2024, 21(2 Special Issue): 0561-0567
https://doi.org/10.21123/bsj.2024.9675
P-ISSN: 2078-8665 - E-ISSN: 2411-7986

Baghdad Science Journal

Natural Language Processing For Requirement Elicitation In University Using Kmeans And Meanshift Algorithm

Devi Yurisca Bernanda*[1,2], Dayang N.A. Jawawi[1], Shahliza Abd Halim[1], Fransiskus Adikara[3]

[1]Department of Computer Science, Faculty of Engineering, Universiti Teknologi Malaysia, Johor Bahru, Malaysia.
[2]Department of Information System, Faculty of Technology and Design, Universitas Bunda Mulia, Jakarta, Indonesia.
[3]Department of Informatics, Faculty of Technology and Design, Universitas Bunda Mulia, Jakarta, Indonesia.
*Corresponding Author.

PARS2023: Postgraduate Annual Research Seminars 2023.

Received 26/09/2023, Revised 10/02/2024, Accepted 12/02/2024, Published 25/02/2024

© 2022 The Author(s). Published by College of Science for Women, University of Baghdad.
This is an Open Access article distributed under the terms of the Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Data Driven Requirement Engineering (DDRE) represents a vision for a shift from the static traditional methods of doing requirements engineering to dynamic data-driven user-centered methods. Data available and the increasingly complex requirements of system software whose functions can adapt to changing needs to gain the trust of its users, an approach is needed in a continuous software engineering process. This need drives the emergence of new challenges in the discipline of requirements engineering to meet the required changes. The problem in this study was the method in data discrepancies which resulted in the needs elicitation process being hampered and in the end software development found discrepancies and could not meet the needs of stakeholders and the goals of the organization. The research objectives in this research to the process collected and integrating data from multiple sources and ensuring interoperability. Conclusion in this research is determining is the clustering algorithm help the collection data and elicitation process has a somewhat greater impact on the ratings provided by professionals for pairs that belong to the same cluster. However, the influence of POS tagging on the ratings given by professionals is relatively consistent for pairs within the same cluster and pairs in different clusters.

Keywords: DDRE, Data Source, Requirement Engineering, System Software, Elicitation Process.

Introduction

Organizations are closely related to the need for information systems, adaptation of information technology, and information systems are the main drivers for organizations to develop their business [1]. Requirement engineering is important initial processes when developing software for an organization, including how data plays an important role in requirement engineering[2,3].

Requirement engineering (RE) is a collection of activities to identify and communicate the goals of the system, specifically the software, and the context in which the software will be used [3]. RE is bridge between the real-world needs of users, customers, and other constituents affected by a software system, and the capabilities and opportunities provided by software-intensive technologies [4-6].

The initial process of RE development is very useful for obtaining system functions that will be developed in software [5]. Software requirements engineering activities must be able to run correctly, completely and accurately so that the information system developed does not become backward, over budget, or even fails to be completed [6]. The quality of the requirement engineering process is an important factor that can cause errors in software engineering projects [7].

Failures in system development are often caused by misunderstandings that are misinterpreted when the RE stage cannot meet user expectations. The better the requirements specification given, the better the software system developed [8]. Therefore, a RE process is needed that is able to solve these problems. RE is often referred to as the most important phase in software engineering because errors in this phase are very expensive if not detected at a later stage [9].

Currently, RE not only on key stakeholders but also on large-scale data, which comes from a number of operations and customer feedback [10]. For example, user reviews on mobile apps platform becomes an important target of analysis, because it contains a lot of information. However, data-driven requirements technology is a new domain that must be continuously researched and developed [11].

According to the data, data management and analytics will grow rapidly according to their role. The compound annual growth rate is 21%, twice as fast as the business software. As data grows rapidly, decision makers and stakeholders are demanding computerized support for their work by asking for intelligent solutions that can analyze and visualize their data to achieve their goals [12]. To meet these needs, it is necessary to enhance RE elicitation process that utilizes available data as the main source for determining software requirement [5]. This study uses Natural Language Processing (NLP) combined with the Kmeans and MeanShift algorithms to find out whether there are significant differences in the data used and in the end these data can be used in the need elicitation process [13].

Materials and Methods

Test the effectiveness of different concepts and combinations of POS tagging and clustering techniques in automatically managing large amounts of user input. Ten participants, consisting of five students and five lecturers, were asked to evaluate pairs of existing feedback data [13]. The selected pairs were designed to test all variables equally. The participants' familiarity with the domain and their level of interest were considered, and four combinations of POS tagging were chosen for testing. The supervised and unsupervised clustering algorithms were used to partition the data after extracting the POS tagging combinations, and the performance differences between these approaches were also examined [14]. After the experiment, participants were interviewed to gather their motivations for choosing specific items. The evaluation metrics focused on intra-cluster relatedness, measuring the similarity of items within the same cluster, and inter-cluster relatedness, assessing the differences between items in different clusters. These metrics were utilized to evaluate the effectiveness of different NLP settings and techniques in generating groupings [15 ,16].

In Fig. 1, the research used is university case studies and will eventually produce a dashboard in the form of the distribution of existing data sets. To achieve this goal, a process is followed where 20 feedback items are extracted from the dataset. These items are then transformed using POS tagging, which involves utilizing NLP tools and domain-specific dictionaries. The POS tagging helps in identifying and categorizing different parts of speech in the text.

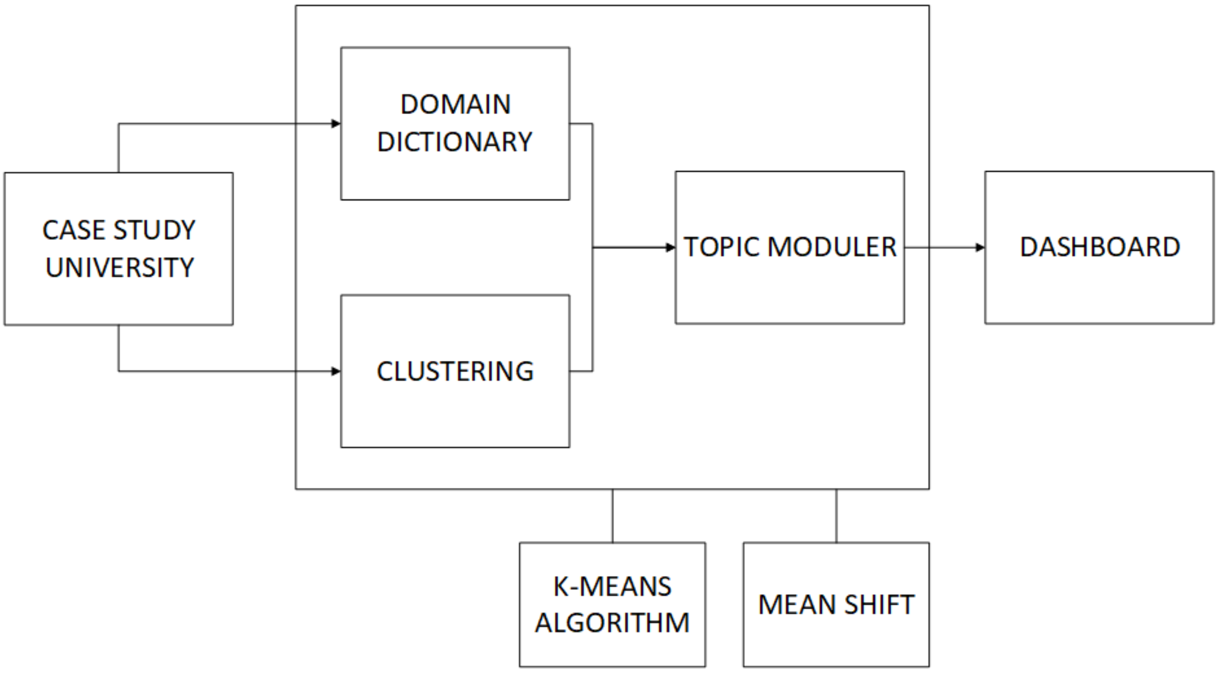

Figure 1. Component Diagram Architecture

After the text transformation, two clustering algorithms, namely K-means and Meanshift, are applied to assign the transformed text to cluster [17]. These clustering algorithms group similar feedback items together based on their textual features[18].

In addition to clustering, topic modeling is also performed to identify topics within each cluster. Topic modeling helps in uncovering underlying themes or subjects that are prevalent within the feedback data [19].

Overall, the analysis focuses on automatically organizing the unstructured text data by applying POS tagging, clustering algorithms (K-means and MeanShift), and topic modeling [20]. These techniques aim to provide insights into the structure and content of the feedback data without relying on human-generated classifications.

Results and Discussion

In Fig. 2, the overall evaluation of the pairs based on the assessments from the test subjects, it is observed that 48% of the pairs are rated as "Not at all Associated." When including the category of "Somewhat Unrelated," this percentage increases to 68%, indicating a significant portion of the data being considered as not associated based on evaluating clustering algorithm. It is crucial to consider these findings since the effectiveness of our metrics relies on accurately grouping or separating the feedback items. The presence of a substantial number of unrelated pairs affects the results and may indicate that our grouping approach is successful in identifying and handling unrelated tickets, particularly when more than two components of a pair are unrelated.

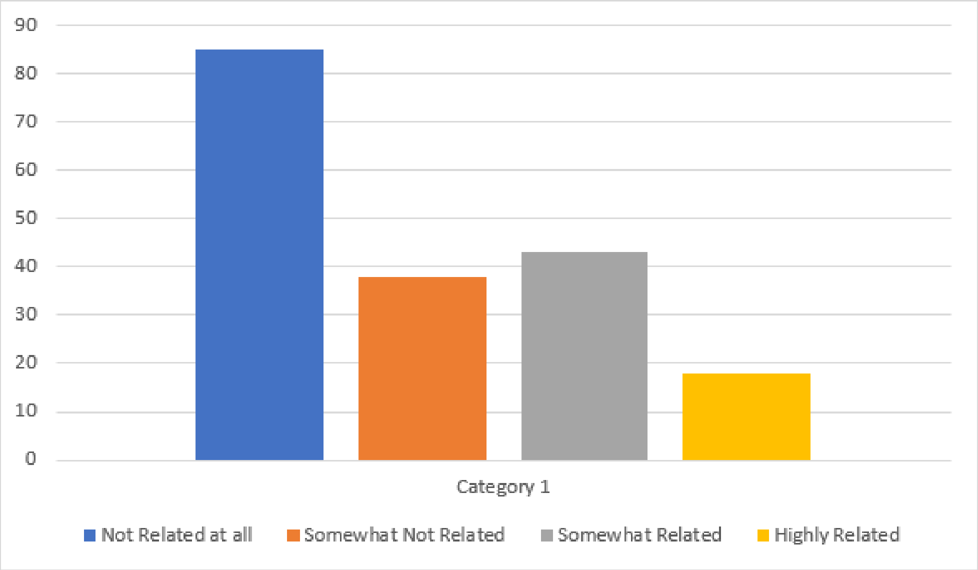

Figure 2. Distribution of Ratings Across Data Set

The Wilcoxon Total Rating test was conducted on the distribution of ratings, comparing pairs that were in the same cluster versus pairs in different clusters. The Wilcoxon Rank Sum Test results in N=639992, W=397301, and a p-value of less than 2.2e-15, indicating a statistically significant difference between these two groups. This suggests that there is a significant distinction between the ratings of paired items within the same cluster and those in different clusters.

Further analysis was performed by dividing the data into pairs within the same cluster and pairs in different clusters, and conducting the Wilcoxon Rank Sum Test on the difference in ratings between students and professionals. For pairs in the same cluster, N=160931, W=86389, and the p-value is 0.06285. For pairs in different clusters, N=158531, W=72621, and the p-value is 0.02135. The W-score represents the sum of the ratings, indicating how many ratings are larger in one population compared to the other. Although no statistical significance was found, it was observed that students and professionals tended to agree more on ratings for pairs within the same cluster and had more disagreements on ratings for pairs in different clusters.

Overall, these findings suggest a tendency for higher agreement between students and professionals when evaluating pairs within the same cluster and lower agreement for pairs in different clusters, although the differences did not reach statistical significance

In Fig.3, the lecturer rate tickets in the same cluster. Three combinations the K-Means is related by more than 51%, whereas Meanshift has only one distribution which violates the 50%. However, all distributions are tied between 42% and 56%. When examining the results for lecturers the saw pairs clustered in the same cluster and different clusters separately, because interested in the performance of the POS clustering and tagging algorithms regarding ticket grouping and segregation.

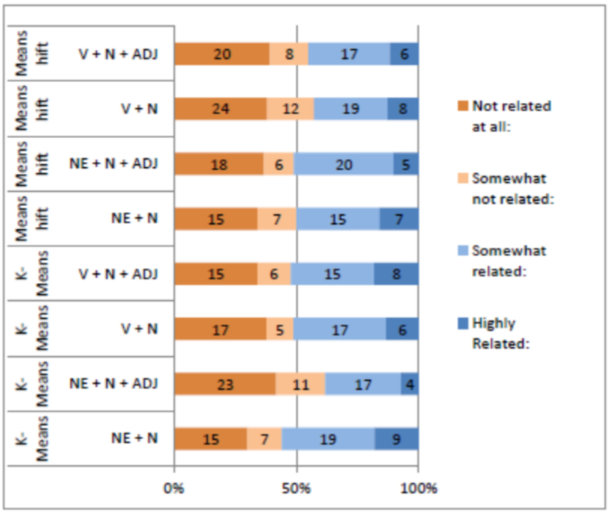

Figure 3. Distribution of Ratings Across Data Set

In Fig.4, the distribution of pairs in a different cluster is analyzed to evaluate the success of ticket separation. A high proportion of unrelated pairs indicates a successful distribution. The most effective combination is the Meanshift algorithm with the Verbs and Nouns feature set, achieving a separation of unrelated tickets in 92% of cases. Conversely, the worst performing combination is the K-means algorithm with Named Entities, Nouns, and Adjectives, achieving a separation of unrelated tickets in only 80% of cases. Therefore, the MeanShift algorithm with the Verbs and Nouns combination is the most successful approach for accurately separating tickets into different clusters.

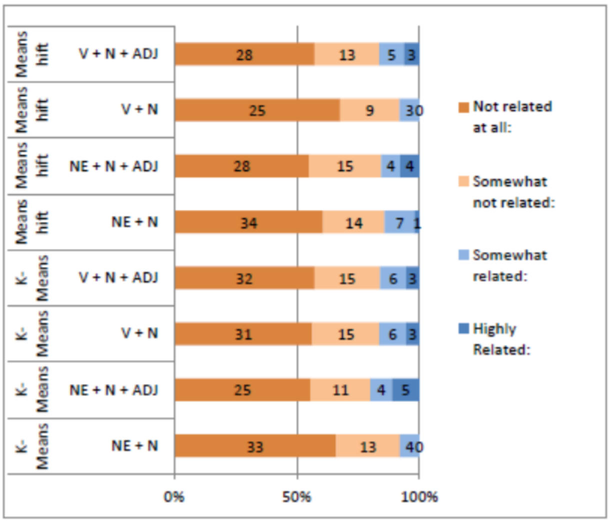

Figure 4. Lecturer For Pair in The Different Cluster Rating

Conclusion

In the evaluation of the pairs based on test subjects' assessments, it was found that a significant percentage of the data, 68%, was considered unrelated or not associated. This finding is important as it affects the accuracy of the metrics used to group or separate feedback items. A statistical test showed a significant difference in ratings between pairs within the same cluster and pairs in different clusters. However, when comparing ratings given by students and professionals, there was no statistical significance, but a tendency for higher agreement within the same cluster and lower agreement in different clusters was observed.

In line with the shift towards a data-centered approach, there is a need for automated tools that can process feedback. These tools could build on the findings of this thesis, particularly in the area of directed clustering, which focuses on words relevant to requirements analysts in software development.

The Wilcoxon Sum Rating Test was conducted separately on pairs within the same cluster and pairs in different clusters, considering the ratings given by professionals and students. For pairs within the same cluster, N=40159, W=20539, and the p-value is 0.6773. For pairs in different clusters, N=39759, W=20135, and the p-value is 0.803. These results suggest that the clustering algorithm has a slightly more noticeable effect on the ratings given by professionals when evaluating pairs within the same cluster compared to pairs in different clusters. The distribution of ratings appears to be less uniform in the former category.

The Kruskal-Wallis test was performed to assess the influence of the four distributions of POS tagging on the ratings given by lecturers. The test was conducted separately for pairs within the same cluster and pairs in different clusters. For pairs within the same cluster, the test yielded a chi-squared value of 2.6496, df=3, and a p-value of 0.4488.

The clustering algorithm appears to have a slightly stronger influence on the ratings given by lecturers for pairs within the same cluster, while the effect of POS tagging on the ratings given by lecturers is similar for pairs within the same cluster and pairs in different clusters.

Acknowledgment

The author thanks the Universiti Teknologi Malaysia, for providing the laboratory to complete this study.

Authors' Declaration

- Conflicts of Interest: None.
- We hereby confirm that all the Figures and Tables in the manuscript are ours. Furthermore, any Figures and images, that are not ours, have been included with the necessary permission for re-publication, which is attached to the manuscript.
- Ethical Clearance: The project was approved by the local ethical committee in Universiti Teknologi Malaysia.

Authors' Contribution Statement

D.Y.B and D.N.J conceived of the presented idea. D.Y.B, D.N.J., and F.A. development theory and performed computation. D.N.J and S.A.H. verified the analytical methods. F.A. encouraged D.Y.B. to investigate and supervised finding of this work. All author discussed the result and contributed to the final manuscript.

References

1. Shafiq M, Zhang Q, Akbar MA, Khan AA, Hussain S, Fazal-E-Amin, et al. Effect of Project Management in Requirements Engineering and Requirements Change Management Processes for Global Software Development. IEEE Access. 2018;6(May 2018):25747–63. http://dx.doi.org/10.1109/ACCESS.2018.2834473.
2. Mavin A, Mavin S, Penzenstadler B, Venters CC. Towards an ontology of requirements engineering approaches. Proc IEEE Int Conf Requir Eng. 2019;2019-Septe:514–5. http://dx.doi.org/10.1109/RE.2019.00080.
3. Andry JF, Hadiyanto, Gunawan V. Intelligent Decision Support System for Supply Chain Risk Management Process (SCRMP) with COBIT 5 in Furniture Industry. Int J Adv Sci Eng Inf Technol. 2023;13(2):736–43. http://dx.doi.org/10.18517/ijaseit.13.2.17359.
4. Hemmati A, Al Alam SMD, Carlson C. Utilizing product usage data for requirements evaluation. Proc - 2018 IEEE 26th Int Requir Eng Conf RE 2018. 2018;432–5. http://dx.doi.org/10.1109/RE.2018.00056.
5. Adetoba Bolaji T, Ogundele Israel O. Requirements Engineering Techniques in Software Development Life-Cycle Methods : A Systematice Literature Review. Int J Adv Res Comput Eng Technol. 2018;7(10):733–43.
6. Andry JF, Hadiyanto, Gunawan V. Critical Factors of Supply Chain Based on Structural Equation Modelling for Industry 4.0. J Eur des Systèmes Autom. 2023;56(2):187–94. http://dx.doi.org/10.18280/jesa.560202.
7. Petersen P, Stage H, Langner J, Ries L, Rigoll P, Philipp Hohl C, et al. Towards a Data Engineering Process in Data-Driven Systems Engineering. ISSE 2022 - 2022 8th IEEE Int Symp Syst Eng Conf Proc. 2022;1–8. http://dx.doi.org/10.1109/ISSE54508.2022.10005441.
8. Kourla SR, Putti E, Maleki M. REBD: A Conceptual Framework for Big Data Requirements Engineering. 2020;(2018):79–87.AIRCC. http://dx.doi.org/10.5121/csit.2020.100608.
9. Berry DM. The requirements engineering reference model: A fundamental impediment to using formal methods in software systems development. Proc - 2019 IEEE 27th Int Requir Eng Conf Work REW 2019. 2019;17(3):109. http://dx.doi.org/10.1109/REW.2019.00024.
10. Hansch G, Schneider P, Brost GS. Deriving impact-driven security requirements and monitoring measures for industrial IoT. CPSS 2019 - Proc 5th ACM Cyber-Physical Syst Secur Work co-located with AsiaCCS 2019. 2019;37–45. http://dx.doi.org/10.1145/3327961.3329528.
11. Juneja P, Kaur P. Software Engineering for Big Data Application Development: Systematic Literature Survey Using Snowballing. 2019 Int Conf Comput Power Commun Technol GUCON 2019. 2019;492–6. https://ieeexplore.ieee.org/document/8940574.
12. Guzmán L, Oriol M, Rodríguez P, Franch X, Jedlitschka A, Oivo M. How can quality awareness support rapid software development? - A research preview. Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics). 2017;10153 LNCS:167–73. http://dx.doi.org/10.1007/978-3-319-54045-0_12.
13. Altarturi HH, Ng KY, Ninggal MIH, Nazri ASA,

Ghani AAA. A requirement engineering model for big data software. 2017 IEEE Conf Big Data Anal ICBDA 2017. 2018;2018-Janua(November):111–7. 1 http://dx.doi.org/0.1109/ICBDAA.2017.8284116.

14. D'Aloisio G. Quality-Driven Machine Learning-based Data Science Pipeline Realization: a software engineering approach. Proc - Int Conf Softw Eng. 2022;291–3. http://dx.doi.org/10.1109/ICSE-Companion55297.2022.9793779.

15. Andry JF, Sibaran R, Yefta VN. Analysis of Big Data Football Club Market Value Using K-Means and Linear Regression Mining Methods. J Comput Sci. 2023;19(2):286–94. http://dx.doi.org/10.3844/JCSSP.2023.286.294.

16. Zhao L, Alhoshan W, Ferrari A, Letsholo KJ, Ajagbe MA, Chioasca EV, et al. Natural Language Processing for Requirements Engineering. ACM Comput Surv. 2021;54(3):1–41. http://dx.doi.org/10.1145/3444689.

17. Wang Q, Du W, Ma C, Gu Z. Gradient Color Leaf Image Segmentation Algorithm Based on Meanshift and Kmeans. IEEE Adv Inf Technol Electron Autom Control Conf. 2021;2021:1609–14. http://dx.doi.org/10.1145/3444689.

18. Allala SC, Sotomayor JP, Santiago D, King TM, Clarke PJ. Generating Abstract Test Cases from User Requirements using MDSE and NLP. IEEE Int Conf Softw Qual Reliab Secur QRS. 2022;2022-Decem:744–53. http://dx.doi.org/10.1109/QRS57517.2022.00080

19. Muhajir M. MyBotS Prototype on Social Media Discord with NLP Imam Al Maksur Abstract : Introduction : Materials and Methods. Baghdad Sci. J. 2021;18(1):753–63. https://bsj.uobaghdad.edu.iq/index.php/BSJ/article/view/5954.

20. Patwary MKH, Haque MM. A semi-supervised machine learning approach using K-means algorithm to prevent burst header packet flooding attack in optical burst switching network. Baghdad Sci J. 2019;16(3):804–15. http://dx.doi.org/10.21123/bsj.2019.16.3(Suppl.).0804.

معالجة اللغة الطبيعية لاستنتاج المتطلبات في الجامعة باستخدام خوارزمية KMEANS وMEANSHIFT

ديفي يوريسكا برناندا[1,2]، دايانغ إن.إيه جواوي[1]، شهليزا عبد الحليم[1]، فرانسيسكوس أديكارا[3]

[1]قسم علوم الحاسب، كلية الهندسة، الجامعة التكنولوجية الماليزية، جوهور باهرو، ماليزيا.
[2]قسم نظم المعلومات، كلية التكنولوجيا والتصميم، جامعة بوندا موليا، جاكرتا، إندونيسيا.
[3]قسم المعلوماتية، كلية التكنولوجيا والتصميم، جامعة بوندا موليا، جاكرتا، إندونيسيا.

الخلاصة

تمثل هندسة المتطلبات المبنية على البيانات DDRE رؤية للتحول من الأساليب التقليدية الثابتة للقيام بهندسة المتطلبات إلى الأساليب الديناميكية التي تعتمد على البيانات والتي تركز على المستخدم. البيانات المتاحة والمتطلبات المتزايدة التعقيد لبرامج النظام التي يمكن أن تتكيف وظائفها مع الاحتياجات المتغيرة لكسب ثقة مستخدميها، تحتاج إلى نهج في عملية هندسة البرمجيات المستمرة. هذه الحاجة تدفع إلى ظهور تحديات جديدة في مجال هندسة المتطلبات لمواجهة التغييرات المطلوبة. كانت المشكلة في هذه الدراسة هي طريقة تناقضات البيانات التي أدت إلى إعاقة عملية استنباط الاحتياجات وفي نهاية تطوير البرمجيات وجدت تناقضات لا يمكن أن تلبي احتياجات أصحاب المصلحة وأهداف المنظمة. يهدف البحث إلى جمع ودمج البيانات من مصادر متعددة وضمان قابلية التشغيل البيني. الاستنتاج في هذا البحث هو أن خوارزمية التجميع تساعد في جمع البيانات وعملية الاستنباط لها تأثير أكبر إلى حد ما على التقييمات المقدمة من قبل المتخصصين للأزواج التي تنتمي إلى نفس المجموعة. ومع ذلك، فإن تأثير (POS tagging) على التقييمات التي يقدمها المحترفون يكون متسقًا نسبيًا بالنسبة للأزواج داخل نفس المجموعة والأزواج في مجموعات مختلفة.

الكلمات المفتاحية: DDRE، مصدر البيانات، هندسة المتطلبات، برمجيات النظام، عملية الاستنباط.