


Multi-Label Learning for Aspect Category Detection of Arabic Hotel Reviews Using AraBERT

Asma Ameer^{1,2} ^a, Sana Hamdi² ^b and Sadok Ben Yahia³ ^c

¹*Polytechnic School of Tunisia, Tunisia*

²*Faculty of Sciences of Tunis, Tunisia*

³*Tallinn University of Technology, Estonia*

Keywords: Aspect Detection, Contextual Embedding, AraBERT, Imbalanced Data, Arabic Hotel Reviews.


Abstract: Studying people's satisfaction with social media is vital to understanding the users' needs. Nowadays, textual hotel reviews are used to evaluate the hotel's e-reputation. In this context, we are interested in Aspect Category Detection (ACD) as a subtask of aspect-based sentiment analysis. This task needs to be investigated through multi-label classification, which is more challenging, in natural language processing, than single-label classification. Our study leverages the potential of transfer learning with the pre-trained AraBERT model for contextual text representation. We are based on the Arabic SemEval-2016 data set for hotel reviews. We propose a specific preprocessing for this Arabic reviews dataset to improve the performance. In addition, as this data suffers from an imbalanced distribution, we use a dynamically weighted loss function approach to deal with imbalanced classes. The carried-out results outperform the pioneering state-of-the-art of the Arabic ACD with an F_1 score of 67.3%.


1 INTRODUCTION


Social media has become an essential part of our daily life as a tool for communication in various situations (Hamdi et al., 2022). In particular, it encourages emotional self-expression towards the hotels by providing the user's reviews on opinion websites such as Tripadvisor, Booking.com, etc. This textual data through the analysis system helps measure the users' satisfaction towards the visited hotels. These analyses can be classified into three levels: document, sentence, and aspect. The fine-grained level is called Aspect Based Sentiment Analysis (ABSA) (Pontiki et al., 2016a), which aims to provide precise information regarding each aspect.

Most of the published research on ABSA is in English. However, there is a lack of studies in Arabic, whereas Arabic is the official language of 22 countries and presents many speakers. Few works have been published because of the complexity of Arabic morphology and a lack of data resources (Guellil et al., 2021), making the ABSA tasks more challenging.

That's why we are interested in this study on the ACD of the Arabic hotel reviews as a subtask of the ABSA. Indeed, this task aims to detect the topics (subjects) discussed by the reviewer to fix the aspect category in the comments about the hotels. In this context, we aim to apply a Multi-label Classification (MLC) to extract the different categories of Arabic hotel reviews. Solving MLC problems can be managed using different techniques, such as Problem Transformation and algorithms adaptation methods. The PT aims to transform a multi-label problem into one or more single-label problems (as in the case of Binary Relevance and Classifier Chains). At the same time, the algorithms adaptation seeks to modify an algorithm directly for the multi-label predictions. Recently, multiple models using Deep Learning (DL) based on the pre-trained models have been applied based on the attention mechanism (Vaswani et al., 2017). However, these techniques are under-investigated for the ACD task and are still in their early stages, especially in Arabic. This paper introduces an approach for the Arabic ACD task of the hotel reviews based on the AraBERT fine-tuning. This pre-trained model provides dynamic contextual word embedding for Arabic. The proposed approach is evaluated using the Arabic SemEval-2016 dataset for the hotel reviews.

^a  <https://orcid.org/0000-0002-2175-9310>

^b  <https://orcid.org/0000-0001-6439-2275>

^c  <https://orcid.org/0000-0001-8939-8948>

Leveraging the Arabic SemEval-2016 dataset for this MLC task, we note that the data suffer from a skewed distribution. We propose an approach of Dynamically Weighted Loss Function (DWLF) to deal with this imbalanced multi-learning dataset.

The contributions of this study are summarized as follows: (i) proposing specific data preprocessing for Arabic for this ACD task of the hotel reviews; (ii) investigating the contextual semantic embedding with the AraBERT fine-tuning on the SemEval-2016 for hotels; and (iii) proposing the DWLF to deal with the issue of the imbalanced labels to improve the model performance.

The remainder of this paper is organized as follows: Related works of ACD, using the Arabic SemEval-2016 dataset for hotel reviews, are summarized in Section 2. Next, our proposed AraBERT-based approach is thoroughly presented in Section 3. Next, Section 4 discusses the harvested results. Finally, Section 5 concludes the paper and sketches issues for future work.

2 RELATED WORKS

The ABSA tasks and precisely the ACD subtask could be investigated based on shallow ML and DL-based methods. The ML-based techniques are effective, but they rely on handcrafted features such as lexicons to well train the classifier. Recently, different methods based on neural networks and word embedding layers were developed to provide better results. It is paramount to mention that the development of such studies is not that much developed for the Arabic language (Al-Dabet et al., 2021).

This section presents the related works for the Arabic ACD using the SemEval-2016 dataset. The baseline model is based on the Support Vector Machine (SVM) classifier and provides an F_1 of 40.33%. This ACD task has paid more attention by emerging the pioneering work of INSIGHT-1 method (Ruder et al., 2016). This study proposed an MLC model based on the CNN and GloVe representation to extract the most informative features. The latter work achieves an F_1 improvement of 11.78%. The UFAL model is then proposed in (Tamchyna and Veselovská, 2016) by investigating the MLC with a binary classification for many languages, including Arabic. This model is based on the Long Short-Term Memory network (LSTM), which helps to detect the long text's distance relationship. The UFAL result outperforms the baseline model by 12.26%. Recently, an MLC model is proposed in (Al-Dabet et al., 2021) based on the Binary Relevance (BR) classification and achieves

an F_1 of 58.05%. This is based on CNN and the Independent Long-Short Term Memory (IndyLSTM) (Gonnet and Deselaers, 2020). This model can extract local and sequential features to learn long text dependencies to explore the final sentence representation.

Table 1 sketches these surveyed works based on the following criteria:

- **Preprocessing:** indicates whether or not data preprocessing was conducted.
- **Features:** presents the used word embedding.
- **Models:** enumerates the employed models.
- **Imbalanced data:** checks whether or not authors have dealt with imbalanced data.
- **F_1 score:** presents the achieved F_1 score.

As underscored in Table 1, the preprocessing, features representation, and the imbalanced data issue were not well examined during the recent studies of the Arabic ACD. Still, these features are crucial steps in Natural Language Processing (NLP). Starting with the preprocessing criterion, none of the authors in (Al-Dabet et al., 2021), and (Tamchyna and Veselovská, 2016) have proposed specific preprocessing techniques for the Arabic SemEval-2016 dataset, despite its compelling necessity. Nevertheless, authors in (Pontiki et al., 2016a) pay attention to only stop words removed and (Ruder et al., 2016) uses tokenization. As the morphology of the Arabic language is deeply rooted, specific preprocessing for the Arabic language is proposed in this study. Moreover, we note that all these previous research for ACD with SemEval-2016 did not investigate dynamic contextual embedding. However, it is critical to detect the context as the static feature representation can provide a loss of information.

Regarding the imbalance data problem, all of the mentioned related works in Table 1 do not try to solve this issue when investigating the Arabic SemEval-2016 dataset for hotel reviews. However, this issue poses a severe challenge to predictive modeling because learning algorithms will be biased toward the majority class than other samples in the data. Indeed, most ML algorithms are based on the inherent assumption of balanced data (the data is equally distributed among all its classes). The DL models have recently achieved excellent learning success but still cannot escape the negative impact of imbalanced data (Huang et al., 2016). However, when training a model with an imbalanced dataset, the learning becomes biased toward most classes. In this way, the model performs in the majority classes and fails to learn meaningfully in the minority classes due to a lack of examples for these categories.

Several techniques have been explored to mitigate the imbalance of class impact. We mention

Table 1: Comparison of the studies using the SemEval-2016 dataset for Arabic hotel reviews.

References	Preprocessing	Features	Models	Imbalanced data	F_1 score in%
(Pontiki et al., 2016a)	Stop-words	N-grams	SVM	No	40.33
(Ruder et al., 2016)	Tokenization	GloVe	CNN	No	52.11
(Tamchyna and Veselovská, 2016)	Not declared	Word2Vec	LSTM	No	52.59
(Al-Dabet et al., 2021)	Not declared	AraVec	IndyLSTM	No	58.05

the data level re-sampling (over-sampling and under-sampling), and the algorithm level with the cost-sensitive for re-weighting learning (Cui et al., 2019). For the re-sampling methods, the number of examples is directly adjusted by over-sampling the minor class, under-sampling the major class, or both. On the one hand, under-sampling the majority class can remove certain samples associated with the majority classes. This could lead to the model missing out on learning certain essential concepts from these removed samples. On the other hand, oversampling the minority classes entails the repetition of samples associated with the minority classes. This could quickly slow the training and lead to overfitting in a model. Consequently, the under-sampling can be preferred over the over-sampling, as underscored in (Drummond et al., 2003). Yet neither method directly solves the problem of unequal classes, and both can be risky because they can cause new problems. Among the most used techniques for data resampling, we cite the Synthetic Minority Oversampling Technique (SMOTE) (Chawla et al., 2002) and its variants, the Multi-Label Synthetic Minority Oversampling Technique (MLSMOTE) (Charte et al., 2015). They are based on an interpolation strategy that uses the nearest neighbors of samples to oversample minority instances by averaging between them. Indeed, the MLSMOTE is helpful for multi-label sampling, but this method presents severe flaws and does not work well with textual data.

The cost-sensitive approach is an alternative to avoiding these issues by directly proposing a penalty for minority misclassification (Huang et al., 2016). As justified in (Kaur et al., 2019), these learning techniques aim to find the costs associated with the misclassified examples. This option represents a classical method in statistics that assigns higher misclassification costs to the minority class than to the majority. Sampling methods are easy to implement and more popular than cost-sensitive learning, but the latter is considered a more computationally effective technique (Kaur et al., 2019).

To downplay the disadvantages of re-sampling, we are interested in this study on the techniques of the cost-sensitive solution for the imbalanced class of

the Arabic SemEval-2016 for hotel reviews. Consequently, our enhanced proposal is based on weighing the samples. This assigned weights to samples to match a given data distribution of the Arabic SemEval-2016 dataset. In this context, we deal in this study with the different challenges discussed in this section using the Arabic SemEval-2016 dataset for hotel reviews. More details of this proposed approach are described in the next section.

3 METHODOLOGY

The proposed approach leverages the critical technical innovation of pre-trained language modeling based on AraBERT fine-tuning. In this context, our proposed architecture for this ACD task is mainly broken down into three components: 1- Data processing for data preprocessing and tokenization. 2- AraBERT fine-tuning for feature extraction, and 3- Classification model for aspect category prediction combined with the proposal of a DWLF-based re-weighting strategy to deal with the imbalanced class. Figure 1 presents the overall framework of this proposed approach. It's important to note that this study is evaluated using the Arabic SemEval-2016 dataset for hotels.

3.1 Arabic SemEval-2016 Dataset

This data involves the ABSA's multilingual tasks in 8 languages and seven domains. The SemEval 2016, an international workshop in NLP, introduced multilingual datasets, a total of 39 datasets from seven domains and eight languages for the ABSA task (Pontiki et al., 2016a). It included datasets of restaurants, hotels, laptops, mobile phones, museums, digital cameras, and telecommunication domains in English, Arabic, Spanish, French, Chinese, Dutch, Turkish, and Russian.

This study uses the SemEval-2016 dataset for Arabic hotel reviews¹. The latter contains a set of Arabic

¹<https://github.com/msmadi/ABSA-Hotels>

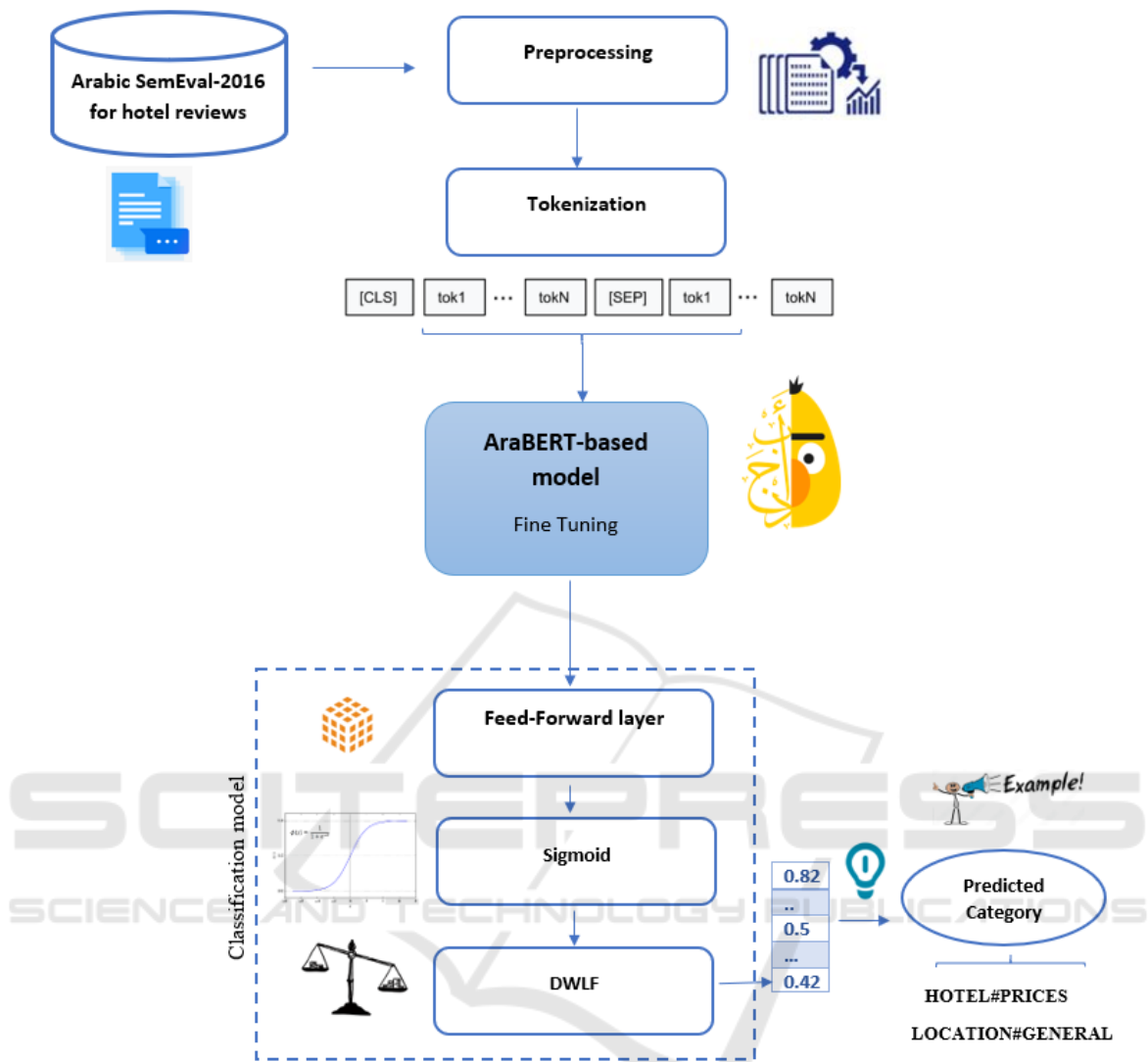


Figure 1: The architecture of the proposed approach of the Arabic ACD for the hotel reviews.

hotel reviews, where each review presents several sentences. For each sentence, we have three tuples: aspect category, aspect target, and aspect polarity. Figure 3 shows an example of hotel reviews in Arabic from the SemEval-2016 dataset. Each category is defined as a pair of entities (E) and attributes (A) providing a unified unit (E#A) (Al-Dabet et al., 2021; Pontiki et al., 2016a). The E and A terms are already defined, depending on the domain. For example, in the hospitality industry, the entities can be ‘HOTELS,’ ‘FACILITIES,’ ‘LOCATION,’ etc. The attribute labels can be ‘CLEANLINESS,’ ‘GENERAL,’ ‘QUALITY,’ etc. As presented in Figure 2, the total number of predefined aspect categories in this dataset is 34.

Each sentence can be assigned to more than one

	GENERAL	PRICES	DESIGN& FEATURES	CLEANLINESS	COMFORT	QUALITY	STYLE& OPTIONS	MISCELLANEO US
HOTEL	✓	✓	✓	✓	✓	✓	x	✓
ROOMS	✓	✓	✓	✓	✓	✓	x	✓
ROOM_ AMENITIES	✓	✓	✓	✓	✓	✓	x	✓
FACILITIES	✓	✓	✓	✓	✓	✓	x	✓
SERVICE	✓	x	x	x	x	x	x	x
LOCATION	✓	x	x	x	x	x	x	x
FOOD& DRINKS	x	✓	x	x	x	✓	✓	✓

Figure 2: The possible E#A pairs for the Arabic SemEval-2016 dataset for hotels (Pontiki et al., 2016b).

category, as expressed in Figure 3. In this context, this sentence provides information about the two cate-

```

<Review rid="890">
  <sentences>
    <sentence id="890:0">
      <text>فندق ممتاز عالي شوي بس ايجابياته كثيره.</text>
      <Opinions>
        <Opinion target="فندق" category="HOTEL#QUALITY" polarity="positive" from="0" to="4"/>
        <Opinion target="NULL" category="HOTEL#PRICES" polarity="negative" from="0" to="0"/>
      </Opinions>
    </sentence>
    <sentence id="890:1">
      <text>فيه بعض النواقص في الغرف يتم تداركها اذا طلبتها .. المواقف صعبه ولاتتناسب مع فندق خمس نجوم</text>
      <Opinions>
        <Opinion target="الغرف" category="ROOMS#MISCELLANEOUS" polarity="negative" from="19" to="24"/>
        <Opinion target="المواقف" category="FACILITIES#GENERAL" polarity="negative" from="51" to="58"/>
      </Opinions>
    </sentence>
  </sentences>
</Review>

```

Figure 3: An example of the Arabic SemEval-2016 dataset for hotel reviews.

Table 2: The distribution of the Arabic SemEval-2016.

	Reviews	Sentences	Tuples
Training data	1,839	4,802	10,509
Testing data	452	1,227	2,604

gories 'HOTEL#PRICES' and 'HOTEL#QUALITY'. The dataset distribution is described in Table 2.

To facilitate this MLC, some data transformations are applied with an encoding of the label categories for each review. If the category is verified in the review, we assign 1 else, 0, as presented in Figure 4.

3.2 Data Processing

To prepare the textual data, preprocessing is essential for the NLP task. However, it depends on one application to another one. This is critical, especially for the Arabic language, as it is a morphologically rich language with several characteristics (Oudah et al., 2019). Indeed, the preprocessing of Arabic is a challenging task compared to other languages. In this context, we note that the order of applying the preprocessing for Arabic text impacts the final result of the model.

Among the considered Arabic text preprocessing, we do the data cleaning, dealing with repeated letters, emojis, stop words, etc. In our study, several preprocessing steps are proposed and investigated, as presented in Figure 5.

- **Data Cleaning:** This includes text normalization, removing the URLs, numbers, mentions, HTML, diacritics, and extra white space. Also, letter normalization was applied to unify the different letters in the Arabic language, such as [أأأ].

- **Repeated Letters:** To better normalize the text, we deal with repeated characters, which can be important to express an intense meaning such as [جميبيبييل] (beautiful). To confirm the mentioned sentiment, we

replace the repeated letter with the term [جدا] (very) to become [جميل جدا] (very beautiful).

- **Emoticons Transformation:** For social media in general, and the hotel opinion website in particular, the emojis preprocessing represents a challenging problem. Replacing the emojis with their Arabic descriptions can improve the model by providing more detailed meaning. So, we transform the emoticons into their Arabic meaning words based on a dictionary that we develop manually. Based on the emoticon list in Wikipedia², we try to cluster the used emoticons that resemble the expressed meaning. Some examples can be summarized in Figure 6.

- **Stemming:** With the morphologically rich Arabic language, we use the stemming technique to reduce a word to its word root. In our study, we apply Farasa, among the recommended techniques that outperform the state-of-the-art (Abdelali et al., 2016).

- **Stop Words Dealing:** To minimize the non-subjective vocabulary in our corpus, we investigated the stop words. Indeed, the default Arabic stop words list has been removed except for the terms expressing negation or intensity meaning as the case of the following terms [لا، لن، ليس، لم]. These terms play a crucial role in the text's information. So, we remove stop words that do not express any opinion that can convert the context.

As in the example from the user data set, we mention this review [لا توجد أضياء ولا مناشف]; where the word's signification will be changed if we remove the term [لا].

All these mentioned techniques are applied to the reviews to improve the performance of the classifier model.

- **Tokenization:** Followed to the data preprocessing, the textual data is then tokenized to be adapted with

²https://en.wikipedia.org/wiki/List_of_emoticons

text	FACILITIES#CLEANLINESS	FACILITIES#COMFORT	FACILITIES#DESIGN_FEATURES	FACILITIES#GENERAL
..فعلا هذا المكان الصغير في حجمه الكبير في خدماته وجماله يجعلك تقضي أياما لا تنسى فموقعه ممتاز	0	0	1	0
..الفندق هادئ .. العامالين ودوديين جدا .. اذا كانت زيارتك للاسكندريه تتطلب فقط مكانا نظيف ومرتب	0	0	0	0
بعد وعود من احد موظفي الفندق بالخدمة الرائعة والطلبات المجابة لكي بعد ان قدمت للفندق ودفعت	0	0	0	0
، تنظيف الغرف ، الشاطي ، راعع وواسع ، حمامات السياحه مريحه والمياه نقيه ، السونافوق الممتازة ،	0	1	1	1
،اللاجة سيئة وطعام رخيص وخدمة رخيصة كذلك ومكيف هواء ضعيف وليس نظيف تماما، لا توجد خيا	0	0	0	0
، صعوبة الوصول للغرفة وتبديل المصاعد ،	0	0	0	1

Figure 4: The Arabic SemEval-2016 dataset after the format transformation.



Figure 5: The proposed preprocessing for the AraBERT-based approach of the Arabic ACD task.

Icons	Emoticons in English	Emoticons in Arabic
:-) :D =D	smiley	مبتسم
:-D :D =D	laughing	ضاحك
:(;(:-(sad	حزين
:'-(:'(:=(crying	بكاء
>:(>:[angry	غاضب

Figure 6: Examples of the emoticons translation from English to Arabic.

the pre-trained-based model. Indeed, it is essential to convert the input data into an appropriate format to be sent to the pre-trained AraBERT model to obtain the corresponding embedding. This step concerns replacing the considered data with a unique identification that retains all the essential information.

Dealing with the AraBERT model, each token in the input sentence is mapped to its corresponding unique IDs using the pre-trained vocabulary. In this context, it is essential to mention that when applying a pre-trained model to some other data, some tokens in the new dataset can not appear in the fixed vocabulary of the pre-trained model. This problem, known as the Out-Of-Vocabulary (OOV), is resolved with the advantage of the BERT pre-trained model. Indeed, BERT uses the WordPiece algorithm, which aims to break a word into several subwords, where the model can commonly represent subwords. However, this tokenization is based on converting these tokens to the BERT’s format and adding the specials token [CLS] and [SEP], respectively, at the beginning of each text and between sentences.

This proceeded data is then used for the features representation based on the contextual word representation with AraBERT fine-tuning.

3.3 Contextual Words Embedding

Several techniques can be used for word representation in NLP, such as TF-IDF, n-grams, and word em-

bedding. Nowadays, embedding techniques present a vital role in model performance. They are used to capture the semantic relations between words. These techniques can be regrouped into static word embedding (Word2Vec, FastText, etc.), and contextual embedding such as BERT in (Devlin et al.,), and Embeddings from Language Models (ELMo) in (Matthew et al., 2018).

Recently, pre-trained language models have shown an essential role in NLP, such as ELMo, GPT (Radford et al., 2018), and BERT. BERT is one of the most popular pre-trained language models armed with Transformers (Vaswani et al., 2017). It is defined as DL techniques for NLP that use unsupervised language representation and bidirectional models. It considers a word’s context from both the left and right side simultaneously (Devlin et al.,). From a sequence, BERT extracts more context features than training left and right separately, as in the case of ELMo. BERT is valid for ABSA, but its Arabic models’ research is still slower than English. For this objective, we investigate the power of BERT, mainly using AraBERT (Antoun et al., 2020) for the Arabic language.

3.4 AraBERT Fine-Tuning

In this study, we use the AraBERT, as an Arabic pre-trained language model based on Google’s BERT architecture. AraBERT is a multi-layer bidirectional transformer encoder that uses the same BERT-Base configuration and is retrained in Arabic. The experimental evaluation of this study was investigated with the AraBERT-v02. This version of the pre-trained model was trained on 200M sentences with a size of 77GB and 8.6B words. It has 12 encoder blocks, 768 hidden dimensions, 12 attention heads, 136M parameters, and 512 maximum sequences.

In our study, the AraBERT is fine-tuned on the downstream task of ACD using Arabic hotel reviews.

Following the data preprocessing, the tokenization process is then applied. For the latter step, we fix the *'max_token_length'* equal to 128, based on the maximum token count. After tokenizing the hotel reviews into tokens, the feature representation is applied to each token via multiple transformed layers using AraBERT. These vectors are provided to fine-tune the AraBERT's parameters based on the used labeled data.

The AraBERT output is connected to an additional feed-forward linear layer involving the sigmoid classifier to predict the aspect category. These contextualized representations are fed to a task-specific layer to identify the aspect categories of hotel reviews. As we deal with an MLC for the ACD, the Binary Cross Entropy (BCE) is used as a loss function. This BCE aims to minimize the error for each category label in model training. The output is the probability distribution over all the categories. A threshold is then fixed for this objective to predict the most relevant classes as a final result.

3.5 DWLF for Imbalanced Classes

In the real-world classification tasks, the imbalanced class represents an inherent issue, where the minority class is the class of interest (Fernando and Tsokos, 2021). This data type is characterized by a long tail with a few dominant classes, while most other classes are represented by relatively few examples (Cui et al., 2019). In this context, the authors in (Jafari et al., 2019) proposed a weighted loss function by generating a weight based on the predicted value and error obtained for each instance of the image segmentation. In addition, an existing DWLF, focal loss (FL), is designed for predicting probabilistic outputs in (Lin et al., 2017). Based on both of these research, the DWLF was investigated in (Rengasamy et al., 2020) to overcome the issue of imbalanced data in prognostic and health management. Moreover, the authors in (Alturayef and Luqman, 2021) investigated the DWLF for an imbalanced dataset of tweets.

In our study, we mention that the SemEval-2016 dataset for hotel reviews shows an imbalanced distribution of the different classes, as presented in Figure 7. Consequently, we propose the weighting approach of the Inverse Number of Samples (INS) to deal with the issue of skewed data. We examine the performance of this approach on the ACD task using the Arabic SemEval-2016 for hotel reviews. This DWLF technique is considered cost-sensitive learning that uses weighting by inverse class frequency (Cui et al., 2019). It aims to directly influence the loss function by assigning relatively higher costs to examples from

minor classes. Indeed, weights are computed for the different samples based on the class these samples belong to (majority or minority classes). We essentially want to assign a higher weight to the loss encountered by the samples associated with minor classes. Consequently, the applied weighting on the BCE can be formally expressed as presented in Equation 1, where x_i is the input, y_i is the ground truth label, N is the number of batch size, and w_i is the sample weight that we wish to compute for every sample.

$$L(x, y) = (l_1, \dots, l_N)^T \quad (1)$$

$$l_i = -w_i[y_i \log x_i + (1 - y_i) \log(1 - x_i)]$$

Regarding the INS method, the weights of the samples are presented as the inverse of the class frequency for the class they belong to. This enables us to weigh the contribution of a particular sample toward the overall loss. Based on the number of classes' samples, this method aims to have different weights for each class in the loss function. The implementation computes these weights and normalizes them over different classes. The class weight w_c is calculated as explained in Equation 2. Then, each sample weight w_i is deduced as the average of the weights of the classes that the sample is in. The calculation of the sample weight is underscored in Equation 3; C represents the global number of classes the sample belongs to.

$$w_c = \frac{1}{\text{Number of samples in class } c} \quad (2)$$

$$w_i = \frac{\sum_{j=1}^C w_j}{C} \quad (3)$$

4 EXPERIMENTAL RESULTS

This part highlights the results and discussions of the developed explorations in this study using the Arabic SemEval-2016 for hotel reviews. For the evaluation, we used the F_1 score metric, which is more robust to class imbalance than accuracy.

The parameters of the AraBERT pre-trained model are fine-tuned to transfer its knowledge into this ACD of the hotel reviews. In this context, the parameter of *'max_token_length'* is essential as they help us to specify the maximum length of the classified reviews for the AraBERT model. Indeed, the sentences are padded for those shorter than this maximum length parameter, while it will be truncated from the right side when the sentence is longer. As a smaller *'max_token_length'* helps to have faster training and lower resource, we choose it with a value of

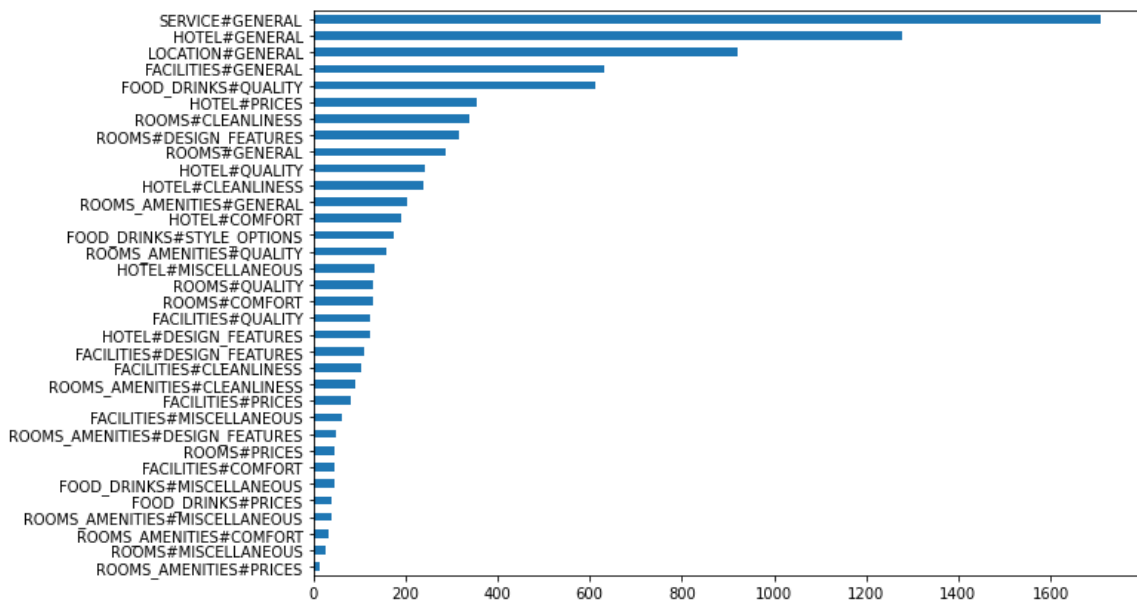


Figure 7: Imbalanced distribution of the Arabic SemEval-2016 dataset for hotel reviews.

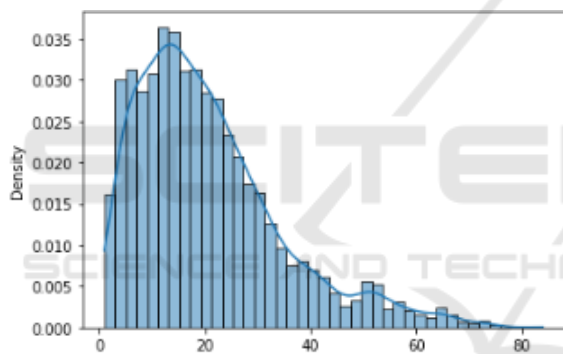


Figure 8: Text distribution in SemEval-2016 dataset.

128, as the smallest power of 2 captures 100% of our reviews. The chosen value of this parameter is justified in Figure 8.

To stress the hyper-parameters used for our model training, we consider AdamW as the optimizer with a batch size of 16, 10 epochs and a learning rate of $1e-3$. In this study, to quickly converge to an optimum, we considered a learning rate scheduler with linear warm-up steps, as specified in Figure 9.

To identify the aspect categories of the hotel reviews, the generated contextualized word embedding was then fed into a simple one-hidden linear layer, as a task-specific layer on top of the AraBERT pre-trained model. In this context, our training approach provides accurate results for the model of the multi-label learning of the hotel categories. As shown in Figure 10, we compared different models, considering the impact of the preprocessing and the proposed loss function weighting to deal with imbal-

anced classes. We achieved promising results for all the investigations. This confirms the usefulness of using the contextual word representation by fine-tuning the pre-trained AraBERT for this MLC task. Furthermore, it helps to detect the context and association between terms to predict the aspect category. In addition, the fine-tuned AraBERT model for this task exceeds the state-of-the-art with an F_1 of 64.3%.

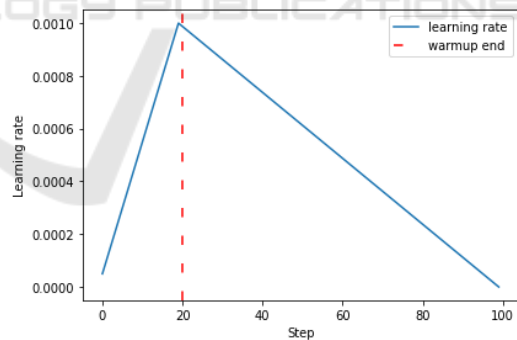


Figure 9: Linear learning rate scheduler.

In this study, we also evaluate the impact of the detailed preprocessing combined with the AraBERT fine-tuning. When applying the proposed preprocessing as dealing with the stop words, emoticons transformation, stemming, etc. (AraBERT+Prep.), the model performance overcomes that just based on AraBERT with an F_1 of 65.2%, as shown in Figure 10. Although this improvement is slight on this DL-based model, this can justify the importance of the data preprocessing investigation to handle the most challenges for Arabic morphology complexity.

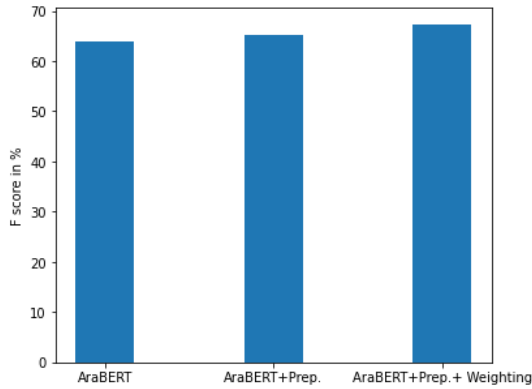


Figure 10: Comparison of AraBERT-based models.

Comment= 'التعرف فسيحة ودورات المياه ممتازة وفخمة ومتكاملة ، لم يعجبني المسبح لأنه اصغر مما يجب على فندق خمس نجوم وكذلك عدم وجود واي فاي مجانا'

Output :

FACILITIES#GENERAL: 0.6068637371063232
ROOMS#DESIGN_FEATURES: 0.7377303838729858

Figure 11: An example of the output prediction.

Regarding the issue of imbalanced labels of the SemEval-2016 dataset, the proposed INS technique for the DWLF was applied to the sample weighting. We mention that the finding of our proposed cost-sensitive solution improves the model performance. Consequently, penalizing the loss function directly based on the proposed weighting approach positively impacts the model. Finally, our proposed approach based on AraBERT combined with the preprocessing and the weighting of the samples achieves an F_1 of 67.3%, as confirmed in Figure 10 (AraBERT+Prep.+ Weighting).

Figure 11 shows an example of the provided output for our proposed approach. Compared to the truth aspect categories of this example, this test shows global acceptable predicted categories ('FACILITIES#GENERAL' and 'ROOMS#DESIGN_FEATURES'). To confirm the added value of our study, we compared our proposed approach to the related works using the same dataset for the Arabic ACD, as shown in Table 3. The final result of our proposed AraBERT-based approach outperforms the previous related works for Arabic ACD with more than 9% in terms of F_1 score.

5 CONCLUSION

This paper proposed an enhanced MLC approach for the ACD using the Arabic SemEval-2016 for hotel reviews. The methodology used was based on the trans-

Table 3: Comparative results of our proposed approach versus the related works for the ACD task.

Models	F_1 %
Baseline (Pontiki et al., 2016a)	40.33
INSIGHT-1 (Ruder et al., 2016)	52.11
UFAL (Tamchyna and Veselovská, 2016)	52.59
IndyLSTM (Al-Dabet et al., 2021)	58.05
Our AraBERT-based approach	67.30

fer learning of the AraBERT-based model. This proposed approach was improved by using specific preprocessing for Arabic text and investigating the INS weighting for the loss function to deal with imbalanced classes. Our study's results were state-of-the-art in the Arabic ACD task using the hotel reviews dataset. In the future, we want to try out other ways to deal with the imbalanced data and use other Arabic BERT-based models.

REFERENCES

- Abdelali, A., Darwish, K., Durrani, N., and Mubarak, H. (2016). Farasa: A fast and furious segmenter for arabic. In *the North American chapter of the association for computational linguistics: Demonstrations*.
- Al-Dabet, S., Tedmori, S., and Mohammad, A.-S. (2021). Enhancing arabic aspect-based sentiment analysis using deep learning models. *Computer Speech Language*.
- Alturayef, N. and Luqman, H. (2021). Fine-grained sentiment analysis of arabic covid-19 tweets using bert-based transformers and dynamically weighted loss function. *Applied Sciences*, 11(22):10694.
- Antoun, W., Baly, F., and Hajj, H. (2020). Arabert: Transformer-based model for arabic language understanding.
- Charte, F., Rivera, A. J., del Jesus, M. J., and Herrera, F. (2015). Mlsmote: Approaching imbalanced multi-label learning through synthetic instance generation. *Knowledge-Based Systems*.
- Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*.
- Cui, Y., Jia, M., Lin, T.-Y., Song, Y., and Belongie, S. (2019). Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9268–9277.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding.
- Drummond, C., Holte, R. C., et al. (2003). C4. 5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling. In *Workshop on learning from imbalanced datasets II*, volume 11.

- Fernando, K. R. M. and Tsokos, C. P. (2021). Dynamically weighted balanced loss: class imbalanced learning and confidence calibration of deep neural networks. *IEEE Transactions on Neural Networks and Learning Systems*.
- Gonnet, P. and Deselaers, T. (2020). Indylstms: Independently recurrent lstms. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.
- Guellil, I., Saâdane, H., Azouaou, F., Gueni, B., and Nouvel, D. (2021). Arabic natural language processing: An overview. *Journal of King Saud University-Computer and Information Sciences*.
- Hamdi, S., Hamdi, A., and Ben Yahia, S. (2022). Bert and word embedding for interest mining of instagram users. In *Advances in Computational Collective Intelligence*, pages 123–136, Cham. Springer International Publishing.
- Huang, C., Li, Y., Loy, C. C., and Tang, X. (2016). Learning deep representation for imbalanced classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.
- Jafari, M., Li, R., Xing, Y., Auer, D., Francis, S., Garibaldi, J., and Chen, X. (2019). Fu-net: multi-class image segmentation using feedback weighted u-net. In *International Conference on Image and Graphics*, pages 529–537. Springer.
- Kaur, H., Pannu, H. S., and Malhi, A. K. (2019). A systematic review on imbalanced data challenges in machine learning: Applications and solutions. *ACM Computing Surveys (CSUR)*, 52(4):1–36.
- Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P. (2017). Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.
- Matthew, E. P., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Kenton, L., and Zettlemoyer, L. (2018). Deep contextualized word representations.
- Oudah, M., Almahairi, A., and Habash, N. (2019). The impact of preprocessing on arabic-english statistical and neural machine translation.
- Pontiki, M., Galanis, D., Papageorgiou, H., Androutsopoulos, I., Manandhar, S., Al-Smadi, M., Al-Ayyoub, M., Zhao, Y., Qin, B., De Clercq, O., et al. (2016a). Semeval-2016 task 5: Aspect based sentiment analysis. In *International workshop on semantic evaluation*, pages 19–30.
- Pontiki, M., Galanis, D., Papageorgiou, H., Manandhar, S., and Androutsopoulos, I. (2016b). Semeval 2016 task 5: aspect based sentiment analysis (absa-16) annotation guidelines.
- Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. (2018). Improving language understanding by generative pre-training.
- Rengasamy, D., Jafari, M., Rothwell, B., Chen, X., and Figueredo, G. P. (2020). Deep learning with dynamic weighted loss function for sensor-based prognostics and health management. *Sensors*.
- Ruder, S., Ghaffari, P., and Breslin, J. G. (2016). Insight-1 at semeval-2016 task 5: Deep learning for multilingual aspect-based sentiment analysis.
- Tamchyna, A. and Veselovská, K. (2016). Ufal at semeval-2016 task 5: recurrent neural networks for sentence classification. In *Proc. of the 10th international workshop on semantic evaluation*.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, ., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*.