Rewayatech: Saudi Web Novels Dataset

Aseel Addawood
Information Science Department
Imam Mohammad Bin Saud University
Riyadh, Saudi Arabia
aasdawood@imamu.edu.sa

Daliyah Alzeer
English Language Institute
Taif University
Riyadh, Saudi Arabia
dalia.h@tu.edu.sa

Abstract

The internet has changed the way people perceived fiction to a new level. For instance, online forums have given people the opportunity to write without revealing their real identities. Especially in the Saudi context, online users were using these forums to write web novels that reflect their culture, lives, concerns, hopes and dreams. In this paper, we describe a dataset that was collected from one of the online forums that was used for sharing web novels among its readers. The collected dataset contains 1,267 novels between 2003-2015. This data set is available to the research community to analyze to gain a better understanding of the social, economic, and behavioral mindset that was manifested in the community in that decade.

1 Introduction

“The stories that we tell about our own and others’ lives are a pervasive form of text through which we construct, interpret and share experience” (Schiffrin, 1996). In telling stories, we share experiences as it is part of the essence of human interaction and a way of expressing their perspectives of life (Ochs and Capps, 2009). Needless to say, people’s beliefs and emotions can manifest itself in the novels they write, through words we configure their experience and identity (Bamberg, 2004). The importance of novels lies in the social and environmental changes that can be traced through them which Sultan Alqahtani emphasises (1994).

In this paper, we focus on Saudi web novels that are written on online forums, the history of Saudi novel has begun and flourished since the first published novel "The Twins" in 1930 (Al-Qahtani, 1994). However, as the internet has became a part of every Saudi house, web-based Saudi novels has changed how people think of novels at that time. The difference between web-based novels and printed novels can be concluded in few points. First, the scarcity of references in this type of data has created an interesting gap for researchers to explore especially in light of the short history of Saudi novels. In addition, the dataset is peerless as the writers were publishing their work while they interacted simultaneously with the readers without necessarily the risk of revealing their true identities. Moreover, the writers and the readers are mostly young; "Granted, increasing numbers of young writers are turning to the Web for publication" (2006, Bishop Starkey). It is important to understand that some Saudi web-based novels are similar to printed novel in their style and length. Few of these web novels have turned into printed novels such as "You are mine" by Muna Almarshood (2013). Accordingly, we can see some resemblance between both the Saudi web novels and printed novels. The availability of such corpus can help particularly in the field of digital humanities.

Previous studies have used similar datasets for genre classification (Samothrakis and Fasli, 2015; Henny-Krahmer, 2018; Yu, 2008; Kim et al., 2017), story clustering (Reagan et al., 2016), mapping emotions to geographical locations in literature (Heuser et al., 2016), and construction of social networks of characters (Nalisnick and Baird, 2013; Jhavar and Mirza, 2018). Other studies use emotion analysis as a starting point for stylometry (Koolen, 2018), inferring psychological characters’ traits (Egloff et al., 2018), and analysis of the causes of emotions in literature (Kim and Klinger, 2018, 2019).

For the best of our knowledge there is no availability of Arabic literature dataset. Arabic datasets are limited to newspapers (Einea et al., 2019; Ababneh et al., 2014), book reviews (Aly and Atiya, 2013), and poems (Ahmed et al., 2018), which have different characteristics, readers and writers from fiction in literature.

The dataset shared contains 1,267 stories written in Arabic using Saudi dialects. In the following sec-
tions we describe data collection, dataset statistics, and information about how to access the data set.

2 Corpus construction

The data collection started by browsing online forums for stories written in Arabic. One forum was used for the data collection\(^1\). The forum contained word files for each story written by an online user with a total of 1,267 files, novels. In average, each novel contains 73,798 words. These novels where written by 913 unique authors, with a minimum of one story per writer and a maximum of seven stories per author. Authors who authored more then one novel usually have novels with multiple series.

3 Dataset Access

The data set is accessible through GitHub at this address: https://github.com/aseelad/Rewayatech-Saudi-Stories/

Acknowledgements

The author gratefully thanks Razan Alhjjy for their help in data collection and paper organization.

References


\(^1\)https://forums.graaam.com