

SMVS: A Web-based Application for Graphical Visualization of Malay Text Corpus

Noor Bazilah Ahmat Baseri
School of Computing,
Universiti Utara Malaysia
Kedah, Malaysia
noorbazilah9616@gmail.com

Juhaida Abu Bakar
School of Computing,
Universiti Utara Malaysia
Kedah, Malaysia
juhaida.ab@uum.edu.my

Azizah Ahmad
School of Computing,
Universiti Utara Malaysia
Kedah, Malaysia
azie@uum.edu.my

Hawa Jafferi
School of Computing,
Universiti Utara Malaysia
Kedah, Malaysia
hawajafferi@gmail.com

Muhammad Faiz Zamri
School of Computing,
Universiti Utara Malaysia
Kedah, Malaysia
faizamri.muhd@gmail.com

Abstract— Information visualization is an interesting field nowadays. A good information visualization ensures distraction of misleading information is not included in the visualization. Many studies have been conducted on the Quranic corpus. The advancement technology coupled with modern approach of the computer technology can support the learners to understand Qur'an easily. Smart Malay Visualization System (SMVS) is a Python Flask framework web application which help users efficiently to produce the most basic data visualization from a big data. This web application displayed information from the state-of-the-art corpus which is identified through text. Agile development has been adapted to prepare this web application. Six phases of the methodology have been implemented in this study which are requirements, analysis, planning, design, implementation, testing, and deployment. Natural Language Processing approach has been used to visualize the data. Twenty most informative word from each verse has been visualized using Frequency Distribution and has been embedded to the web application. This work focuses on the Malay translation of the Qur'an corpus.

Keywords— *Big data, data visualization, knowledge representation, Qur'an knowledge, natural language processing*

I. INTRODUCTION

Information visualization is the process of representing data in a visual and meaningful way in order to make a better understanding. Visualising aspects of corpus data can be useful for discovery as well as for communicating results [1]. Meanwhile, data visualization is a similar study which focused on the presentation of data in a pictorial or graphical format. As stated in [2], the volumes of digital collections continue to grow and the traditional methods become increasingly ineffective, leading to a transition on the use of graphs and maps to interpret textual data. Visualizations are useful to ease understanding of large amounts of data quickly.

The views of information and visualization deduce two important aspects: (1) information visualization is used to discover new insights and knowledge from abstract data through graphical, and (2) information visualization can be considered a representation of data that amplifies cognition [3]. Importantly, right visualization types must be chosen to ensure distract or mislead information are eliminated.

In addition, the major problem with data visualization is confusion of data [4]. Study in [4] stated that graphical visualization with no accompanying text and lack of clear

overall logic may confuse the viewers. Furthermore, many visualizations may have implicit meaning which adds to the problem of not explained thoroughly and may be misinterpreted. Some sentences are fundamentally difficult to understand which lead to visualizations that depict many complex relationships and are not optimally represented.

Meanwhile, Information extraction is the process of extracting specific (pre-specified) information from textual sources. It remains a fundamental challenge for any system that works with structured data [5]. The process involves transforming an unstructured text or a collection of texts into sets of facts. The good information extraction solutions are a combination of automated methods and human processing. Tokenization process is used in the extraction information before representing the visualize data.

In computational linguistic, text corpus is a large and structured sets of texts within a specific language territory [1]. Researchers used state-of-the-art corpus in various type of subject such as speech recognition and machine translation [6-7].

Many studies have been on Qur'an corpus such as studies in [8-10]. The understanding of the Quranic knowledge does not only require appropriate teachers, computer technology can also support the learners to understand Qur'an easily, especially in the web and mobile-based environment [11]. The Qur'an consists of 30 divisions (Juz), 114 chapters (Surah), and 6236 numbered verses (ayat). Many efforts to create online information systems for the Qur'an knowledge has been implemented such as work by [8-11]. In this study, the focus is on the final division, or known as Juz Amma.

Three objectives are developed for this study, which are (i) To gather requirements for Malay data visualization system which can be operate online, (ii) To built-up web application for the Malay data visualization, and (iii) To evaluate web application for the Malay data visualization system.

II. RELATED STUDY

In recent years, data visualization for the big data studies were popular to support literary scholars and experts from other domains. Visualization is the process of transforming data, information, and knowledge into visual forms of which we could have insights [12].

Visualization is not an end but a means toward an end, which is understanding. Basically, one does not speak about

visualizing a diagram but visualizing a concept or problem. To visualize a diagram means simply to form a mental image of the diagram, but to visualize a problem to understand the problem in terms of a diagram or visual image [13]. A graph visualization for Text Variant Graphs model which highlight in a data structure that represent various editions of a text is proposed by [14].

Information extraction methods can be used by employing approaches such as Natural Language Processing (NLP), Text Mining and Data Mining. In this study, the NLP approach has been used to extract valuable information. Natural language processing (NLP) method extracts structure from textual representations of information. As in [15] the interface can be used to generate visual interpretations of the semantic content of a given natural language that can be then visualized either as a static scene or a dynamic animation. Many studies have used the NLP approach such as studies in summarization [16], semantic understanding [17], name entity identification [18-19], and discourse [20]. By using all these body of knowledge, creating useful visual representation from the textual information sources is a current new direction.

III. METHODOLOGY

The study was conducted following agile methodology. Agile development methodology is used in order to implement this web application since it has iteration. This methodology consists of requirements analysis, planning, design, implementation, testing, and deployment. In this project, agile model is used as the methodology of the SMVS. In this Agile Model, prior planning, the requirement for the system is identified. Then, the model is developed and tested, and the design is ready for coding and testing. This will be based on user requirement. The flow of the phases is illustrated in Fig. 1.

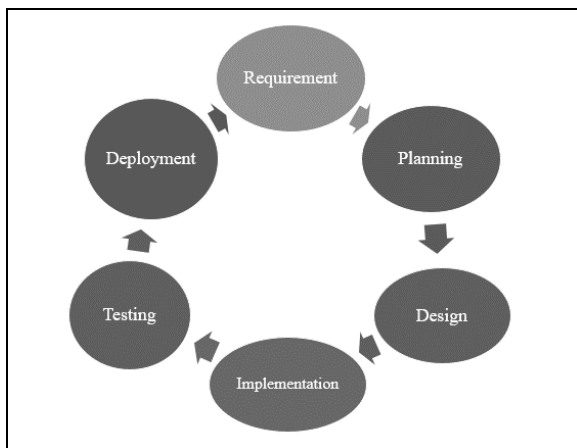


Fig. 1. Agile development methodology.

There are several techniques that can be used to analyze the user **requirements** such as case diagram that shows the users that had been assigned to the requirements. Then, the sequence diagram is derived as user manual because it explains the function of the systems based on user requirements. Sequence diagram is a design based on the use case specification. The **planning** of the system has to be made when all the requirement are already identified. At this

phase, system will be planned and visual will be identified for the data visualization system. This system will display information from a big data into data visualization.

Next, the **design** phase is about the design of the system before proceeding to low-fidelity based on the user requirements. Class diagram will be used to help in identifying and understanding the requirements of the problem domain and to identify its components.

The installation is needed in the **implementation** stage. Class diagram has been used to convert a model into code in this phase. Then, the developer would install the right software in order to develop the system. There are several software can be used to develop high fidelity application such as Python, Flask framework, Dreamweaver, and Notepad.

In the development process, **testing** is an important phase. This phase ensures that the software and hardware function well closely to the target performance. Test case should include ID, action, expected result, actual result, and to decide the functionality of the test. **Deployment** focuses on production and technical support. It is the stage where the system has been developed and ready to be released to the user. The user can use and give feedback on the developed system. Since developer uses agile development, it has iteration where it will start from the first stage of next iteration if any changes or updates occur.

IV. DESIGN AND DEVELOPMENT OF SMVS

This section describes the design and development of a web based for creating visualization system. The section is divided into two sub-sections; (1) the requirements of the web based for creating visualization system, and (2) the prototype development of SMVS; a web-based application developed to demonstrate the gathered requirements.

Two ways were employed to gather the requirement process such as (1) to select the respondents based on the suitability of the respondents, and provide the questionnaire through Google form link, while the respondents answer all questionnaires given, and (2) from the internet, systems that are related to the visualization system are analysed. with the questionnaire also included open-ended questions on the features of the web-based application. The questionnaire consisted of four sections, namely, usefulness, ease of use, satisfaction, and security.

The other requirement for gathering data was through search engine platforms. The article was being search at Google search engine platform by using many key words such as, “visualization”, “Malay visualization system”, “data visualization”, “verse Qur’an” and etc. Other than that, related article of system was searched in relevant websites. All the findings were analyzed and ideas were formed to solve the problem. Then, ideas for this system requirement such as register, login, manage content, manage usability experience design and manage visualization process were also deduced. Table I shows the requirement involve in SMVS.

TABLE I. LIST OF REQUIREMENTS FOR SMVS

ID	Requirement Description	Priority
1	Account Registration	
1.1	Users must register their details such as username, email and password	M
	The system shall notify if the registration form is incomplete	D
2	Login	
2.1	The user that have registered, must login using email and password	M
2.2	The user that forgot their password, they can retrieve password using email	D
2.3	Admin must verify the validity of user's email and password	M
3	Manage Content	
3.1	Users (students and lecturer) can add data or information	M
3.2	Users (students and lecturer) can delete data or information	D
3.3	Users (students and lecturer) can update the data or information	D
3.4	Users (students and lecturer) can search the types of visual	D
3.5	Users (students and lecturer) can click cancel before the system translate into visual	D
3.6	Users (students and lecturer) will get the result when admin confirms for visualize the data	M
4	Manage Usability Experience Design	
4.1	Admin can create contact us information	M
4.2	Admin can read contact us information	M
4.3	Admin can update contact us information	O
4.4	Admin can delete contact us information	O
4.4	The system will display contact us information	O
5	Manage Visualization Process	
5.1	Admin can add the data or information	M
5.2	Users (students and lecturer) can view the result of the information efficiently	M
6	Social Media Marketing	
6.1	Admin can create embed to social media icon	O
6.2	Admin can read embed to social media icon	O
6.3	Admin can update embed to social media icon	O
6.4	Admin can delete embed to social media icon	O
6.5	Users (students and lecturer) can click social media icon for sharing	M
7	Logout	
7.1	Admin and users shall be able to logout	M

The requirements presented in Table I were translated into the computer system functionality. The next process is visualizing and modelling the requirements using the appropriate modelling method and tools. In this work, the Unified Modelling Language (UML) was used to visualize and model the requirements. The models used in this work are two behavioural diagrams namely use case and sequence diagram, and a class diagram that represents the structural components of the system. The diagrams were drawn using StarUML and Violet. Seven major use cases are register, login, manage content, manage usability experience design, manage visualization process, social media marketing and logout.

Visualization process starts when the web application is started. The text corpora are fed into the linguistic processing

pipeline that applies a tokenizer, a sentence splitter and stopword. After tokenization and sentence splitting, the pipeline identifies all token that occur in more than one of the source texts. After pre-processing, identification of the most informative about the text corpus are made. The frequency distribution technique has been used in this study which informs the frequency of each vocabulary item in the text. The distribution information informs how the total number of word tokens in the text are distributed across the vocabulary items. In this study, Natural Language Toolkit (NLTK) using Python 2.0 provides built in support for the *FreqDist* function that has been embed to the web application. Based on the twenty most informative occurrence, an interactive visualization shows the cumulative frequency plot of each verses in the web application. Fig. 2 shows the flowchart for visualize process of SMVS.

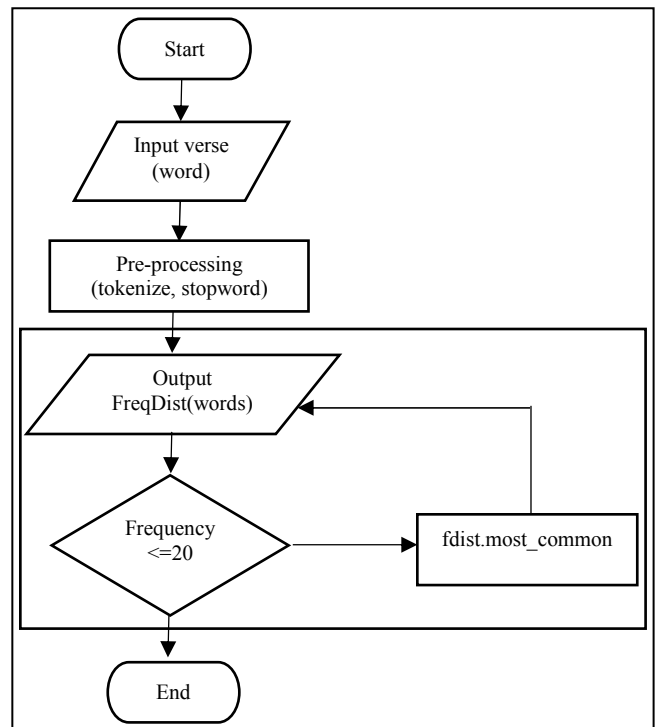


Fig. 2. The flowchart for visualize process of SMVS.

V. THE SMVS PROTOTYPE DEVELOPMENT

A prototype of the web application for creating visualization system named SMVS was developed. It represents the requirements explained in the previous subsections. This system used DB SQLite as a platform to save the data like authentication and database for storage. Here, Fig. 3-6 are screenshot and selected from SMVS.



Fig. 3. Main page.

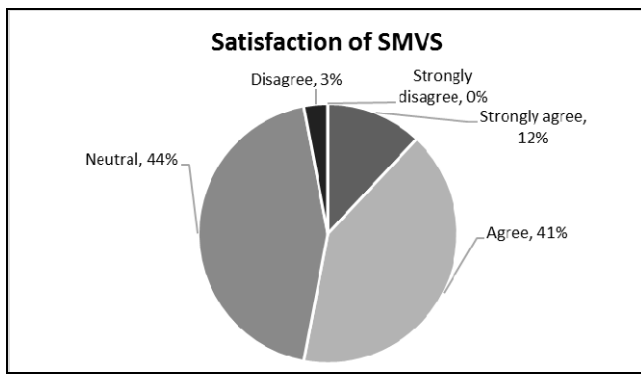


Fig. 9. The average of satisfaction of SMVS.

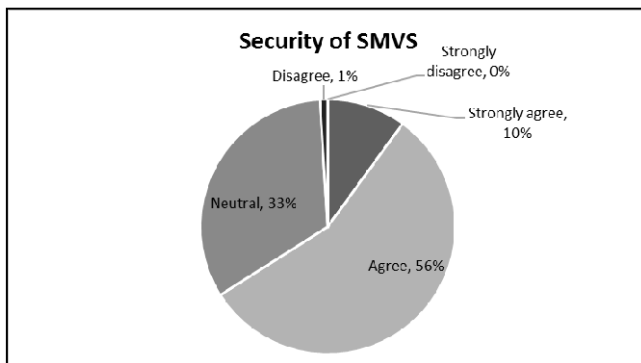


Fig. 10. The average of security of SMVS.

VI. EVALUATION OF SMVS

This paper describes how a system can help students, and lecturers to understand the information created through text. There are many aspects of visualization that can be studied. In the future, we plan to expand the functionality of SMVS by providing more text form verses in the Qur'an. Other than that, we also plan to make an improvement with the SMVS' design and function. Furthermore, we also plan to make the web application to be visualized from more sources and any text information such as articles or books. This can encourage users to use the web application and increase their knowledge. Hence, the users will be attracted to use this web application, because it can give them new knowledge.

ACKNOWLEDGMENT

We thank the Malaysian Ministry of Higher Education as well as UUM under SO Code 13849 for their support of this study.

REFERENCES

- [1] Allen, W. (2017). Making corpus data visible: Visualising text with research intermediaries. *Corpora*, 12(3), 459-482.
- [2] Scrivner, O., & Davis, J. (2017). Interactive Text Mining Suite: Data Visualization for Literary Studies. In *CDH@ TLT* (pp. 29-38).
- [3] Chen, H. M. (2017). . An Overview of Information Visualization. *Library Technology Reports*, 53(3), 5-7.
- [4] Eppler, M. E., & Burkhard, R. A. (2004). Knowledge visualization. Towards a new discipline and its fields of application. In: *ICA Working Paper*, Lugano, University of Lugano.
- [5] Dalvi, B., Bhakthavatsalam, S., Clark, C., Clark, P., Etzioni, O., Fader, A., & Groeneveld, D. (2016, June). IKE-an interactive tool for knowledge extraction. In *Proceedings of the 5th Workshop on Automated Knowledge Base Construction* (pp. 12-17).

- [6] Di Gangi, M. A., Cattoni, R., Bentivogli, L., Negri, M., & Turchi, M. (2019). MuST-C: a multilingual speech translation corpus. In *2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 2012-2017). Association for Computational Linguistics.
- [7] Takamichi, S., Mitsui, K., Saito, Y., Koriyama, T., Tanji, N., & Saruwatari, H. (2019). JVS corpus: free Japanese multi-speaker voice corpus. arXiv preprint arXiv:1908.06248.
- [8] Bakar, J. A., Omar, K., Nasrudin, M. F., & Murah, M. Z. (2016). NUWT: Jawi-Specific Buckwalter Corpus for Malay Word Tokenization. *Journal of Information & Communication Technology*, 15(1).
- [9] Husin, M. Z., Saad, S., & Noah, S. A. M. (2017, November). Syntactic rule-based approach for extracting concepts from quranic translation text. In *2017 6th International Conference on Electrical Engineering and Informatics (ICEEI)* (pp. 1-6). IEEE.
- [10] Yunus, M. A. M., Mustapha, A., Iqbal, R., & Samsudin, N. A. (2017). An Ontological Approach towards Dialoguebased Information Visualization System: Quran Corpus for Juz'Amma. In *MATEC Web of Conferences* (Vol. 135, p. 00070). EDP Sciences.
- [11] Ali, B. B. M., & Ahmad, M. (2013). Al-Quran themes classification using ontology. *ICOI. Cms. Net. My*, 74, 383-389.
- [12] N. Gershon, S. G. Eick, S. Card/ (1998). "Information Visualization", *Interactions*, vol. 5, no. 2.
- [13] Zimmermann, W., & Cunningham, S. (1991). Editor's introduction: What is mathematical visualization? *Visualization in teaching and learning mathematics*, 1-7.
- [14] T. L. Andrews, J. J. Van Zundert. (2013). "An Interactive Interface for Text Variant Graph Models", *Proceedings of the Digital Humanities*.
- [15] Hassani, K., & Lee, W. S. (2016). Visualizing natural language descriptions: A survey. *ACM Computing Surveys (CSUR)*, 49(1), 17.
- [16] M. John, E. Marbach, S. Lohmann, F. Heimerl, and T. Ertl. (2018). "Multicloud: Interactive word cloud visualization for the analysis of multiple texts," in *Proceedings of Graphics Interface 2018*, pp. 34 – 41, Canadian Human-Computer Communications Society / Societ' e canadienne du dia- logue humain-machine.
- [17] Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training. URL https://s3-us-west-2.amazonaws.com/openai-assets/researchcovers/languageunsupervised/language_understanding_paper.pdf.
- [18] Salleh, M. S., Asmai, S. A., Basiron, H., & Ahmad, S. (2017, May). A Malay Named Entity Recognition using conditional random fields. In *2017 5th International Conference on Information and Communication Technology (ICoICT)* (pp. 1-6). IEEE.
- [19] Sazali, S. S., Rahman, N. A., & Bakar, Z. A. (2016, August). Information extraction: Evaluating named entity recognition from classical Malay documents. In *2016 Third International Conference on Information Retrieval and Knowledge Management (CAMP)* (pp. 48-53). IEEE.
- [20] Alias, S., Mohammad, S. K., Hoon, G. K., & Sainin, M. S. (2018, March). Understanding Human Sentence Compression Pattern for Malay Text Summarizer. In *2018 Fourth International Conference on Information Retrieval and Knowledge Management (CAMP)* (pp. 1-6). IEEE.

APPENDIX A

TABLE II. THE RESPONDENTS' RESPONSES ON THE PERCEIVED USEFULNESS OF SMVS

The post-task questionnaire items	Strongly disagree	Disagree	Neutral	Agree	Strongly agree	Average
SMVS enhances my effectiveness on managing corpus.	0 (0.00)	0 (0.00)	1 (3.33)	22 (73.33)	7 (23.33)	4.20
I can register an account in SMVS without any error.	0 (0.00)	0 (0.00)	3 (10.00)	12 (40.00)	15 (50.00)	4.40
I can login the app with registered email and password.	0 (0.00)	0 (0.00)	1 (3.33)	16 (53.33)	13 (43.33)	4.40
The search button can function well.	0 (0.00)	0 (0.00)	0 (0.00)	14 (46.67)	16 (53.33)	4.53
I can choose time and date to set reminder.	0 (0.00)	0 (0.00)	0 (0.00)	16 (53.33)	14 (46.67)	4.47
The reminder functions according to set date and time.	0 (0.00)	0 (0.00)	3 (10.00)	11 (36.67)	16 (53.33)	4.43
It saves my time when I use this app to manage my corpus.	0 (0.00)	0 (0.00)	5 (16.67)	10 (33.33)	15 (50.00)	4.33
SMVS meets my needs.	0 (0.00)	0 (0.00)	5 (16.67)	17 (56.67)	8 (26.67)	4.10
SMVS does everything I would expect it to do.	0 (0.00)	0 (0.00)	2 (6.67)	23 (76.67)	5 (16.67)	4.10
SMVS is useful in overall.	0 (0.00)	0 (0.00)	0 (0.00)	19 (63.33)	11 (36.67)	4.37

TABLE III. THE RESPONDENTS' RESPONSES ON THE PERCEIVED EASE OF USE OF SMVS

The post-task questionnaire items	Strongly disagree	Disagree	Neutral	Agree	Strongly agree	Average
SMVS is easy to use.	0 (0.00)	0 (0.00)	1 (3.33)	22 (73.33)	7 (23.33)	4.20
SMVS is user friendly.	0 (0.00)	0 (0.00)	1 (3.33)	21 (70.00)	8 (26.67)	4.23
SMVS is flexible.	0 (0.00)	0 (0.00)	1 (3.33)	20 (66.67)	9 (30.00)	4.27
SMVS is easy to learn how to use it.	0 (0.00)	0 (0.00)	3 (10.00)	12 (40.00)	15 (50.00)	4.40
I can use SMVS without written instructions.	0 (0.00)	0 (0.00)	7 (23.33)	6 (20.00)	17 (56.67)	4.33
I can easily remember how to use SMVS.	0 (0.00)	0 (0.00)	3 (10.00)	17 (56.67)	10 (33.33)	4.23
I don't notice any inconsistencies as I use SMVS.	0 (0.00)	0 (0.00)	2 (6.67)	18 (60.00)	10 (33.33)	4.27
My interaction with the app would be clear and understandable.	0 (0.00)	0 (0.00)	3 (10.00)	17 (56.67)	10 (33.33)	4.23
I can use SMVS successfully every time.	0 (0.00)	0 (0.00)	3 (10.00)	14 (46.67)	13 (43.33)	4.33

TABLE IV. THE RESPONDENTS' RESPONSES ON THEIR SATISFACTION OF SMVS

The post-task questionnaire items	Strongly disagree	<i>a) Disagree</i>	Neutral	Agree	Strongly agree	Average
I am satisfied with SMVS.	0 (0.00)	0 (0.00)	3 (10.00)	22 (73.33)	5 (16.67)	4.07
I would recommend SMVS to my friends.	0 (0.00)	0 (0.00)	3 (10.00)	19 (63.33)	8 (26.67)	4.17
SMVS works the way I want it to work.	0 (0.00)	0 (0.00)	1 (3.33)	17 (56.67)	12 (40.00)	4.37
I feel I need to have SMVS app in my smartphone.	0 (0.00)	0 (0.00)	1 (3.33)	10 (33.33)	19 (63.33)	4.60
SMVS is wonderful and pleasant to use.	0 (0.00)	0 (0.00)	1 (3.33)	13 (43.33)	16 (53.33)	4.50