Enhancing Arabic NLP:
A Comparative Study of AI-Driven Text Preprocessing Tools

Suha Khalil Assayed
http://orcid.org/0000-0002-9924-0324
The British University in Dubai, UAE

Safwan Maghaydah
http://orcid.org/0000-0003-2477-1187
Abu Dhabi University, UAE

Khaled Shaalan
http://orcid.org/0000-0003-0823-8390
The British University in Dubai, UAE

ABSTRACT

Arabic is a first language for more than 300 million people. It has some unique features that can make it one of the most complex languages, such as multiple derivatives, unlimited vocabulary, diacritics, and others. Preprocessing Arabic text is an essential step in order to prepare text for Natural Language Processing (NLP) purposes. This article provides a comparison study of several preprocessing tools for Arabic text. It explains the challenges in pre-processing the Arabic language as well as the techniques that used in every particular tool. However, the authors used the PRISMA for reporting the systematic reviews, which they started with screening 200 articles and ended-up with including only 30 articles. After reviewing these articles deeply, the results show that different tools such as AMIRA, CAMel ,and NLP packages added value in text-preprocessing. However, most of this papers considered that the ambiguity in Arabic orthography as well as the dialectal variants are the most challenges in Arabic NLP.

INTRODUCTION

As artificial intelligence and social media platforms grow rapidly, enormous amounts of data and textual content flow across the Internet every second. The large amount of digital data requires the development of efficient tools to process and organize it effectively for the best possible use. In particular, artificial intelligence (AI) techniques such as machine learning (ML) and natural language processing (NLP) are increasingly needed in this domain. These technologies enable automatic analysis, categoriza-

DOI: 10.4018/407607

tion, and extraction of meaningful patterns from vast datasets, facilitating tasks like sentiment analysis, trend detection, and content summarization.

Interestingly, that within the Arabic-speaking community, Twitter is recognized as a powerful and influential social media platform (Alruily, 2021). It serves as a hub for discussions, news dissemination, and social interaction, making it an important source of data. However, processing Arabic text from platforms like Twitter presents specific challenges. The Arabic language is morphologically rich, has complex syntax, and contains dialectal variations, all of which complicate the effective extraction of meaningful insights from text. Moreover, the absence of diacritics in most online content further adds to the difficulty.

To address these issues, AI-driven NLP tools designed specifically for Arabic have gained importance. These tools utilize algorithms to identify patterns in text, analyze sentiment, classify content, and extract relevant information, improving our ability to derive insights from Arabic-language social media.

It has been used for different purposes like business, entertainment, information sharing, news, and others (Hegazi et al., 2020). Arabic is a first language for more than 300 million people and the second language of about 250 million people (Husain & Uzuner 2021).

We chose the Arabic language, because it is highly derivative where several words could be formed using only the same root. Also, a single word may be derived from multiple roots (Attia, 2000, Kaddoura & Ahmed, 2022).). The Arabic language has some features like multiple derivatives, unlimited vocabulary, diacritics, and others. The Arabic language consists of three forms (Classical Arabic, Modern Standard Arabic and Dialect Arabic language) (Jarrar, 2021). So, processing of Arabic language text is one of the challenges in Natural language processing, and in the last few years the research works that are done in processing Arabic texts in the social media field have gotten highly important (Hegazi et al., 2020).

Objectives

Recently, several tools and techniques developed by several scientists in order to reduce the challenges and the ambiguities in preprocessing the Arabic Languages into NLP applications. In this research, we will evaluate and understand these techniques by exploring several research papers from different journals and accordingly we can highlight the most essential processes that can add value to the resources that support Arabic NLP.

Research Questions

1. What is the most important Arabic' preprocessing functions that are developed to solve the challenges?
2. What are the features in Arabic Language that produced the challenges for preprocessing the Arabic texts?
3. What are the most effective NLP tools for Preprocessing the Arabic texts?

APPROACH

We applied Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) as it shown in figure 2, it demonstrates the phases of a systematic review by identifying the number of re-