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, economical, 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, peoples' 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

Jafar Ababneh, Omar Almomani, Wael Hadi, Nidhal Kamel Taha El-Omari, and Ali Al-Ibrahim. 2014. Vector space models to classify arabic text. *International Journal of Computer Trends and Technology (IJCTT)*, 7(4):219–223.

Alfalahi Ahmed, Ramdani Mohamed, and Bellafkih Mostafa. 2018. Use smo svm, lda for poet identification in arabic poetry. In *International Conference on Advanced Information Technology, Services and Systems*, pages 161–169. Springer.

Sultan SM Al-Qahtani. 1994. *The novel in Saudi Arabia: emergence and development 1930-1989; an historical and critical study.* Ph.D. thesis, University of Glasgow.

Mohamed Aly and Amir Atiya. 2013. Labr: A large scale arabic book reviews dataset. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 494–498.

Michael Bamberg. 2004. Form and functions of 'slut bashing'in male identity constructions in15-year-olds. *Human development*, 47(6):331–353.

Mattia Egloff, Antonio Lieto, and Davide Picca. 2018. An ontological model for inferring psychological profiles and narrative roles of characters. In *Digital Humanities 2018: Conference Abstracts, Mexico City, Mexico.*

Omar Einea, Ashraf Elnagar, and Ridhwan Al Debsi. 2019. Sanad: Single-label arabic news articles dataset for automatic text categorization. *Data in brief*, 25:104076.

Ulrike Henny-Krahmer. 2018. Exploration of sentiments and genre in spanish american novels. In *DH*, pages 399–402.

Ryan Heuser, Franco Moretti, and Erik B Steiner. 2016. *The emotions of london.* Literary Lab.

Harshita Jhavar and Paramita Mirza. 2018. Emofiel: Mapping emotions of relationships in a story. In *Companion Proceedings of the The Web Conference 2018*, pages 243–246.

Evgeny Kim and Roman Klinger. 2018. Who feels what and why? annotation of a literature corpus with semantic roles of emotions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1345–1359.

Evgeny Kim and Roman Klinger. 2019. Frowning frodo, wincing leia, and a seriously great friendship: Learning to classify emotional relationships of fictional characters. *arXiv preprint arXiv:1903.12453.*

Evgeny Kim, Sebastian Padó, and Roman Klinger. 2017. Investigating the relationship between literary genres and emotional plot development. In *Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 17–26.

Corina Koolen. 2018. Women's books versus books by women.

Eric T Nalisnick and Henry S Baird. 2013. Extracting sentiment networks from shakespeare's plays. In *2013 12th International Conference on Document Analysis and Recognition*, pages 758–762. IEEE.

Elinor Ochs and Lisa Capps. 2009. *Living narrative: Creating lives in everyday storytelling.* Harvard University Press.

Andrew J Reagan, Lewis Mitchell, Dilan Kiley, Christopher M Danforth, and Peter Sheridan Dodds. 2016. The emotional arcs of stories are dominated by six basic shapes. *EPJ Data Science*, 5(1):1–12.

Spyridon Samothrakis and Maria Fasli. 2015. Emotional sentence annotation helps predict fiction genre. *PloS one*, 10(11):e0141922.

Deborah Schiffrin. 1996. Narrative as self-portrait: Sociolinguistic constructions of identity. *Language in society*, pages 167–203.

Bei Yu. 2008. An evaluation of text classification methods for literary study. *Literary and Linguistic Computing*, 23(3):327–343.

[1] https://forums.graaam.com