SentDialects: A Sentiment Analysis Dataset in Multiple Dialects

Ahmed Ali[1], Peter Bell[2]
[1]QCRI, [2]HBKU

February 8, 2025

Abstract

SentDialects is a novel Arabic sentiment analysis dataset comprising 10,000 sentences across multiple dialects: Syrian (4,000), Iraqi (4,000), and Lebanese (2,000). The dataset was curated from the AAA dataset and supplemented with data collected from grammar websites. SentDialects aims to enhance sentiment analysis research in Arabic dialects, addressing the linguistic diversity and challenges inherent in processing dialectal Arabic. The dataset is publicly available under the CC BY-NC 4.0 license at GitHub.

1 Introduction

SentDialects, an Arabic sentiment analysis dataset, comprises 10,000 sentences in multiple dialects. See the following table for a description of the dialects

Dialect	**Number of Sentences**
Syrian	4,000
Iraqi	4,000
Lebanese	2,000
Total	**10,000**

Table 1: Description of the Dataset by Dialect

2 Data Collection

ABCD collected this dataset using the AAA dataset and crawling grammar websites. It's available using this link [1] under the license CC BY-NC 4.0. Also accessible on HuggingFace via the link [2] under the same license.

[1] https://github.com/ahmedali/SentDialects

[2] https://hf.co/datasets/ABCD/SentDialects