Sentiment Analysis and Visualization of Amazon Books' Reviews

Aljoharah Almjawel
Department of Computer Science
CCIS, Shaqra University
Shaqra, Saudi Arabia
aalmjawel@ksu.edu.sa

Sahar Bayoumi
Department of Information technology.
CCIS, King Saud University
Riyadh, Saudi Arabia
Institute of Graduate Studies & Research
Alexandria University, EGYPT
sahali@ksu.edu.sa

Dalal Alshehri
Department of Information Technology
CCIS, King Saud University
Riyadh, Saudi Arabia
dalal.h.it@gmail.com

Soroor Alzahrani
Department of Information Technology
CCIS, King Saud University
Riyadh, Saudi Arabia
soosooo79@windowslive.com

Munirah Alotaibi
Department of Information Technology
CCIS, King Saud University
Riyadh, Saudi Arabia
Munirah_alotaibi@yahoo.com

***Abstract*— Nowadays user-generated content is growing in a digital world, where many individuals can post texts about various products in order to express their feedback. Reviews about products are critical as they guide others to use it or not which affect the production profit, and popularity. Amazon is one of the companies that provide reviews of users regarding all its products. Since reviews are textual and varying; an interactive representation would be supportive to users through summarizing opinions and providing feedback. In this paper, a visual analytics tool proposed; as a dashboard; a user-friendly analysis of books' reviews on Amazon. Interactive Packed bubbles, Linear chart, Stacked bars, and Word-cloud are four visualization techniques incorporated together enabling users to explore relationships between multiple objectives at the same time in an interactive manner. Besides, visual sentiment analysis of opinions extracted from the user reviews of books whether positive, neutral or negative is provided to support users' decision.**

Keywords*— *Text Visualization, Tableau, Rstudio, Amazon Reviews, Opinion Analysis, Sentiment Analysis.

I. INTRODUCTION

The internet is considered one of the leading sources of a customer's opinion allowing releasing a variety of websites. These websites enable customers to share their reviews and opinions about different products such as movies, restaurants, hotels, devices, and books. Amazon is an example of increasing availability and increasing popularity for opinion-rich resources where it provides millions of users' reviews of different products categories. Users with different backgrounds wrote the text of reviews. Useful visual analysis of online customer feedback is needed, which significantly affect the analysis of reviews, has gained a high interest in the area of sentiment analysis or determining the appropriate decision-making [1]. Information visualization generally aims to provide different insights and helps in understanding the structure of the data [2]. Sentiment visualization is a subfield of information visualization that deals with the presentation and visual analysis of opinions extracted from the user's opinions [3]. The objective of opinion visualization systems is to provide a simple way to provide users with meaningful summaries of reviews.

This paper aims to provide a practical visual analysis of customer feedback sentiment analysis and visualization techniques integrated. The Amazon book reviews in this paper used to emphasize the usefulness of the proposed interactive dashboard in supporting decision makers. In order to achieve the aim, a sentiment analysis system implemented of textual reviews using a large dataset of the reviews. Furthermore, visualization techniques used to provide users with a summary in an interactive representation using various representations. The proposed system tackles the following questions: (i) what the book that gets the highest reviews is? (ii) Can I find out whether earlier book reviews tend to receive better feedbacks than later ones? (iii) Can I compare several books according to their sentiment reviews?

The following sections organized as follow: section II show previous works about sentimental analysis and visualization of reviews, section III describes the proposed visualization system; section IV shows the results of using the proposed system with real data from Amazon. Finally, section V summarizes the proposed system and results and future work.

II. LITERATURE REVIEW

Our review to research papers concerned with visualizes representation of reviews classified into two main sections. Firstly, researches about systematic reviews regardless of the subject area. Secondly, researches about users' feedback on different products.

A. Systematic Literature Reviews

Godwin in 2016, visualized a systematic literature review in order to identify which areas have sizeable published work as well as which gaps not covered by the current pieces of literature. The dataset gathered from various fields which are Engineering Village, Scopus, ERIC, Education Full Text, and Web of Science databases. After that, a visualization technique developed from the linguistic analysis which called co-occurrence network. This technique presented the network with each cluster as a different color and node shape. Finally, the co-occurrence network helped researchers to highlight significant

978-1-7281-0108-8/19/$31.00 ©2019 IEEE

gaps, and therefore presented new and quick ways to promote research in under-researched areas [4].

Other systematic reviews provided by Salameh and Aljammal on software evolution visualization (SEV) tools. The authors aim to explore the main target of SEV, analyze the classifications and taxonomies that are used to represent SEV tools, and find out what are the primary sources of information used to visualize software's evolution. Different types of visualization techniques used on (SEV) which are graph-based, notation-based, matrix-based, metaphor-based and others were used. Graph-based were the most popular whereas notation-based were the least. This study achieved a high level of improvement for helping a developer, maintainer, researcher to get proper knowledge about the state of software evolution and visualization as a whole [5].

Mikhailov et al. [6] presented Systematic Literature Reviews to summarize the current state of the art in the field of ontology visualization. The authors collected papers that are relevant to ontology visualization from various fields. The selected papers included 52 papers from 2007 to 2015. After that, an analyze processes started for the papers to determine which visualization techniques used in ontology visualization. The authors found that the conventionally used visualization techniques were: 2D graphs, 3-dimensional graph, UML classes, Radial tree, and Concept diagrams. The most common way of ontology visualization was plain 2D graphs [6].

B. Reviews from users

Gundla and Otari in 2015, proposed an interactive system that visualized sentiment analysis to help users in the decision-making process. The dataset was collected from online customer reviews of products from different websites periodically. Their system depends on analyzing reviews in which each word in the review tagged with its part of speech (i.e., noun, adjective, etc.) and they would easily retrieve the nouns as product features, and the adjectives as opinion words. After that, the product features from each sentence extracted then the semantic orientation of each opinion word was identified to know whether an opinion word classified as a positive, negative or neutral. The summary of the products was visualized positive and negative sentiments and offered as graphs such as pie charts and bar graphs [7].

In 2016, Mugdha and Pradnya proposed a system to improve the quality of sentiment analysis on textual reviews using the Hadoop framework. The reviews on Kindle extracted from Amazon. Then, they used the Natural Language Tool Kit (NLTK) library for data preprocessing. For sentiment classification, a map-reduce environment used to implement Naïve-Bayes classifier to identify the polarity of review whether it is positive, negative or neutral [8].

Aashutosh et al. in 2015, proposed a system that visualized sentiments extracted from reviews about iPhone5. The authors applied a classification rule to classify all reviews into service, product and feature reviews. After the classification process, they used POS tagging and regular expressions to extract the sentiments from each feature review. The polarizer method used in which it receives extracted sentiment and returns (1) for positive sentiment and (-1) for negative sentiment. The user profile picture on the reviews shown as green smiley, yellow smiley, and red smiley for helping the user to recognize the review with good, average, and bad rating respectively. They displayed an attractive dashboard that consists of bar charts and pie charts which helps users to understand the sentiment easily and quickly [9].

In 2015, Kamal designed a system to summarize and visualize the feature based on opinion, which is called (OSVS). A supervised machine learning technique used for classification of review sentences. The author applied natural language processing techniques to mine information components of reviews as a feature, opinion or optional modifier. Then, he determined the polarity values: negative, positive or neutral of opinion words. Kamal used methods for distinguishing subjective and objective sentences, extracting feature opinion and classifying sentiment. Finally, the results presented in a graphical representation such as a bar chart and pie chart. Similarly, the color scheme was used to highlight the extracted information components from review documents [10].

Chen et al. in 2016, proposed an interactive system that visualizes sentiment pairs extracted from reviews about hotels. The reviews included customers' satisfaction with cleanliness, service, comfort, condition, and neighborhood. In their study, they used Latent Dirichlet Allocation (LDA) to get a better understanding of visualization. Therefore, users can choose any topic to filter irrelevant reviews on a topic basis. Also, they proposed an interactive visualization for offering summaries of reviews. The system visualized sentiment polarity of reviews. It shows topics of reviews and interactive function to modify sentiment that is not suitable. A node for each sentiment pair used. Nodes have different colors according to sentiment polarity. Additionally, they employed a dropdown list to select a topic to focus on particular sentences. Also, this study adopted the highlighting of keywords. The positive, negative, and neutral sentiment words had colored with green, red, and gray, respectively [11].

In 2015, Qiu et al. showed their interest in public opinions and sentiment trends in higher education services in the US. Through applying their study, they enhanced their previous system that is called (eduMRS). eduMRS is an education monitoring and ranking system. The study introduced means for aggregating and visualizing of public textual comments to empowering education ranking systems. An interactive map view of sentiment analysis of public opinions applied. The different size of circles on the map represents sentiment in each state in the US. The size of a circle varies with the aggregated value of sentiments calculated according to a college at the time when the data was collected. As a result, the study indicated that aggregation and visualization of public opinions would lead to help stakeholders understand the quality of provided education services [12].

In 2016, Li et al. provided a visual analytic system that

visualizes the opinions of Chinese travelers about everything related to a tourism domain. In this study, three views were included in the system, the interactive filtering view, which can select travel notes and comments, the content view, which is used to express comments and tourists' emotion changes, and the pop-up information view, it shows social relationships of tourist and tag cloud of comments. The authors used a temporal histogram, map, sentiment analyzer, comments and notes list, hierarchical structure and word cloud. The sentiments were separated into three (positive, neutral and negative) and used different colors corresponding to different sentiments. The system could provide three of analysis tasks like tourist's regional tendency analysis, tourists' sentiment changes analysis and tracing and analyzing the social networking relationships from specific tourist notes. The result showed that the proposed system has usefulness and effectiveness to analyze tourists' regional tendency and emotional changes [13].

In 2014, Lee et al. proposed a visualization approach to find hidden relations between movies and their evaluation. The authors analyzed the patterns with reviews, to find out the influence of the word-of-mouth effects. Thus, their approach will provide users with a clear identification of which movie is favorite or not. Together with the plotting, they placed a bar chart beside the customer's plot to compare pre-attentively which movie has the most or the least number of customers. They proposed two kinds of visualizations on the screen. Firstly, they draw the line chart with normalized values from the number of customers and reviews. Secondly, they analyzed the difference between any movies; they implemented the users' option to be able to select any movies from the dataset [14].

In 2017, Amigud et al. proposed a novel approach proposed to visualize the textual data that depicts information on a continuum (temporal or spatial), it should allow inferences to made about the thematic organization of a document and its structure. ThemeTrack has been used to allow the user to create a visual map of the textual data and obtain a succinct summary of the information it carries. The authors conducted experiments on the dataset that consists of academic texts and journal articles. These experiments aim to compare the visual representation of information extracted using contiguous bigrams and trigrams to represent themes, and syntactic parsing technique utilizing verb-noun pairs to represent actions. Consequently, these visualizations provide a summary of how content is structured [15].

Xu et al. in 2016, proposed a new approach based on semantic word cloud, considering the semantic meaning of words for reviews of three different fields: restaurant, movie, and product. They applied the distributed word representation to describe the semantic meaning of words accurately. Furthermore, constructing a word similarity graph based on the semantic distance between words to lay out words more compactly and aesthetically. The motivation behind the semantic word cloud was to preserve the semantic meanings of words to reveal general themes of texts for summary and exploration intuitively. The proposed word cloud applied to user-generated reviews in different fields to study the effectiveness of the method [16].

In conclusion, many papers discussed the different types of reviews and opinions about movies, books, shopping centers, restaurants, hotels, and devices like the iPhone and Kindle. In visualization context, these studies employed various visualization techniques to present reviews and opinions to meet the needs of the user and the stakeholders. Word cloud, parallel coordinate plot, line chart, map, pie charts, bar graphs, and histogram conducted. Meanwhile, Word cloud and line chart were the most techniques used extensively since they provide a good representation of reviews as shown in [8] [10] [11] [13] [14]. Therefore, the system utilizes these types of visualization techniques and appends other techniques that particularly suitable for the system purpose where the user can effectively visual analysis of customer feedback. Also, the user can interact with these techniques, for example, selecting a specific book and determine which the amount of rating it has and if the rating is positive, negative or neutral where the user can choose one of the three sentiments and all techniques directly are adjusting based on user choice. Therefore, the system answers many of questions about the products and take quick making decisions.

III. System Design

The proposed system aims to provide a visual approach for book reviews to assist users in finding understandings on different books. The system consists of two main parts integrated to answer the research questions: sentiment analysis and representations of visualization techniques (see Figure.1). The system passed on three phases: handling with raw data then sentiment analysis and the last phase provides dashboard has many visualization techniques where the user interacts with it. The method involved a combination of the visualization processes of Tableau [17] with the analytics of R.

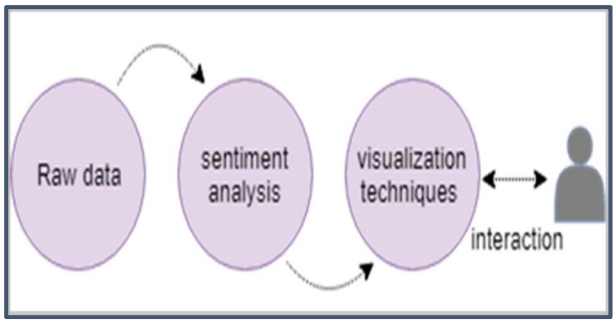

Figure.1. System Architecture

A. Raw Data Preprocessing

The dataset used in this research about the reviews of books from Amazon [18], which is available to the public. The books' reviews obtained are in a computer-ready form, similarly in JSON format. The used dataset formed from two separate datasets. The first dataset represents the product description and includes 9 attributes (see Table I) while the second dataset is about the review and has 8 attributes (see Table II). In this study, we focused on four attributes: overall, summary, title and time of reviews from both datasets.

TABLE I: REVIEW STRUCTURE

Features	Description
ReviewerID	ID of the reviewer, e.g., A2SUAM1J3GNN3B.
Asin	ID of the product, e.g., 0000013714.
ReviewerName	Name of the reviewer.
Helpful	Helpfulness rating of the review, e.g., 2/3.
ReviewText	Text of the review.
Overall	Rating of the product.
Summary	Summary of the review.
UnixReviewTime	Time of the review (Unix time).
ReviewTime	Time of the review (raw).

TABLE II: PRODUCT DESCRIPTION STRUCTURE

Features	Description
asin	ID of the product, e.g., 0000031852
title	Name of the product
price	Price in US dollars (at time of crawl)
imUrl	URL of the product image
related	Related products (also bought, also viewed, bought together, buy after viewing)
salesRank	Sales rank information
brand	Brand name
categories	List of categories the product belongs to

B. *Sentiment analysis process*

Sentiment analysis, which is also called Opinion mining, means the handling of natural language and text analysis. It is a classification process to discern the subjective opinion about a specific product, topic, etc. In this paper, we carried out a sentiment analysis on several books' reviews to determine whether a review of a book is positive, negative or neutral. R is a powerful analysis language where several researchers and data scientist uses it. Therefore, we use Rstudio IDE 1.1.383 [19] to classify the summary of 12 books' reviews into positive, negative and neutral. There are two packages in R that could use for the sentiment purpose: *sentiment*, which used in our study, and *qdap*. The *sentiment* package requires installation of *tm* and *Rstem* packages. After that, the R script created as a calculated field Classification applied using the *classify_polarity* function in Tableau.

C. *Visualization techniques*

According to Spence [20], "Visualization is the formation of a mental model of something." Several visualization techniques have been combined to provide a visual system that aids users to explore different information about many books. Packed bubbles, linear chart, stacked bars, and word-cloud are the four charts used in the study in order to achieve the aim.

- *Packed bubbles* used to present the total number of records of each book.
- *Linear chart* utilized to offer the time when the reviews were written.
- *Stacked bars* used to present the ratings of books and whether they are positive, neutral or negative.
- *Word-cloud* used to show the most frequently used word in the reviews of books. It also shows the books that get the most reviews.

The four components are synchronized in one interface and provide the following functionalities:

- Meta overview using four visualization techniques of books' reviews.
- Interactive filtering, to select a specific book and determine the reviews positive, neutral or negative.
- Details on demand, the user can have detailed information about a book that in his/her interest.

IV. RESULTS

The following subsections describe the visualization techniques employed to analyze the data and the overall interface.

A. *Identify the highly reviewed book*

The dataset includes 12 books reviews selected from Amazon, and the total records were about 1000 records. The word cloud which famous can also be used to compare different bodies of text together where we utilized to get a basic view of the title book which is mostly reviewed by customers (see Figure.2). As shown from the figure, the book *Water for elephant* has the most reviews between the 12 books. The user can click on the title to filter the other visualization techniques on the dashboard concerning the selected book.

Figure.2. Reviews Word Cloud

B. *Determine the frequency of words of the sentiment on a book or all books*

In order to get a general look about how many positive, negative and neutral reviews for the books, packed bubbles are used with different colors as shown in *Figure.3*. In packed bubbles, the frequency of each review counted for collecting the different sentiment words. The size of the bubbles shows reviews for different books. The color of the bubbles shows the tendency of sentiment to positive, negative or neutral (the green is positive, the red is negative the orange is neutral). As demonstrated, the positive reviews are the highest reviews whereas; the negative and neutral reviews have few reviews compared to the positive.

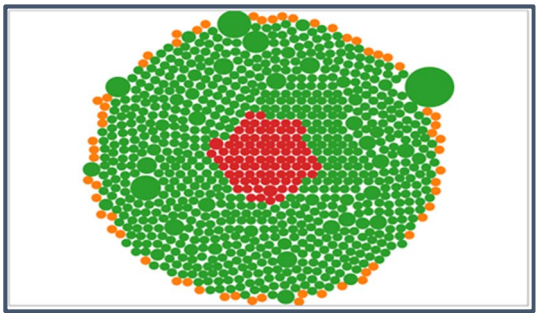

Figure 3. Packed Bubbles of Sentiment Reviews

On the other hand, the word cloud of the sentiment reviews is generated to illustrate the frequency as words with different colors as shown in *Figure.4*. The function of visualizing the most common words in the review summary of each book can help the user to distinguish the overall users' opinions of that book easily.

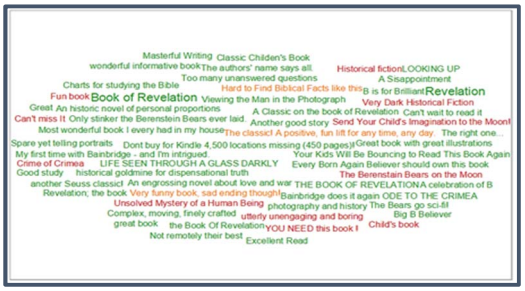

Figure 4.Word Cloud of Sentiment Reviews

C. *View the review's sentiment over the years*

The line chart is more appropriate to use to track changes over periods. *Figure.*5 includes three lines: positive, neutral and negative for all the reviews from 1998 to 2014. We can see that the number of totally negative and neutral reviews of books increased slower than the increase in positive reviews. Accordingly, positive reviews show more significant variance than others and we can discover the earlier book reviews received better reviews than later ones.

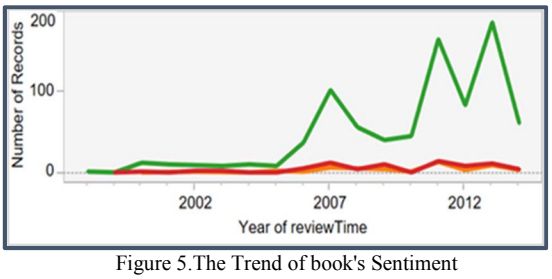

Figure 5.The Trend of book's Sentiment

D. *Identify the rating for a book or all books according to the sentiment*

Stacked bar graph provides an excellent mean to represent and show comparisons of multiple types of data within a single bar. This type of visualization allows the users to know the total rating of a book and the relationship between the rating and the sentiment whether positive, neutral or negative. *Figure.6* shows that the book *Water for elephant* has the most reviews with a rank value of 5 and simultaneously most of them are positive as shown in *Fig.6*.

Figure 6.The Rating of Books

E. *Overall Dashboard*

All the previous visualizations combine in one dashboard to answer questions and provide a better direct interactive system for the users. As shown in *Figure.7*, the user able to select a book to see the total reviews, with its classification into positive, neutral and negative in different representations. Each representation can give a clear and flexible recommendation about one or more books. Also, the user able to filter the results according to the sentiment that allows more interesting and subtle sentiment analysis. All visualizations techniques could simultaneously be controlled.

V. CONCLUSION

In recent years with the vastly increasing amount of customer data around the digital world, text analysis is gaining more adoption. In this paper, we discussed many papers in the systematic approach to finding out the importance of using visualization. Additionally, several papers in the text visualization field have reviewed especially, in opinion and reviews visualization. Using Tableau software integrated with R, we presented about 1000 of reviews from Amazon's books using visualization. Different visualization techniques have been used for saving user's time and effort in terms of knowing the most sentiment reviews on the specific book whether positive, negative or neutral. The user can visualize the relationship between the previous reviews and the later and know which book has the most reviews in comparing with other books in the dataset.

Moreover, the user can easily compare the different sentiment reviews of the different books. Some of the results obtained are not too perfect, where some reviews are positive but classified as neutral. The reason behind that is R's sentiment package follows a lexicon based approach. Therefore, the incorrect classifications are more likely to be slang, colloquial and jargon words that are being used in the reviews we are analyzing since these words are not covered extensively in common lexica.

VI. FUTURE WORK

The authors are planning to evaluate the developed system with different products reviewed through different vendors. Also, conduct a usability study in order to verify systems' results.

Furthermore, the authors wish to extend the study to develop an adaptive system for visualizing sentimental reviews of different products from different vendors so that to recommend

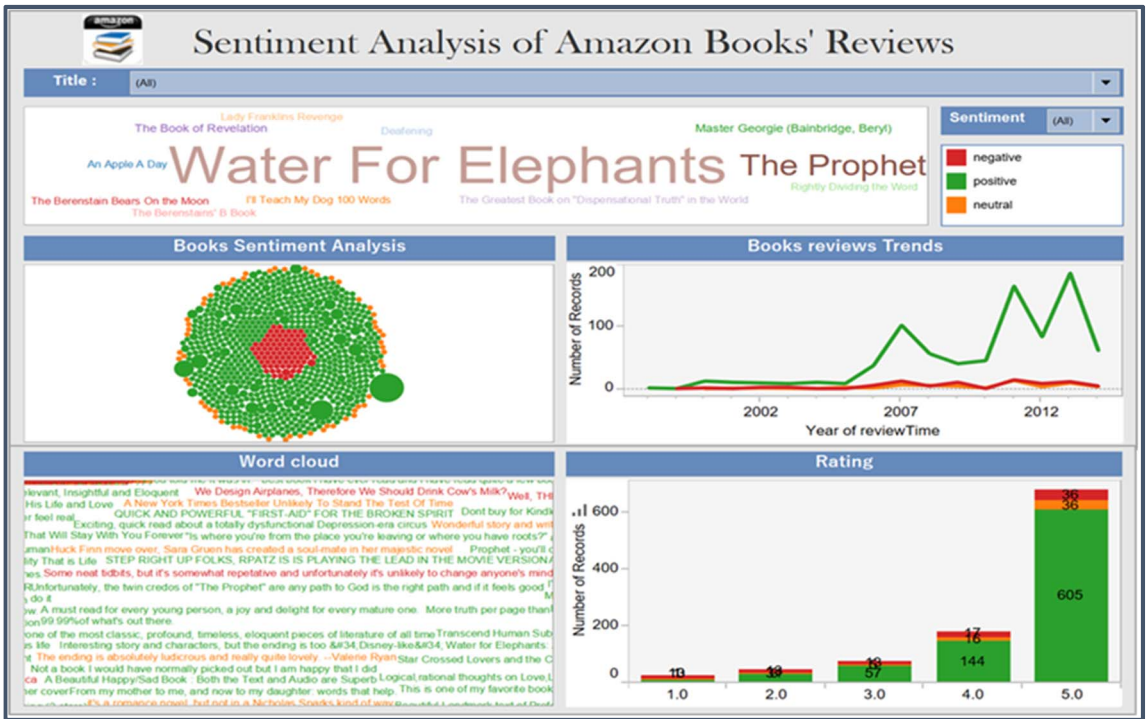

Figure.7. The Dashboard for Amazon Books Sentiment Analysis

.

The best product from the best seller. Also, adding new words to the lexicon will support the results. Including more semantic features in text, mining may enhance the results

REFERENCES

[1] K. Dave, S. Lawrence, and D. M. Pennock, "Mining the peanut gallery: Opinion extraction and semantic classification of product reviews," in Proceedings of the 12th international conference on World Wide Web, 2003, pp. 519-528: ACM.

[2] S. Liu, W. Cui, Y. Wu, and M. Liu, "A survey on information visualization: recent advances and challenges," The Visual Computer, vol. 30, no. 12, pp. 1373-1393, 2014.

[3] A. Šilić and B. D. Bašić, "Visualization of text streams: A survey," in International Conference on Knowledge-Based and Intelligent Information and Engineering Systems, pp. 31-43: Springer,2010.

[4] A. Godwin, "Visualizing Systematic Literature Reviews to Identify New Areas of Research," Front. Educ. Conf., p. 8, 2016.

[5] H. B. Salameh and A. Aljammal, "2016 7th Int. Conf. Comput. Sci. Inf. Technol., pp. 1–6, 2016.

[6] S. Mikhailov, M. Petrov, and B. Lantow, "Ontology visualization: A systematic literature analysis," CEUR Workshop Proc., vol. 1684, pp. 1–12, 2016.

[7] A. Gundla, M. Otari, "A Review on Sentiment Analysis and Visualization of Customer Reviews," International Journal Of Engineering And Computer Science, 2015.

[8] M. Jinturkar, P. Gotmare, "Sentiment Analysis of Customer Review Data using Big Data: A Survey," IJCA Proceedings on Emerging Trends in Computing, pp. 3–8, 2016.

[9] A. Bhatt, A. Patel, H. Chheda, and K. Gawande, "Amazon Review Classification and Sentiment Analysis," International Journal of Computer Science and Information Technologies, vol. 6, pp. 5107–5110, 2015.

[10] A. Kamal, "Review Mining for Feature-based Opinion Summarization and Visualization," International Journal of Computer Applications, vol. 119, no. 17, pp. 6-13, 2015.

[11] Y. S. Chen, L. H. Chen, and Y. Takama, "Proposal of LDA-Based Sentiment Visualization of Hotel Reviews," Proc. - 15th IEEE Int. Conf. Data Min. Work. ICDMW 2015, pp. 687–693, 2016.

[12] R. G. Qiu, R. R. Ravi, and L. L. Qiu, "Aggregating and visualizing public opinions and sentiment trends on the US higher education," Proc. 17th Int. Conf. Inf. Integr. Web-based Appl. &Services - iiWAS '15, pp. 1–5, 2015.

[13] Q. Li, Y. Wu, S. Wang, M. Lin, X. Feng, and H. Wang, "VisTravel: visualizing tourism network opinion from the user-generated content," J. Vis., vol. 19, no. 3, pp. 489–502, 2016.

[14] L. Jaehoon, G. Noh, and C. Kim, "Analysis & visualization on movie's popularity and reviews," International Conference on Big Data and Smart Computing (BIGCOMP), pp. 189–190, 2014.

[15] A. Amigud, J. Arnedo-Moreno, T. Daradoumis, and A.-E. Guerrero-Roldan, "A Method for Thematic and Structural Visualization of Academic Content," in Advanced Learning Technologies (ICALT), 2017 IEEE 17th International Conference on, 2017, pp. 230-234: IEEE.

[16] J. Xu, Y. Tao, and H. Lin, "Semantic word cloud generation based on word embeddings," in Pacific Visualization Symposium (PacificVis), 2016 IEEE, 2016, pp. 239-243: IEEE.

[17] "Answer questions as fast as you can think of them with Tableau | Tableau Software," Tableau Software, 2017. [Online]. Available: https://www.tableau.com [Accessed: 20-Oct-2017].

[18] J. McAuley, "Amazon review data," UCSD. [Online]. Available: http://jmcauley.ucsd.edu/data/amazon/links.html. [Accessed: 17-OCT-2017].

[19] "Download RStudio", RStudio, 2017. [Online]. Available: https://www.rstudio.com/products/rstudio/download/. [Accessed: 24- Oct- 2017].

[20] R. Spence, Information Visualization. Human-Computer Interaction: Design Issues, Solutions, and Applications. 2001.