المؤتمر الوطني الخامس لكليات الحاسب الآلي بالجامعات السعودية
The Fifth National Conference of Saudi Computers Colleges 2022
21-22 December 2022

Enhance the Aspect Category Detection in Arabic Language using AraBERT and Text Augmentation

Miada Ahmeddeb Almasre
Information Technology departmatnt
King Abdulaziz university(KAU)
malmasre@kau.edu.sa

Introduction:

Data Augmentation (DA) has recently been incorporated in NLP processes due to interest in working with low resource domains, novel tasks, and the implementation of neural networks that demand large sets of training data. Despite this current focus on DA methods, this field is still somewhat under-researched, probably hindered by the discrete nature of text data.

Moreover, augmenting text data processes that are performed on languages like Arabic might pose challenges of a different type which relates to the language's inherent morphological and syntactical complexities.
ACD studies on the Arabic language indicated a preoccupation with Machine Learning and DL models but a lesser interest in data augmentation.

H name	rating	user type	room type	nights	review	sent_tokenized	no_s
فندق 72	4	زوج	غرفة قياسية مزدوجة	أقمت ليلة واحدة	جيد. الخدمه كانت ممتازه والافطار ممتاز اشكر طاقم الفندق كان muhabatوخصوصا السيد انسان لطيف وخدوم. لا شي	[جيد.,, الخدمه كانت ممتازه والافطار ممتاز اشكر طاقم الفندق كان muhabatوخصوصا السيد انسان لطيف وخدوم.,, لا شي]	3
فندق 72	5	-	غرفة ديلوكس مزدوجة أو توأم	أقمت 3 ليالي	استثنائي. الغرف والنظافه والاستقبال. وقت المساج ووقت السباحه	[استثنائي.,, الغرف والنظافه والاستقبال.,, وقت المساج ووقت السباحه]	3
فندق 72	4	زوج	غرفة ديلوكس مزدوجة أو توأم	أقمت ليلتين	جيد. المكان جميل وهاديء. كل شي جيد ونظيف بس كان حوض السباحه لايعمل في هذي الفتره حسب كلامهم يقولوا فيه صيانه والله اعلم	[جيد.,, المكان جميل وهاديء.,, كل شي جيد ونظيف بس كان حوض السباحه لايعمل في هذي الفتره حسب كلامهم يقولوا فيه صيانه والله اعلم]	3

Original:
مبكره رحله طيران قبل للاقامه مثالي انصح بالنوم ليس تناول الطعام موقع
Augmented Text:
ص9 رحله طيران بيومين للاقامات رائع انصح بالنوم ليس تناول الطعام موقع

Fig. SemEval2016- dataset sample + Augmentation example

Results:

The baseline and augmentation pipelines achieved close results. With the baseline, for example, an f1-score of 0.663 has been recorded, and an f1-score of 0.661 for the pipeline with augmentation. This generally suggests that augmenting the dataset using the Word2Vec model does not lead to better model performance or accurate ACD.

Though the augmentation-based model did not outperform the baseline, the f1-scores of both pipelines show an improvement when compared to results cited in previous research that investigated ACD using DL methods on the Arabic SemEval2016 review datasets. The highest results reported by [12], [11], and [10] are f1-scores of 65%, 53%, and 47.3% respectively.

In addition, investigating the aspects (by-class) f1-scores leads to a better understanding of the previously reported results, Only seven aspect categories have f1-scores that demonstrate an improvement when applying augmentation

Research problem:

The availability of robust text data (dataset) becomes crucial to the feasibility and the accuracy of DL applications, however, one of the issues impeding the realization of this target is the sacristy of big, labeled text data.

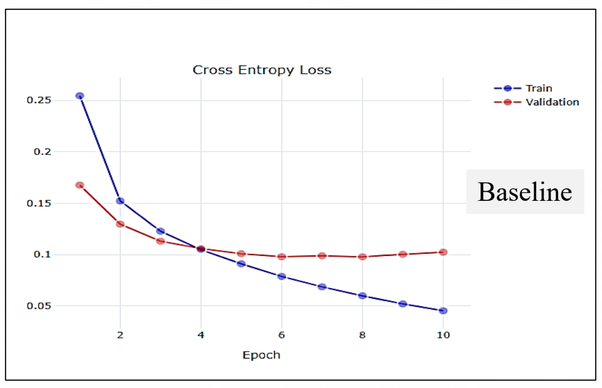
Solution

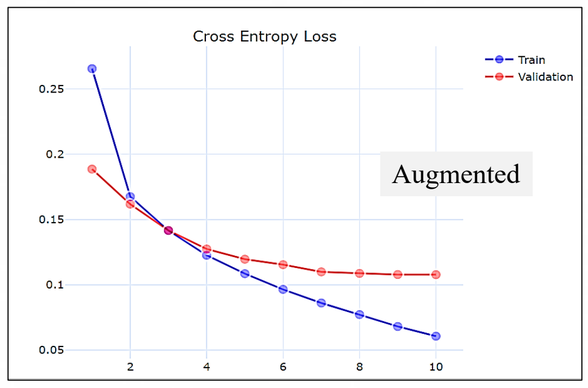
The DL research community constantly explores solutions to labeling big data without human effort, which could be facilitated through applying data augmentation methods and techniques.

Research Methodology:

ACD task focused on opinionated text like reviews, a model detects the aspect category in each sentence based on predefined categories.
In this experiment, a BERT-based model is implemented to predict aspects in an augmented dataset of hotel reviews, where each sentence is supposed to be under one category.
To achieve this objective, a three key phases are proposed:

1. involves the preparation of the datasets.
2. augmentation, which is followed by an implementation of the AraBERT model to predict aspects.
3. an evaluation of the classification results is conducted.





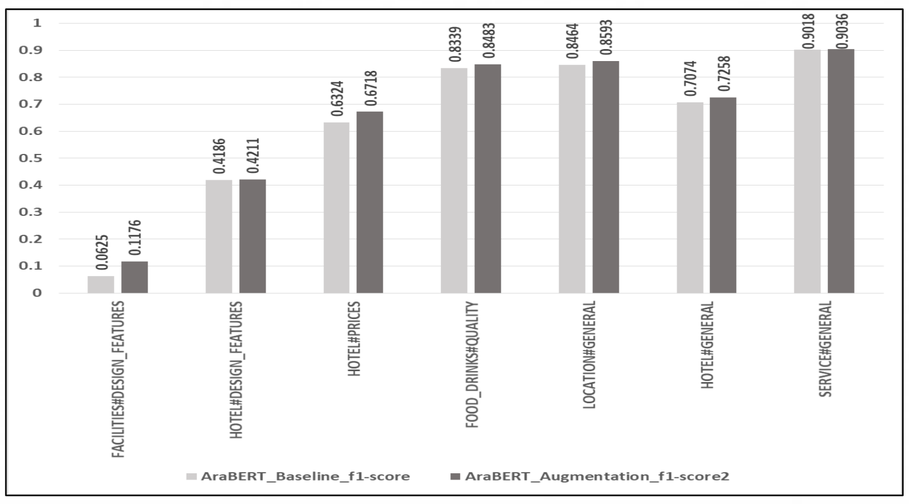


Fig. cross-entropy loss for 10 epochs on both pipelines + aspect categories have f1-scores improvement

Conclusion:

Even though this research does report an improvement in the performance of the augmentation-based model, it presents augmentation as a feasible and relatively successful technique that can be implemented in an ACD task.