vector Machines (SVM) [4] and CRF [13]. Examples of deep learning models that have been used are Convolutional Multiple-Slot Deep Memory network (ConvMSMemnet) [4], RNN-Transformer Hierarchical network (RTHN) [20], CNN -BiGRU -MLP network [21] , BiLSTM-CRF with attention [16] and BERT- CRF [6].

D. Hybrid Approach

Hybrid approach is an approach that uses a combination of the previously mentioned approaches to improve the performance of emotion cause detection. Hybrid approach considers the use of rules or common sense or both as features to the machine learning model. Examples of some hybrid methods that have been used are using the linguistic rules of Lee et al. [19] as features to Max-Entropy as a classifier [22] and to SVM and CRFs as classifiers [12].

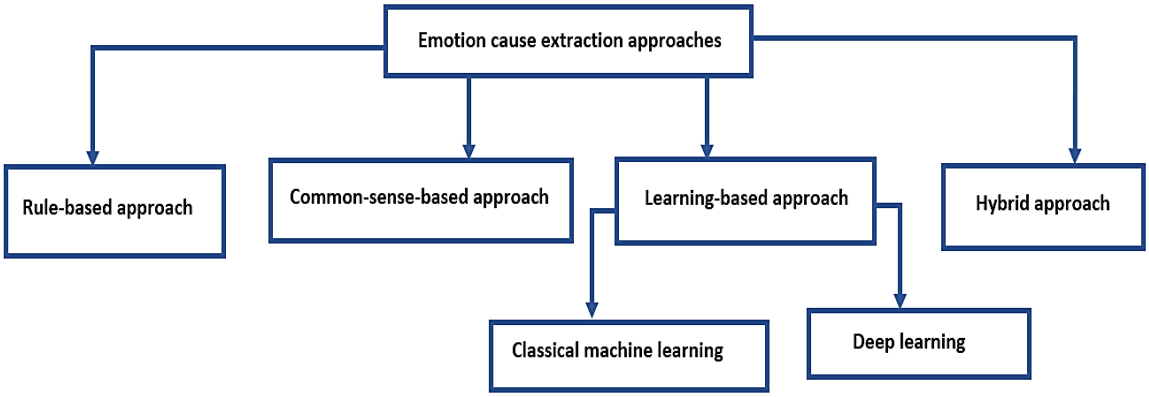

Figure 1: Emotion cause extraction approaches diagram.

4 METHODOLOGY

The overall methodology used in data annotation is shown in Fig. 2. The First stage is getting the Arabic reviews dataset [23]. Keyword spotting technique for emotion recognition (ER) is applied to extract reviews that contain emotions explicitly. Arabic lexicon [24] is used for keyword spotting. We look for reviews not only that convey emotion explicitly but also implicitly. We have selected from the chosen reviews; the reviews contain emotion causes. Causes are extracted manually. The result is our annotated Arabic reviews corpus. Our Emotion Cause Extraction framework is shown in Fig.3.The reviews and extracted causes in our annotated corpus are tokenized with different tokenization types, word tokenization for BiLSTM-CRF and sub-word tokenization for BERT. After tokenizing causes, they are represented in the IOB2 scheme.

IOB1 is a scheme where "I" is used for a token inside a chunk, "O" is used for a token outside a chunk and "B" is only used for the beginning token of a chunk that immediately follows another chunk [25]. While IOB2 is a scheme where "B" tag is given for every token, which exists at the beginning of the chunk, "I" is a token inside a chunk and "O" is a token outside a chunk [25].

Semi-automatic technique has been used for IOB2 representation. We have implemented a method that results in the basis of IOB2 scheme. The method is implemented as follows; the first word in the cause that is in the main review, is tagged as

"B" class. The other words that come after the first one, are tagged as "I" class while the words that are found in the main review but not found in the cause, are tagged as “O" class. This results in a basic representation of the review in IOB2 format. Some of words are not correctly tagged because of some reasons. There may be two words in the same sentence that have same spelling so the first one is tagged as B which is wrong while the second one is the true word that should be tagged as B. There may be more than one cause in the document so each word that is at the beginning of each cause should be tagged as B and this is not handled by our method. The resulted representation has been modified manually. Examples of our constructed corpus reviews, causes and represented reviews in IOB2 scheme are illustrated in Table I.

Sequence labeling models, BiLSTM- CRF and BERT-CRF models are then implemented and trained on the constructed annotated corpus. Testing models shows that BERT outperforms BiLSTM-CRF with a self-attention layer in case of both span-level evaluation and token-level evaluation. Models are evaluated by F1 score, precision and recall using Seqeval library in case of span-level measure evaluation and, scikit-learn library in case of token-level measure evaluation. Span-level evaluation measures the number of exact matches of spans in text while token-level evaluation measures the number of tokens (B, I and O classes) matches in text. Naturally, token-level measures have a higher value than span-level measures.

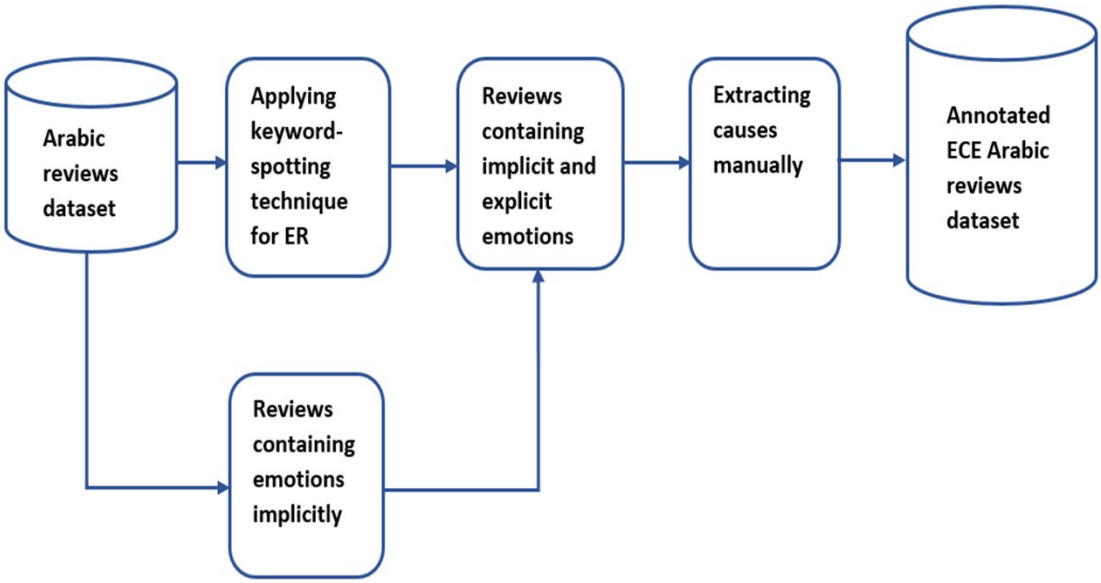

Figure 2: Arabic dataset annotation framework for ECE task.

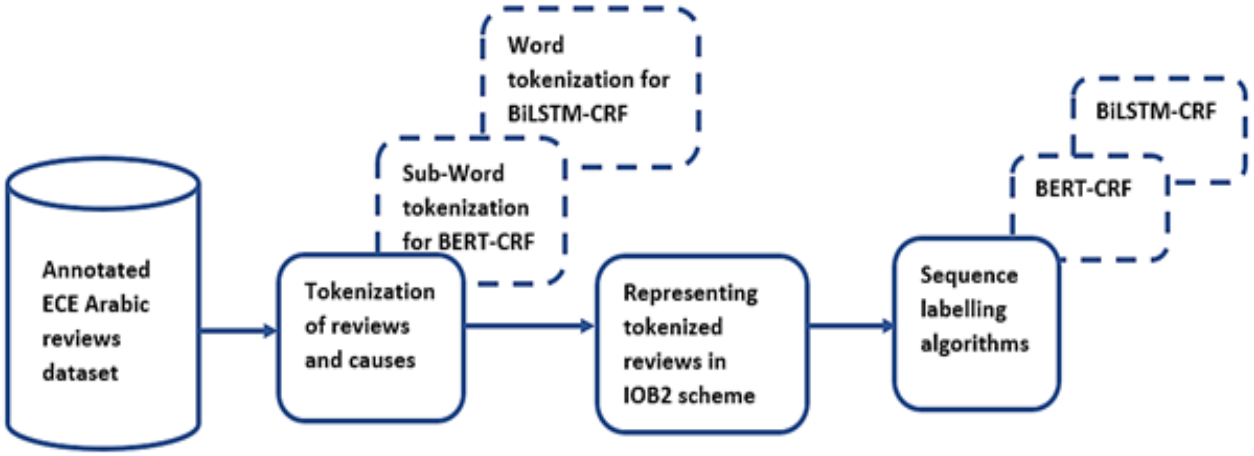

Figure 3: Emotion causes extraction framework.

TABLE I

SAMPLE OF CONSTRUCTED CORPUS REVIEWS WITH TOKENIZATION AND IOB2 REPRESENTATION.

Review	Extracted cause	Tokenization	Tokenized review	Tokenized review represented in IOB2 scheme
استثنائي. الهدوء في الجناح مع مسبح	الهدوء في الجناح مع مسبح	Word Tokenization	['استثنائي', '.', 'الهدوء', 'في', 'الجناح', 'مع', 'مسبح']	['O', 'O', 'B', 'I', 'I', 'I', 'I']
		Sub-word tokenization	['استثنايي', '.', 'الهدوء', 'في', 'الجناح', 'مع', 'مسب', '##ح']	['O', 'O', 'B', 'I', 'I', 'I', 'I', 'I']
أصابني هذا الكتاب **بصدمة** من نوع ما.. توقعته رومانسيا ... وقد كان. لكنها كانت رومانسية موجعة. **اصابني** هذا الكتاب **بقشعريرة**. وداعبت الدموع عيني. كتاب **رائع**.. لكنه حزين جدا ومقبض	هذا الكتاب هذا الكتاب كتاب	Word tokenization	['أصابني', 'هذا', 'الكتاب', 'بصدمة', 'من', 'نوع', 'ما', '..', 'توقعته', 'رومانسيا', '...', 'وقد', 'كان', '.', 'لكنها', 'كانت', 'رومانسية', 'موجعة', '.', 'اصابني', 'هذا', 'الكتاب', 'بقشعريرة', '.', 'وداعبت', 'الدموع', 'عيني', '.', 'كتاب', 'رائع', '..', 'لكنه', 'حزين', 'جدا', 'ومقبض']	['O', 'B', 'I', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O','O', 'O', 'O', 'O', 'O', 'O','O', 'O', 'O', 'B', 'I', 'O', 'O', 'O', 'O', 'O', 'O', 'B', 'O', 'O', 'O', 'O', 'O', 'O']
		Sub-word tokenization	['اصاب', '##ني', 'هذا', 'الكتاب', 'بصد', '##مة', 'من', 'نوع', 'ما', '.', '.', 'توقعت', '##ه', 'رومانسي', '##ا', '.', '.', '.', 'وقد', 'كان', '.', 'لكنها', 'كانت', 'رومانسية', 'موج', '##عة', '.', 'اصاب', '##ني', 'هذا', 'الكتاب', 'بق', '##شعر', '##يرة', '.', 'وداع', '##بت', 'الدموع', 'عيني', '.', 'كتاب', 'رايع', '.', '.', 'لكنه', 'حزين', 'جدا', 'ومق', '##بض']	['O', 'O', 'B', 'I', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'B', 'I', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'B', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O']

A. Corpus Overview

The constructed corpus consists of 512 reviews collected from the main corpus which is a dataset of reviews written in dialectal Arabic. The dialectal Arabic dataset is utilized since it is more realistic and relatable to everyday life. The primary dataset contains hotel, book, movie, product, and airline reviews. It has three classes (Mixed, Negative and Positive). The mappings are based on reviewer ratings, with 3 ratings representing mixed, above 3 representing positive and below 3 representing negative. ach row in the dataset consists of a tab-separated label and text. The reviews are cleaned by removing non-Arabic characters and Arabic diacritics. There are no duplicate reviews in the dataset. The hotel and book reviews are subsets of the datasets Hotels Arabic Reviews Dataset (HARD) [23], Books Reviews Arabic Dataset (BRAD) [24], and Hady ElSahar [25]. The remaining around 100 airline reviews were acquired manually. Our constructed corpus includes six emotions which are joy, anger, surprise, disgust, fear, and sadness.

B. Corpus Analysis

There are two sentence categories in Arabic: nominal and verbal. Nominal sentences start with a noun or a pronoun, while verbal sentences start with a verb. Nominal sentences consist of two elements : a subject (مبتدأ) and a predicate (خبر) [26]. Subjects of nominal sentences are nouns or pronouns. Predicates may be nouns, adjectives, prepositions, or verbs [26].

Based on our annotated Arabic reviews corpus analysis, some observations are found. Most of the reviews contain more than one cause. Fig.4 illustrates the top 5 observed emotion causes types. The analysis of the emotion causes types are done on 350 reviews of our constructed corpus. Most of causes in the constructed corpus are found to be nominal sentences more than verbal sentences. Some causes are just subjects in a nominal sentence; they can be linked with other words by connectives not a complete nominal sentence. Causes can be just a word which in most cases subject in nominal sentences and object in verbal sentences. Some linguistic cues are observed. Causes can follow some expressions, other expressions can come before causes, others can follow or precede causes. Some causes can come between two words. Fig.5 shows the

top 10 used expressions in our dataset where the most observed expression is (رائع / رائعة) (fantastic) which can come at the beginning of a sentence as a predicate and cause comes after it. It can also come as an adjective after the cause where it describes the cause of joy.

The Arabic reviews dataset contains six emotions which are joy, sadness, anger, surprise, fear, and disgust. As shown in Fig.3, the most dominant emotion in the dataset is joy where the most expression observed is (fantastic) (رائع / رائعة). The anger emotion second comes where the most used expression with such emotion is (ضعيف) which means (low standard) in English. The sadness and surprise emotions come third and fourth with the most used expressions (مخيب للأمل) (depressing) and (استثنائي /استثنائية) (Especial) respectively. The least emotions observed are fear and disgust where the most mentioned expressions are (اتوجس منه) (suspicious of it) and (وقح) (منفر) (rude) (repugnant) respectively.

Tables II, III, IV and V show the observed linguistic cues which are represented with expressions in Dialectical Arabic. Dialectical Arabic is different from Modern Standard Arabic where the standard linguistic rules are not always applied. In the tables, we are showing the expressions observed with their English translation, but this doesn't mean that the order of causes and expressions followed in Arabic is the same followed in English language. For example, an expression like (أصابني بالصدمة) which its English translation is (Shocked me) , the cause (هذا الفندق) come in between this expression أصابني (هذا الفندق بالصدمة) but in English, the cause (The hotel) will precede the expression (This hotel shocked me).

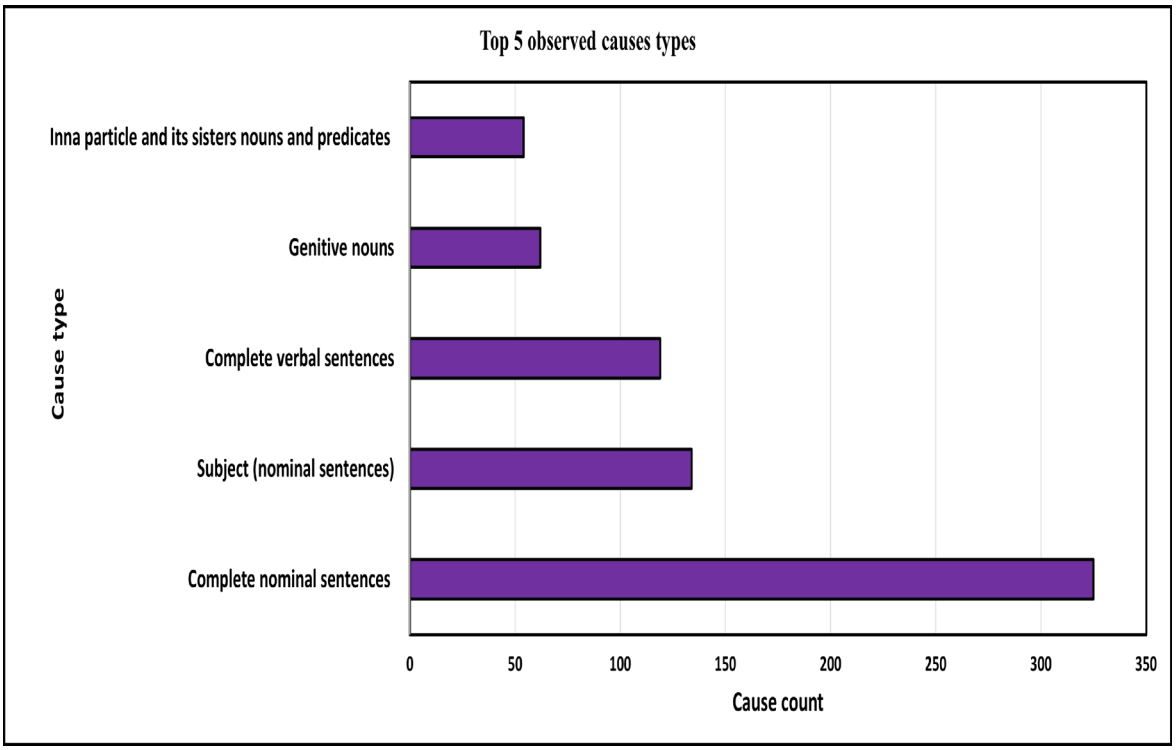

Figure 4: Top 5 emotion causes types.

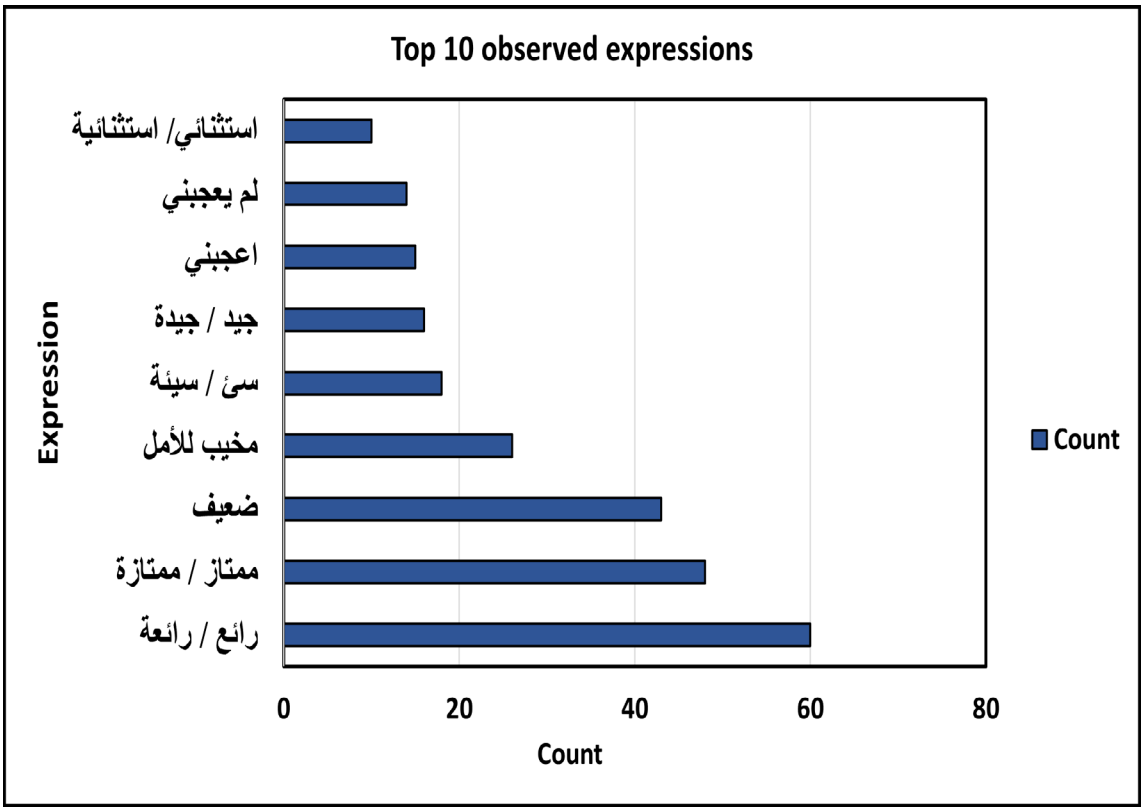

Figure 5: Top 10 observed expressions in our constructed corpus.

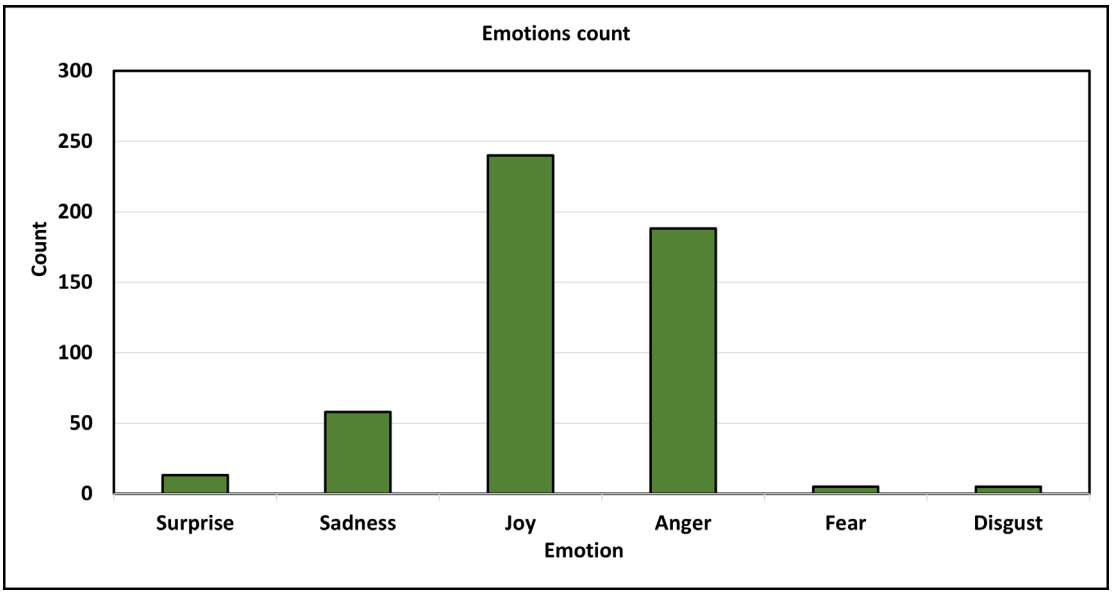

Figure 6: Emotions' count observed in our constructed corpus.

TABLE II

SOME EXPRESSIONS WHERE CAUSES COME IN BETWEEN AND THEIR ENGLISH TRANSLATION.

Arabic	**English**
أصابني بالقشعريرة	Make me feel goosebumps
أصابني بالصدمة	Shocked me

TABLE III

SOME EXPRESSIONS THAT PRECEDE CAUSES AND THEIR ENGLISH TRANSLATION.

Arabic	**English**
استمتعت (بالفعل) (جدا) ب	I have been interested (indeed) (so much) in
أحببت	I have loved
أشكر	Thanks for
سعدت ب	I have been happy
شكرا (جزيلا) ل او لل	Thanks a lot for
ممتاز + ل او لل	Excellent for
اتمني (ان)	I hope to
سبب اختياري هو	The reason for my choice is
أمتعنا ب	We have the fun of
أخاف من	I am afraid of
أتوجس من	I am suspicious of
انفعنا ب	We have benefited from
أحزن مع او ل	I am sad for

TABLE IV

SOME EXPRESSIONS THAT FOLLOW CAUSES AND THEIR ENGLISH TRANSLATION.

Arabic	**English**
يستحق كل تقدير	Deserve all appreciation
فهي مميزة /فهو مميز	It is special
أتوجس منها (منه)	Suspicious of it

TABLE V
SOME EXPRESSIONS THAT CAN FOLLOW OR PRECEDE CAUSES AND THEIR ENGLISH TRANSLATION.

Arabic	English
أعجبتني/أعجبني	I like
ممتاز	Excellent
استثنائي/ استثنائية	Especial
جميل (ة)(جدا)	Beautiful (Very beautiful)
رائع	Fantastic
جيد(ة) (جدا)	Good (Very good)
حلو(ة) (اوي)	Beautiful (Very beautiful)
أكثر ما أعجبني	The most I like

5 DEEP LEARNING MODELS

We have addressed the ECE problem in our experiments using two deep learning models, BiLSTM-CRF with a self-attention layer and BERT-CRF. In this section, we will talk about the models' background and describe the models' architecture.

A. Models Overview

1) BERT

Google introduced BERT in 2018. It is a pretrained model whose substructure is the vanilla transformer language model. BERT has improved the language comprehension level. It has been regarded as a revolution in Natural Language Processing pipeline. BERT pre-trains deep bidirectional representations from unlabeled text by concurrently conditioning on left and right contexts in all layers. The pre-trained BERT model is fine-tuned with just one more output layer to create state-of-the-art models for a variety of tasks. These tasks may include question answering, sequence classification, token classification, and language inference, without the need for significant task-specific architecture adjustments [9].

2) BiLSTM

Bidirectional LSTM (BiLSTM) is a sequence processing model. It consists of two LSTMs layers, the first layer takes the input in a forward direction, and the second one in a backwards direction. The outputs from both LSTM layers are combined in several ways, such as average, sum, multiplication, or concatenation. BiLSTMs effectively utilize information from both directions, and this improves the ability to extract the context. BiLSTMs proposed by [27] are used to access both past and future input features. Using past (through forward states) and future (through backward states) information for a specified time frame helps a lot in sequence labelling tasks. Backpropagation through time (BPTT) is used to train BiLSTM networks [28]. The forward and backward passes over the unfolded network over time performed similarly to conventional network forward and backward passes, with the exception that the hidden states for all time steps must be unfolded.

3) CRF

Conditional Random Fields is a type of discriminative model. It is most suitable for prediction tasks in which contextual information or the neighbor's state influences the current prediction. There are two distinct ways for the using of neighbor tag information in predicting the current tag. The first way is predicting tags distribution for each time step and then using beam-like decoding to find the optimal tag sequences. Maximum Entropy Classifier [29] and Maximum Entropy Markov models (MEMMs) [30] work in such a way. The second way is to focus on sentence-level instead of individual positions which is the work of Conditional Random Fields (CRF) models [31]. The inputs and outputs are connected directly. CRFs have been shown to have the ability to produce higher tagging accuracy in general.

B. Models' Architecture

1) BiLSTM- CRF Model

BiLSTM-CRF model architecture is described as follows (Fig. 7). The embedding layer is the first layer in BiLSTM-CRF model. Each word is represented by a 300-dimension pretrained embedding vector. The longest sentence contains 137 words, so the maximum length is 137. The sentences that have less than 137 words are post-padded with zeros. The second layer is the Bidirectional LSTM layer of 100 neurons and recurrent dropout of rate 0.01**.** The self-attention layer is the third layer with attention width 6 and sigmoid activation function. The following layer is the Time Distributed layer wrapping a dense layer of output neurons equals to the number of output tags (tokens). The last layer is the CRF layer with output neurons equals to the number of output tags (tokens) which is four tags (B, O, I, Pad). Adam optimizer is used, and the loss function is the CRF log-likelihood function. It is computed using "crf_log_likelihood" TensorFlow addons function.

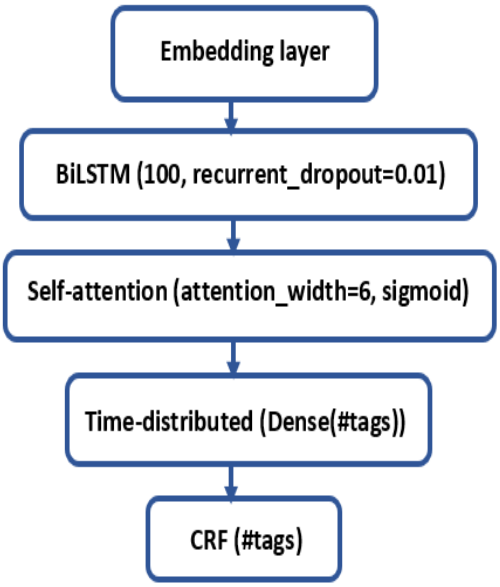

Figure 7: BiLSTM-CRF Model Architecture

2) BERT-CRF Model

We have fine-tuned Arabic BERT model (asafaya/bert-base-Arabic) [32] which is a pretrained BERT base language model for Arabic language. Arabic-BERT-base model was pretrained on about 8.2 billion words which are from Arabic version of OSCAR dataset filtered from Common Crawl and recent dump of Arabic Wikipedia [32]. The architecture of BERT-CRF model is described as shown in Fig.8. The first layer is a pretrained Arabic BERT model. The second layer is a dropout layer of 0.1 rate. Then, the dense layer with number of units equal to the number of tags. Finally, the CRF layer with output neurons equals to the number of output tags.

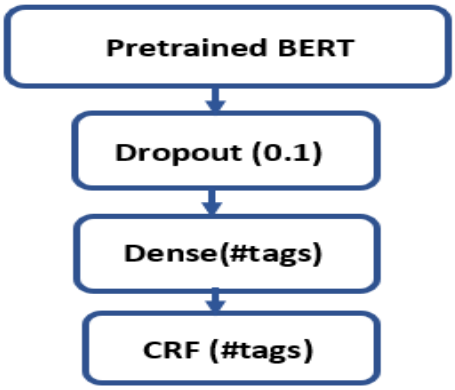

Figure 8: BERT-CRF Model Architecture

6 RESULTS AND DISCUSSION

The experiments have been conducted using Python 3, scikit-learn library, TensorFlow 2 and Keras API. BERT-CRF is implemented using Hugging Face transformers library and PyTorch 1.10.2. All experiments were done using core i7 Processor and 16GB RAM. Models' evaluation metrics used are precision, f-measure, recall using macro-average approach. We have used two evaluation measures to evaluate our models, span-level evaluation measure and token-level evaluation measure. In span detection problems, the evaluation measure can either be based on the number of matching tokens and this evaluation measure is called token-level evaluation measure. It can also be stricter and consider the exact spans and the number of exact matches and this evaluation measure is called span-level evaluation measure. The span-level evaluation measure is represented by equations (1), (2), and (3), where "proposed spans" is the number of emotion-cause spans predicted by the model, "annotated_spans" is the total number of emotion-cause spans labelled in the dataset, and "correct_spans" is the number of spans that are both labelled and predicted. The token-level evaluation measure is where each token (B or I or O) precision is calculated as in eq. (4), recall of each token is calculated as in eq. (5) and F1 score is the tradeoff of both recall and precision as shown in eq. (6). The F1 score of all tokens is calculated as macro-average.

$$P_{\text{span-level}} = \frac{\sum correct\_spans}{\sum proposed\_spans} \quad (1)$$

$$R_{\text{span-level}} = \frac{\sum correct\_spans}{\sum annotated\_spans} \quad (2)$$

$$F1_{\text{span-level}} = \frac{2\ X\ P_{span-level}\ X\ R_{span-level}}{P_{span-level}+R_{span-level}} \quad (3)$$

$$P_{\text{token-level}} = \frac{TP}{TP+FP} \quad (4)$$

$$R_{\text{token-level}} = \frac{TP}{TP+FN} \quad (5)$$

$$F1_{\text{token-level}} = \frac{2\ X\ P_{token-level}\ X\ R_{token-level}}{P_{token-level}+R_{token-level}} \quad (6)$$

We have divided our corpus into 70% of data for training, 20% of data used for validation and 10% of data used for testing using Train-Validation-Test (T-V-T) split training mode. Early stopping is used to know in how many epochs, model will overfit. 5-fold cross validation (CV) is then applied for number of epochs that is detected before having model overfitted.

We have used pretrained Arabic news [33] Word2Vec features for BiLSTM-CRF with self-attention layer. We have tried different number of batch sizes which are 32 and 16 using Adam optimizer. We have trained both BiLSTM-CRF and BERT-CRF models using the different training modes T-V-T and 5-fold CV. BiLSTM-CRF is trained using T-V-T split and 5-fold CV for 27 epochs with batch size 32 and for 23 epochs with batch size 16. Training BiLSTM-CRF for 23 epochs with batch size 16 using 5-fold CV gives us the best results. Training BERT-CRF for 3 epochs using 5-fold CV with batch size 16 also gives us the best results for the model.

Comparison between CV and T-V-T training modes for both BiLSTM-CRF and BERT-CRF models is shown in Fig.9 and Fig.10. The analysis of the difference in performance between CV and T-V-T is done based on different number of parameters which are batch size, training time in minutes and macro-average F1 score. 5-fold CV takes more time in training than T-V-T split generally. The less batch size, the less time taken for training in case of BiLSTM-CRF model and the opposite for BERT-CRF the less batch size, the more time taken for training. The less batch size the better F1 score achieved. Training BERT-CRF model takes more time than BiLSTM-CRF model in general.

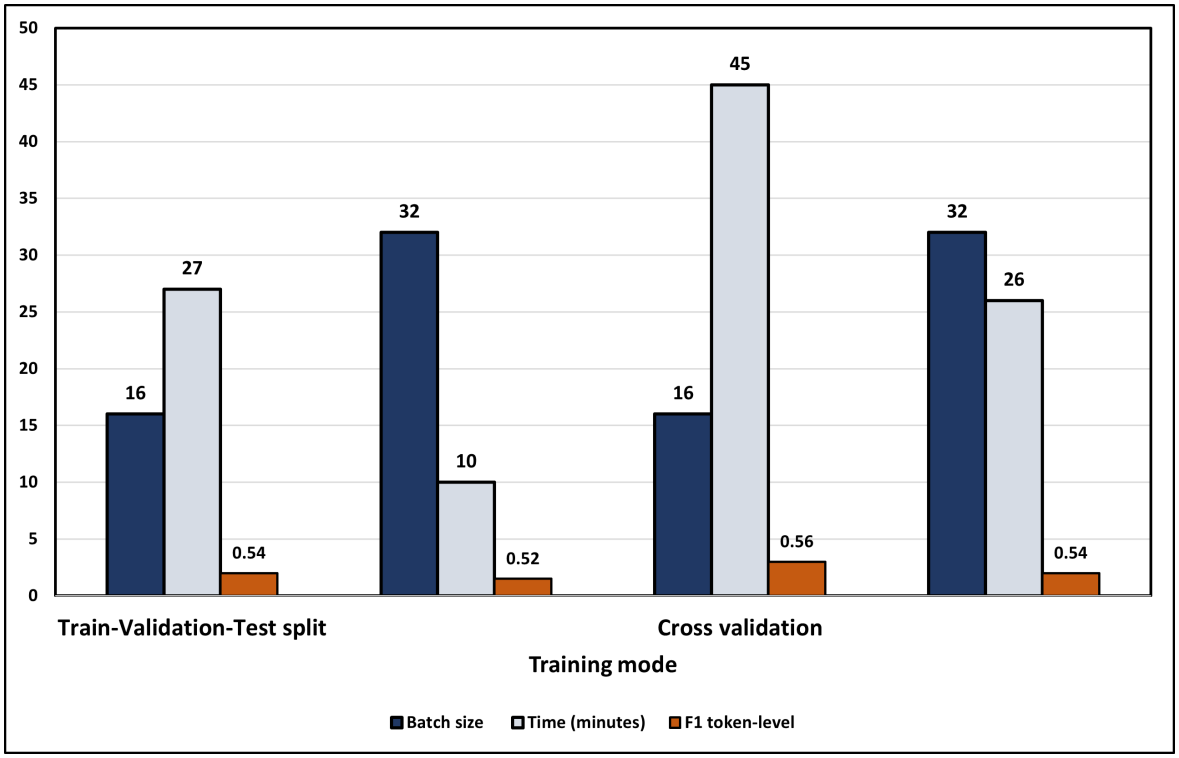

Fig. 9. Training modes comparison for BiLSTM-CRF model.

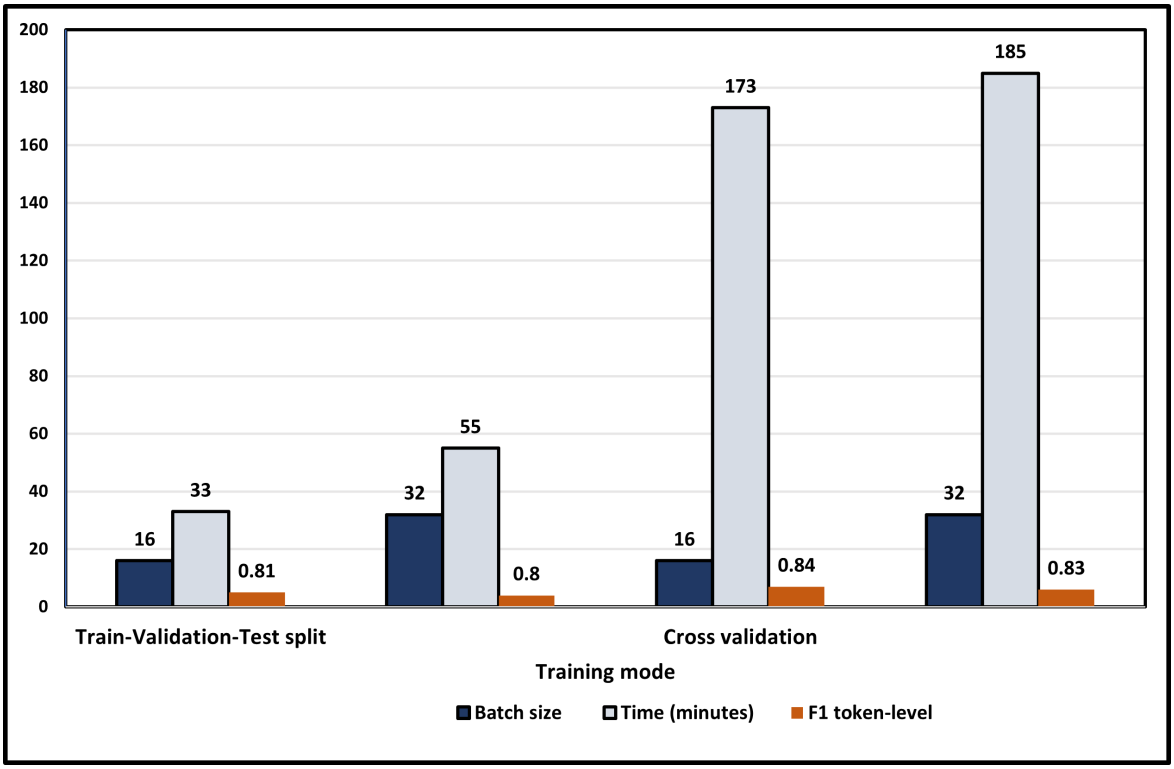

Fig. 10. Training modes comparison for BERT-CRF model.

Best results for both BiLSTM-CRF and BERT-CRF models are illustrated in Table 6 where 5-fold CV training mode is used and batch size is 16. As shown in Table 6, using token-level evaluation in general gives better results than using span-level. BERT-CRF outperforms BiLSTM-CRF in both span-level evaluation resulting in a 0.29 F1 score and token-level evaluation resulting in 0.84 F1 score. Both BERT and BiLSTM have strong contextualized representation abilities, but BERT surpasses BiLSTM. The reason behind getting low span-level evaluation results can be because the context in some reviews in our corpus is long. We can improve both the span-level and token-level evaluation results by increasing the size of our corpus.

TABLE 6

BERT-CRF AND BILSTM-CRF F1 SCORE (F1), RECALL (R) AND PRECISION (P)

Algorithm	Evaluation measure	P	R	F1
BiLSTM with CRF	Span-level	0.1	0.14	0.12
	Token-level	0.55	0.6	0.56
BERT with CRF	Span-level	0.33	0.25	0.29
	Token-level	0.87	0.81	0.84

7 CONCLUSION AND FUTURE WORK

In this paper, we have built an annotated Arabic corpus for ECE task. The constructed corpus consists of 512 reviews that contains six emotions, joy, anger, surprise, disgust, fear, and sadness. We have observed some linguistic cues that help in extracting causes. Some expressions can precede causes, other expressions can follow causes, others can be found before or after causes and causes can be found in between some other expressions. We have addressed the ECE problem as sequence labelling task implementing two models BiLSTM-CRF and BERT-CRF showing that BERT-CRF outperforms BiLSTM-CRF using both span-level measure and token-level measure evaluation. BERT-CRF takes more time during training. Training is done using two different modes T-V-T and CV modes. CV training leads to better results but it takes more time in training. Token-level evaluation give better results than span-level evaluation in both models. In future, we plan to increase the size of our corpus and try different techniques to improve the performance and achieve better performance.

8 REFERENCES

[1] N. M. Hakak, M. Mohd, M. Kirmani, and M. Mohd, "Emotion analysis: A survey," in *2017 international conference on computer, communications and electronics (COMPTELIX)*, pp. 397–402, Jaipur, India, July,2017.

[2] L. Gui, R. Xu, Q. Lu, D. Wu, and Y. Zhou, "Emotion cause extraction, a challenging task with corpus construction," in *Chinese National Conference on Social Media Processing*, vol.1, pp. 98–109, Nanchang, China, October,2016.

[3] I. Russo, T. Caselli, F. Rubino, E. Boldrini, and P. Martínez-Barco, "Emocause: an easy-adaptable approach to emotion cause contexts," In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA 2.011)*, vol.1, pp. 153–160, Portland, Oregon, June 2011.

[4] R. Xu, J. Hu, Q. Lu, D. Wu, and L. Gui, "An ensemble approach for emotion cause detection with event extraction and multi-kernel svms," *Tsinghua Sci. Technol.*, vol. 22, no. 6, pp. 646–659, 2017.

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

[6] X. Li, W. Gao, S. Feng, Y. Zhang, and D. Wang, "Boundary detection with BERT for span-level emotion cause analysis," *Findings of the Association for Computational Linguistics:ACL-IJCNLP 2021*,vol.1, pp. 676–682, Online, , August, 2021.

[7] S. Poria *et al.*, "Recognizing emotion cause in conversations," *Cognit. Comput.*, vol. 13, no. 5, pp. 1317–1332, 2021.

[8] N. Y. Habash, "Introduction to Arabic natural language processing," *Synth. Lect. Hum. Lang. Technol.*, vol. 3, no. 1, pp. 1–187, 2010.

[9] "Top Ten Internet Languages in The World - Internet Statistics." https://www.internetworldstats.com/stats7.htm (accessed May 08, 2022).

[10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv Prepr. arXiv1810.04805*, 2018.
[11] A. Ortony, G. Clore, and A. Collins, *The Cognitive Structure of Emotions*. Cambridge: Cambridge University Press, 1988.doi:10.1017/CBO9780511571299.
[12] L. Gui, L. Yuan, R. Xu, B. Liu, Q. Lu, and Y. Zhou, "Emotion cause detection with linguistic construction in chinese weibo text," in *CCF International Conference on Natural Language Processing and Chinese Computing*,vol.1,pp.457–464, Shenzhen, China, December, 2014.
[13] D. Ghazi, D. Inkpen, and S. Szpakowicz, "Detecting emotion stimuli in emotion-bearing sentences," in *International Conference on Intelligent Text Processing and Computational Linguistics*,vol.2, pp. 152–165, Cairo, Egypt, April, 2015.
[14] L. Xu, H. Lin, Y. Pan, H. Ren, and J. Chen, "Constructing the affective lexicon ontology," *J. China Soc. Sci. Tech. Inf.*, vol. 27, no. 2, pp. 180–185, 2008.
[15] L. Bostan, E. Kim, and R. Klinger, "Goodnewseveryone: A corpus of news headlines annotated with emotions, semantic roles, and reader perception," *arXiv Prepr. arXiv1912.03184*, 2019.
[16] L. A. M. Oberländer and R. Klinger, "Token sequence labeling vs. clause classification for English emotion stimulus detection," in *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics*, vol.1, pp. 58–70, Barcelona, Spain, December, 2020.
[17] L. Gui, R. Xu, D. Wu, Q. Lu, and Y. Zhou, "Event-driven emotion cause extraction with corpus construction," In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1639–1649, Austin, Texas, United States, November, 2016.
[18] M. E. Peters *et al.*, "Deep contextualized word representations," *arXiv Prepr. arXiv1802.05365*, 2018.
[19] S. Y. M. Lee, Y. Chen, and C.-R. Huang, "A text-driven rule-based system for emotion cause detection," in *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text*, pp. 45–53, Los Angeles, California, United States, June 2010.
[20] X. Xiao, P. Wei, W. Mao, and L. Wang, "Context-aware multi-view attention networks for emotion cause extraction," in *2019 IEEE International Conference on Intelligence and Security Informatics (ISI)*, vol.1, pp. 128–133, Shenzhen, China, July,2019.
[21] Z. Yu, Y. Wang, Z. Liu, and X. Cheng, "EmotionX-Antenna: An emotion detector with residual GRU and text CNN," *EmotionX 2019 Challenge, The 7th International Workshop on Natural Language Processing for social media (Social NLP),* Technical report, Macau, China, August 2019.
[22] Y. Chen, S. Y. M. Lee, S. Li, and C.-R. Huang, "Emotion cause detection with linguistic constructions," in *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pp. 179–187, Beijing, China, August,2010.
[23] "Arabic 100k Reviews | Kaggle." https://www.kaggle.com/datasets/abedkhooli/arabic-100k-reviews (accessed June 08, 2022).
[24] M. Saad, "Mining Documents and Sentiments in Cross-lingual Context," Thesis, Université de Lorraine, Metz, France, February 2015.
[25] E. F. Sang and J. Veenstra, "Representing text chunks," *arXiv Prepr. cs/9907006*, 1999.
[26] K. C. Ryding, *A reference grammar of modern standard Arabic*. Cambridge university press, 2005.
[27] A. Graves, N. Jaitly, and A. Mohamed, "Hybrid speech recognition with deep bidirectional LSTM," in *2013 IEEE workshop on automatic speech recognition and understanding*,pp. 273–278, Olomouc, Czech Republic, December, 2013.
[28] M. Boden, "A guide to recurrent neural networks and backpropagation," *Dallas Proj.*, vol. 2, no. 2, pp. 1–10, 2002.
[29] A. Ratnaparkhi, "A maximum entropy model for part-of-speech tagging.," in 1996 *Conference on Empirical Methods in Natural Language Processing (EMNLP 1996)*, vol. 1, pp. 133–142, Philadelphia, Pennsylvania, United States, May, 1996.
[30] A. McCallum, D. Freitag, and F. C. N. Pereira, "Maximum entropy Markov models for information extraction and segmentation.," in *Icml*, 2000, vol. 17, no. 2000, pp. 591–598.
[31] J. Lafferty, A. McCallum, and F. C. N. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," *in Proceedings of the Eighteenth International Conference on Machine Learning (ICML '01)* , pp.282–289,San Francisco, CA, United States, June, 2001.
[32] A. Safaya, M. Abdullatif, and D. Yuret, "Kuisail at semeval-2020 task 12: Bert-cnn for offensive speech identification in social media," in *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pp. 2054–2059, [redacted]December,2020.
[33] A. A. Altowayan and L. Tao, "Word embeddings for Arabic sentiment analysis," in *2016 IEEE International Conference on Big Data (Big Data)*, pp. 3820–3825,Washington, DC,United States, December,2016.

BIOGRAPHY

Yasmin Shaaban received B.Sc. degree in Computer Engineering from Faculty of Engineering, Ain Shams University 2018. She is currently a M.Sc. student in Computer Engineering at Faculty of Engineering, Ain Shams University. She is currently working as a Software Engineer at Ejada Systems Ltd. Company.

Hoda K. Mohamed is a professor at computer and systems engineering Faculty of Engineering, Ain Shams University from 2009 till now. B.Sc. from Faculty of Engineering, Ain Shams University 1978. Msc. from Faculty of Engineering, Ain Shams University 1983, PHD from Faculty of Engineering, Ain Shams University 1992, Assoc. Prof. 2001. Research area are on intelligent systems, E-learning systems, Data Mining, Database systems, Software Engineering, Natural Language processing, Cloud computing, and Image processing. I am a reviewer in IEEE International Conference on Computer Engineering and Systems, Cairo, Egypt.

Walaa Medhat is graduated from Faculty of Engineering, Ain Shams University from computer systems department and got her MSc and PhD from the same department 2002, 2008 and 2015 respectively. She is an Assistant Professor in Nile University since 2019 and affiliated to Benha University since 2017. She is now CS program director in Nile University. The research interest is in Natural Language Processing and software Engineering. She is a member in Egyptian Language Engineering society. She has supervised more than 15 MSc and Ph.D. thesis in the scope of natural language processing, text mining, data mining, machine learning, and big data analytics. She is the author and co-author of more than 13 academic articles in international journals and conferences related to natural language processing, semantic web, text and data mining, machine learning, and big data analytics.

ARABIC ABSTRACT

إستخراج أسباب العاطفة باستخدام التعلم العميق بالعربية

[*1] ياسمين شعبان ،[*2]هدي قرشي محمد، [**3] ولاء مدحت

[*] قسم هندسة الحاسب والنظم ، كلية الهندسة ، جامعة عين شمس ، السرايات العباسية ، القاهرة ، مصر

g18092931@eng.asu.edu.eg[1]

hoda.korashy@eng.asu.edu.eg[2]

[**] قسم تكنولوجيا المعلومات وعلوم الحاسب / قسم كلية الحاسبات و الذكاء الاصطناعي ، جامعة النيل / جامعة بنها ، الجيزة / بنها ، مصر

wmedhat@nu.edu.eg[3]

الملخص

يعتبر استخراج سبب العاطفة مهمة صعبة في الوقت الحاضر. يتم استخراج الأسباب الكامنة وراء العواطف من البيانات النصية. يحتوي استخراج سبب العاطفة على العديد من التطبيقات مثل استخراج الأسباب من المراجعات المستخرجة من الشبكات الاجتماعية ومواقع التوصية حيث يقدم المستخدمون ملاحظاتهم. الموارد في هذا المجال محدودة. هناك بعض المجموعات التي تم إنشاؤها للغات الغربية مثل الإنجليزية ولغات الشرق الأقصى مثل الصينية. موارد اللغة العربية في هذا المجال محدودة للغاية. يقدم هذا البحث اكتشاف العواطف المسببة في اللغة العربية. تم إنشاء مجموعة مشروحة باللغة العربية باللهجة لغرض استخراج اسباب العاطفة. البيانات التي تم جمعها من العديد من الموارد. يتم تطبيق تقنيات وسم التسلسل مع مخطط IOB2 باستخدام خوارزمية BiLSTM-CRF وخوارزمية BERT-CRF . يتفوق BERT-CRF على BiLSTM-CRF في كل من تقييم مستوى الامتداد ومستوى الرمز المميز. يحقق BERT-CRF درجة F1 0.29 في حالة تقييم قياس مستوى الامتداد ودرجة F1 0.84 في حالة تقييم مقياس مستوى الرمز المميز.

الكلمات المفتاحية: ***معالجة اللغة الطبيعية، تحليل المشاعر، استخراج سبب العاطفة، وسم التسلسل، التعلم العميق***