At first we thought to combine 4 datasets as they contain the same columns (context, question, answer, context id) and these datasets are Arabic SQUAD, ARCD, mlqa, tydiqa.

We tried some preprocessing techniques first on the Arabic squad data by luck, and this data showed unpredictable context data, some context text contains strange alone characters which can’t be understood like ا after some words and some other characters, after some time we discovered that this problem was caused by removing the diacritics from the data which is good thing but the problem was the replaceable character which is ‘ ‘ instead of ‘‘ there was white space in the place of every diacritization in the whole dataset context column.

We thought we figured out the problem and we could solve it by removing every ا and combine it the previous word but there wasn’t only the ا , there were many other characters that has no meaning even if u try to combine it the next or previous word, there were some strange characters in the contexts, but we figures out that the diacritization was in ا in the end of word and ى in the beginning of the word, so we may handle those characters, but what about the other characters? And also is there any context that may containا as it is without diacritics? What should we do about these problems?

We agreed on taking the risk in the ا and ى ways to remove the additional spaces before or after despite it may make some wrong words but the majority was of the mis replaceable space, and for the other additional character that has no meaning, we took a quick look at the data at some Arabic characters and found out, there was small occurrence when the separated character really has meaning and been separated due to diacritization, so we decided to remove all characters EXCEPT for the conjunction (حروف العطف) and preposition (حروف الجر) as they do have a meaning when they are separated like (ك،و،ب).

Why didn’t we choose another dataset instead of this one?

We looked into that solution but the data we wanted to combine as mentioned above were small, and they together less than 2000 rows while Arabid Squad’s rows more than 48 thousand, and this is a big difference between the datasets so we need it for its bid data, but we thought about something …

We will take two approaches first approach is working on Arabic squad data with its problem and second approach is to work on the other dataset combined together, it is small but it may work good and give better accuracy, and also first approach may take a lot time and in the end it may give lower accuracy and it will be wasted time, now we will work on these approaches and see what will happen …….

After some searching we made a function that search for the ا and if there space before and after, it joins the ا with the previous word and for the أ and ي if there spaces before and after them, it joins them with the next word and it worked successfully, now the data is good and ready for the preprocessing.

For the preprocessing technique used in the squad data are the function the fix the problem, checking on diacritization, remove extra spaces, remove different variation of أ and replace it with ا and removing punctuation.

For the another preprocessing techniques, we will see how to use it but after trying to get an output form the model

For the tokenizer used to tokenize we tried t5 tokenizer, but It wasn’t good as it was tokenizing spaces and some words into two tokenization, so we tried another tokenizer which is arabert tokenizer and it was better than t5.

Now we are trying to use t5 model small, but we found that the vocab size for the model is smaller than arabert vocab size, vocab size for model is around 24 000 and arabert vocab size is around 64 000, the model limit is so small for the tokenizer, we tried to use t5 tokenizer as it suitable for the model, but its vocab size is much bigger than arabert, around 250 100 which is not good with model also.

We now are trying to try 3 approaches which are try t5 medium, try arabert model, try to resize the model t5 small to the tokenizer vocab size and we will see what the result is will be ………

We tried to run the model t5 with arabert tokenizer with resizing the model to the vocab size of the tokenizer, but this was a great mistake the output of the model was garbage because the model runs with different sequence from the tokenizer, so I tried to try t5 model with its tokenizer, but there isn’t good also as the vocab size was low for the words and the output were null, there was no output.

The way now is to try mt5 model with its tokenizer but the vocab size it too large with 25100, so we are trying to use it and w will see…..