Language Related Research
E-ISSN: 2383-0816
https://lrr.modares.ac.ir
https://doi.org/10.52547/LRR.12.6.16

Tarbiat Modares University

Vol. 12, No. 6, Tome 66
pp. 499-531
January & February 2022

An Overview of Text Mining in Language Studies: The Computational Approach to Text Analytics

Hadi Masjedy[1], *Seyyed Mohammad Reza Adel*[2*], *Seyyed Mohammad Reza Amirian*[3], & *Gholamreza Zareian*[4]

Received: 6 October 2019
Received in revised form: 4 January 2020
Accepted: 25 February 2020

Abstract

Text mining' refers to the computational process of unstructured text analytics for extracting latent linguistic layers and themes. It is especially significant as content or thematic analysis in descriptive and interpretive studies. This process begins with structuring simple texts and proceeds with summarizing, classifiing, modelling, evaluating and interpreting the inherent textual concepts and patterns. Given that this method counts as an interdisciplinary innovation especially in discoursal studies, it is to be pursued more intensively in academic studies. Despite the multitude of English studies in this area, there has been little interest to date in text mining amongst Iranian researchers as evidenced by the critically limited number of local Persian and English studies. Thus looking into the theory and practice of text mining and its major analytic tools and methods in Persian and English, this paper aims to prepare the ground for utilizing this methodology in language studies.

***Keywords*:** Text mining, unstructured texts, content analysis, thematic analysis, natural language processing

1. PhD in TEFL, Hakim Sabzevari University, Sabzevar, Iran; https://orcid.org/0000-0003-3610-6651
2. Corresponding author, Associate Professor of TEFL, Hakim Sabzevari University, Sabzevar,Iran; *Email: sm.adel@hsu.ac.ir*, https://orcid.org/0000-0002-1136-8973
3. Associate Professor of TEFL, Hakim Sabzevari University, Sabzevar, Iran; https://orcid.org/0000-0003-3719-3902
4. Associate Professor of TEFL, Hakim Sabzevari University, Sabzevar, Iran; https://orcid.org/0000-0002-9084-608X

The last two decades faced a major increase in the rate and accuracy of knowledge generation in language studies due to advances in interdisciplinary studies of applied linguistics and computer sciences. At the heart of methodological innovations especially in discourse studies lies 'text mining' whose merits have only recently been appreciated by researchers. 'Text mining', 'text data mining' or 'Text Analysis' is the use of different data mining algorithms and methods like natural language processing and linguistic as well as statistical techniques to derive linguistic features, significant patterns and valuable themes from the unstructured texts through collecting unstructured data, pre-processing and cleansing them to detect and remove anomalies and processing and controlling operations (Zhou et al, 2012). These processes are further broken down into feature extraction, structural analysis, text summary, text classification, text clustering, and association analysis. Text mining is actually a complicated procedure of extracting valuable, significant patterns and trends from a large number of textual data used for such functions as product suggestion analysis, social media opinion mining, and sentiment or trend analysis (He, 2013).

Dating back to Feldman and Dagan (1995), text mining is an innovative methodology with a relatively short history which is often integrated with corpus analysis to computationally analyze a large body of unstructured texts as potential informative sources of insight. As a subfield of data mining in computer sciences and an interdisciplinary method, text mining borrows from corpus and computational linguistics, whose main purpose is to extract the meta-characters representing textual features (Pons-Porrata et al, 2007). Zhou et al (2017) believe that despite its short history, text mining has been remarkably evolved into the mainstream research methodology in many interdisciplinary areas in the wake of increasingly rapid developments in data mining.

Hashimi et al (2015) explained the steps involved in text mining as a semi-automated process of collecting, structuring and then analyzing textual data as follows: (a) collecting unstructured data from a variety of sources like

[DOR: 20.1001.1.23223081.1400.12.6.27.3]

[DOI: 10.52547/LRR.12.6.16]

textual documents, social media, web pages, mails, blogs, etc. using specialized corpora for organization, (b) pre-processing and cleansing the data for removing the anomalies to unveil latent valuable information using text mining tools, (c) unstructured data conversion into relevant structured formats, (d) discovering the underlying data patterns using word structures, sequences and frequency, and (e) extracting useful knowledge and storing them in a secure database for evaluation, later retrieval, trend analysis and possible decision-making. Text mining aslso makes use of lexicometrics dealing with frequency and co-occurrence analysis of vocabulary to derive structures from texts; sentiment analysis is an application of lexicometrics looking for positive or negative emotions in documents and has been used in social media analysis for evaluating public opinion (Shangzhen & Lemen, 2016).

Text mining is an *area of inquiry* that in itself deserves to be pursued more intensively in future studies and this paper, thus, is an attempt to review its basic principles, procedures and top analytic tools and to raise researchers' awareness of the virtues of text mining.

[DOR: 20.1001.1.23223081.1400.12.6.27.3]

[DOI: 10.52547/LRR.12.6.16]

[DOR: 20.1001.1.23223081.1400.12.6.27.3]

[DOI: 10.52547/LRR.12.6.16]