Arabic Dialect Classification Research

Abdulrahman Alshawi, Mohand Alrasheed, Khalid Albader, Abdullah Alsuwailem, Musaad Alqubayl

Supervised by: Dr. Nasser Alsadhan

Abstract:

The Arabic language is one of the oldest languages widely used today, and as a result of that, many Arabic speaking regions have formed dialects exclusive to their own, for example, many countries surrounding the Arabic Gulf have formed a dialect different to countries in the Levantine region. We intend on identifying and systematically determining the dialect of a piece of text.

Introduction:

A dialect is the variation of a language in grammar, pronunciation and vocabulary. Every individual has their own way of talking that is affected by dialect, accent, background and many other factors[1].

The Arabic language has a variety of dialects throughout the Arabic world, dialects could differ not only across countries but also in the same country or even city. Arabic dialects differ from one another in pronunciation and vocabulary, different dialects have different words or different variations of a word that could refer to the same meaning. This research has many applications in Arabic text analysis, such as helping in identifying the regions customers most often come from by analyzing a product's reviews and comments and breaking them down by region, which provides useful intel for a business. One of the major challenges in dialect recognition is dividing data into classes of dialects.

Problem statement:

Dialects are formed mainly due to regional separation between the Arab world. This separation reduces interaction between different regions, and as a result of that, many Arabic speaking regions have formed dialects exclusive to their own. For example, many countries surrounding the Arabic Gulf have formed a dialect different to countries in the Levantine region. The research's main problem is how to identify and predict dialect types from text.

Goals & Objectives:

The goal of this research is to analyze and understand Arabic text to classify the dialect of any piece of Arabic text.

The objective is to implement the most appropriate state of the art NLP model that helps in achieving the best possible accuracy which correlates to correctly classifying what dialect the text is from.

Dataset:

In this research we'll be using the Social Media Arabic Dialect Corpus (*SMADC*) dataset. SMADC's corpus is collected from three different sources, Facebook, Twitter and online newspapers.

The researchers filtered Facebook and Twitter documents automatically by removing hashtags, emojies, redundant characters and so on.

They also started filtering the noises in their dataset, such as writing a nationality that conflicts with the label, non-Arabic characters, etc. [2]

In their final records, SMADC dataset contained 1,088,578 documents. which consisted of 812,849 Facebook comments, 9,440 online newspaper comments, and 266,289 tweets[2]. And each one of them is distributed by the five labels (GLF, EGY, NOR, LEV and IRQ).

Preprocess:

We used preprocessing techniques that help in transforming the data to a representation the model understands, like *tokenization* and *segmentation*.

Before tokenizing our dataset, we will normalize Arabic diacritics such as Fatha, Damma, Kasra and so on. this should help the model group similar words, albeit lose a bit of accuracy.

Arabic word segmentation works by separating the suffixes and prefixes attached to any given word, segmentation has shown to have significant impact in many NLP application such as context understanding, because it gives more information to the model.

Tokenization is an essential task for NLP problems. We'll be using tokenization to transform our dataset to be ready for model input, after applying normalization and segmentation on the data we transform each token to a number, so if one token is repeated more than once then that token is transformed to the same number. This allows the model to understand the input.

Algorithm:

For the purposes of this research, we'll use an existing implementation of BERT to solve our problem. We'll be using an Arabic BERT model called *AraBERT* and fine-tuning it to the problem we need to solve.[3]

In order to use AraBERT, we need to provide a sequence of tokens represented as numbers, after tokenizing our data we'll replace each token with a number representation, this forms a suitable input for AraBERT.

References:

1. Fadi Biadsy. Automatic Dialect and Accent Recognition and Its Application to Speech Recognition. PhD thesis, Columbia University, USA, 2011.
2. Areej Odah O. Alshutayri. Arabic Dialect Texts Classification. PhD thesis, The University Of Leeds, 2018.
3. Wissam Antoun, Fady Baly, and Hazem Hajj. AraBERT: Transformer-based Model for Arabic Language Understanding. American University of Beirut.