A Survey of Text Matching Techniques

Awatif Alqahtani
College of Computer and Information Systems
Umm Al-Qura University
Makkah, Saudi Arabia
s43980248@st.uqu.edu.sa

Hosam Alhakami
College of Computer and Information Systems
Umm Al-Qura University
Makkah, Saudi Arabia
hhhakam@uqu.edu.sa

Tahani Alsubait
College of Computer and Information Systems
Umm Al-Qura University
Makkah, Saudi Arabia
tmsubait@uqu.edu.sa

Abdullah Baz
College of Computer and Information Systems
Umm Al-Qura University
Makkah, Saudi Arabia
aobaz01@uqu.edu.sa

Abstract—Text matching is the process of identifying and locating particular text matches in raw data. Text matching is a vital component in practical applications and an essential process in several fields. Furthermore, several dynamic techniques have been introduced in this context in order to create ease in pattern generation from words. The process involves matching of text files, text mining, text clustering, association rule extraction, world cloud, natural language processing, and text similarity measures (knowledge-based, corpus-based, string-based, and hybrid similarities). The string-based approach forms the most conspicuous form of text mining applied in different cases. The survey attempted in the present study covers a new research premise that uses text-matching to solve problems. The study also summarizes different approaches that are being used in this domain.

Keywords-text mining; similarity measure; matching; clustering; natural language processing; word cloud

I. INTRODUCTION

Nowadays, computer scientists are enjoying the convenience of many advanced tools and technologies to accomplish various tasks and duties, and pattern matching is no exception. Text matching techniques are essential components in practical applications in various fields. Pattern matching refers to the process of checking a specific order or structure of a token for the existence of the elements of a pattern. The matching is usually exact: the result is either a match or not. Even though there are various ways of memorizing data, the text remains the primary form of exchanging information. Before text is searched for patterns, it has to be structured by using various strategies such as text matching. As a result, text matching is the process of identifying and locating particular text matches in raw data. Though there are many methods of applying this process, addressing inherent problems in these methods is needed to boost the efficiency of the process. In this regard, this paper focuses on Natural Language Processing (NLP) algorithms and unsupervised text mining techniques, their inherent challenges, and ways to overcome them. Many approaches have been proposed for Text Matching. Most are mainly based on text mining which involves keyword matching, sequence matching and clustering. Moreover, in matching texts, the similarity measure is used by measuring the similarity between words, sentences, paragraphs and documents. Measuring of similarity is considered to be an important component in various tasks such as text summarizing, text matching, information retrieval, automatic essay scoring, document clustering, and machine translation [1].

II. MATCHING TEXT FILE METHODS

A. Text Mining

Text mining refers to the process of analyzing and exploring large sets of unstructured data to identify and isolate topics, keywords, concepts, patterns and other elements evident in data sets. The process is becoming more practical due to the emergence of large data platforms and various deep learning algorithms with the ability to analyze big sets of unstructured data [2]. Through text mining, it is possible to focus on text data instead of other more structured forms of data examined in data mining. However, both the processes involve organizing and structuring the data in a particular fashion before subjecting it to qualitative and quantitative analysis. Text mining comprises of a wide set of algorithms and topics for analyzing texts and spanning different communities [3]. Some of the common text mining tools include text clustering, text visualization, and association rule extraction. Text clustering rests on the cluster hypothesis that argues that clustering is possible when the relevant documents possess more similarities with each other than with non-relevant documents. It is a trustworthy technique that is largely applied in analyzing and examining big sets of data akin to data mining [4]. Moreover, it is proven that text clustering remains an effective tool for theme analysis of texts, while it also facilitates topic analysis technique where named parties with concurrent occurrences are categorized together and the frequent entity is placed in various sets by using the graph-based approach [5]. Currently, the

Corresponding author: Abdullah Baz

procedure of topic ranking in a dynamic set of data is gaining the attention of researchers using text clustering in the digital field [6, 7]. Even though text clustering remains vital in browsing and navigation processes, various text clustering strategies fail to address challenges such as space complexity and high time, less robustness, privacy risks, inability to comprehend contextual elements of words, etc.

Authors in [8] propose an efficient text-based clustering framework. They determined the similarities between words by using the cosine similarity. The data vector was designed from the component similarities and was later used to compute the clustering particles. Following mutation to maximize clustering, the framework is examined by using Mean Square Error (MSE), Processing Time (PT), and Peak Signal to Noise Ratio (PSNR) [9]. The findings indicate that the proposed framework produces optimal PSNR, MSE and PT compared to the Pair-wise Random Swap (PRS) and Fuzzy CC Means (FCM) approaches. Authors in [10] reported that another way of overcoming challenges to text mining such as data dimension reduction and extraction is by using a deep learning vocabulary network [10]. The vocabulary network's design is based on closely related word sets that contain the concurrence relations of terms or words. Here, the frequency in feature vectors is replaced with the significance of words in terms of page rank and vocabulary network to create exact feature vectors that depict the meanings of text clustering. When the deep-learning vocabulary network was compared to other approaches such as representative algorithms, the results demonstrate that feature vectors determined by using deep learning vocabulary network portray improved clustering performance. Furthermore, in addressing the diversity and complexity of natural language, the semantically responsive text clustering approach has been proposed in [11]. The authors suggest an extension-oriented modeling approach along with a similarity aggregation method and a space construction method. This approach has the ability to improve semantic sensitivity during the text mining process by organizing the generated clusters into different granularities [12]. The effectiveness of this approach is confirmed through recognized examinations of clustering data-sets and algorithms. Authors in [13] collected and textually analyzed various documents from six scientific databases. Text clustering was shown to be effective in matching texts from natural language documents. In cases where similarity was not detected, they attested it to ambiguity or interrelations between analyzed text documents. Cosine similarity is still the most common metric of similarity. The collection of sentences is carried out by selecting sentences from each cluster on of basis of term frequency and inverse term frequency rankings in this same cluster [14].

B. Association Rule Extraction

Approaches associated with association rule extraction focus on identifying the relationships in a big set of variables in data-sets. Association-Rule Mining (ARM) strategies identify the variable-value arrangements that occur repeatedly. It is also called knowledge discovery because it is likened to correlations analysis that focuses on the associations between two variables [15]. The association procedures for text mining focus on exploring the relationships among various factual notions or topics used to describe a body of data. The primary objective is to discover the vital association rules comparative to a body of data in such a manner that the existence of particular topics in articles may resemble the existence of other topics. Some researchers agree that association rule mining is a prominent research field that used to unearth frequent patterns and arrangements in repositories of synthetic or real-world data-sets [16]. Association rule mining underscores that various associations are related or occur among a group of datasets in the database [17]. In many ARM approaches, the rules are evaluated from two main measures, namely confidence and support. Support is the aggregate number of text data that are associated with each other in the database, while confidence is the portion of data-sets that contains associated text data in the database. Here, the researcher employs a prior algorithm to find frequently occurring items over transaction data sets, and the technique is effective in matching similar and dissimilar text data to offer insight for decision-making. Some researchers argue that association rule extraction remains an effective approach of finding accurate and vital knowledge from an assortment of documents.

Today, information technology and the internet are platforms that provide huge amounts of data and information to users [18]. However, a major issue facing text matching is that searching and finding precise and crucial information may be time-consuming or lead to misunderstandings. Creating a method to generate knowledge discovery through the association rule extraction proves to be effective in both saving time and avoiding misunderstandings. Authors in [19] support the use of a temporal approach in extracting negative and positive association rules from texts. Data extraction by using the temporal approach is not popular amongst the ARM researchers, especially for negative associations. However, the researchers insist that the rule allows the researchers to answer essential questions when applying association rules, which boosts the trustworthiness of the resulting association rules for use in matching texts. Traditionally, association analysis is an unsupervised strategy mostly evident in knowledge discovery assignments. Therefore, the majority of the relevant studies focus on simplifying association rules and improving the performance of algorithms. However, other problems arise when association rules are created and applied in various domains. Some of the main challenges the association rules algorithms face include obtaining non-interesting association rules, discovering a huge number of rules, and poor algorithm performance [20]. Troubleshooting for these problems should focus on the objective of an association model and its origin.

C. Word Cloud

Another text mining technique in common use is word cloud. As noted, a word cloud or tag cloud is the visual representation of a word for a particular written content arrangement according to its frequency [21]. It is amongst the most common methods for graphically presenting text data, which makes it vital for examining different kinds of texts such as written opinions or short answers to a questionnaire or survey [22]. It is a preliminary phase during the in-depth analysis of particular text material and the needed information. Nevertheless, this technique faces various challenges. A major

drawback is that word cloud fails to contemplate the linguistic knowhow about text words and corresponding associations with a specific subject when offering a numerical summary to the isolated words [23]. As a result, in the majority of the systems, the word clouds are usually used in a numerical manner to summarize texts and offer little or no chance to correlate the data. Authors in [24] provide the findings of a set of controlled research experiments that demonstrate that layouts where text words are structured into visually and semantically distinctive categories are more efficient and effective for acknowledging underlying topics compared to ordinary world cloud designs. The white space separators and spatially grouped color coding lead to considerably powerful understanding of the fundamental topics as compared to an ordinary word design or layout. As a result, data mining experts are developing data-sets for visually and semantically distinct category identification tasks for use in replicating results for word cloud formats and designs in the future [25]. Authors in [26] agree that word cloud remains an effective text matching tool in writing and reading classes, especially English classes for ESL learners. During the decade, many word cloud applications were designed to provide additional visual appeal to slide shows, posters, and websites amongst others. For writing and reading classes, word clouds are vital for reducing reading time and helping with vocabulary and writing comprehension. Moreover, word clouds are also used for addressing rhetoric elements in instructors' assignment descriptions, syllabuses, and students' essays. The process locates a collection of vocabulary in texts for use in creating word clouds that assist students in understanding and summarizing texts, learning collocations and spellings, finding synonyms, and avoiding repetition, and using word maps for creating tone.

D. Natural Language Processing

NLP is a sub field of artificial intelligence and computer science preoccupied with interactions and associations between human and computer languages. It is also applicable to machine learning algorithms for deciphering texts and speech. It has gained considerable attention in the field of computer science and text matching in particular. Unlike ordinary machine learning roles that deal with structured data, it deals with unstructured text data [27]. Also, one of the types of NLP called sentiment analysis, is a new hybrid approach for the analysis of social problems. It extracts the views from each statement, builds opinion structures of communication, and then decides their social issue alignments [28]. String searching algorithms are the most common NLP tools for use in text matching. In string searching, an algorithm is used to identify and isolate a location where a single or several strings and patters are found in a large body of text data. Some of the most common algorithms are Brute Force, Boyer-Moore Horspool, Rabin-Karp, and Knuth-Morris-Pratt.

Authors in [29] found that a major challenge the pattern matching algorithms face is memory consumption, with the majority of the algorithms focusing on achieving efficiency at the expense of memory use. They offer a new idea to address both by using an algorithm that splits the query configuration into two parts and uses the second part as a query string for use in searching the collection of large text data. The proposed algorithm outperforms the S1 Algorithm on memory use and time efficiency by increasing the length of a query configuration. Authors in [30] proposed Bidirectional (BD) Exact Pattern Matching (EPM) that introduces the novel idea of comparing patterns by using the Selected Text Window (STW) of a text string using a double pointer simultaneously during the searching phase. The researchers identified that during the pre-processing phase, the algorithm boosted the shift decision-making by comparing the mismatched and rightmost character of its Partial Text Window (PTW) to its left at a similar shift length. As a result, it reduced the time complexity of the pre-processing phase and was proved to be more effective than the existing algorithms. The Boyer-Moore-Horspool algorithm increases the efficiency of intrusion detection systems. Many intrusion systems use various string-matching algorithms to carry out searches for mischievous activities, but a major challenge these algorithms face is longer processing time. Authors in [31] found that improving the algorithm by adding a hashing function called HBMH which is vital in reducing the comparison time. As a result, the findings attest that an enhanced algorithm performs faster and lowers processing time. It was also discovered that a variation of the Boyer-Moore algorithm led to other versions with comparative advantages, disadvantages, and performance standards in various conditions. The variants included BMH, BMHS, BMSH2, BMI, and CBM. A BMH algorithm only applies the bad character shift [31]. Regardless of the location of a mismatch, the distance of a shift to the right it calculated by the characters in text strings that are aligned to the last text or string on the pattern strings. Since it removes the concept of good-suffix, BMH is easy to apply [32]. Moreover, more jumps are realized by using this algorithm. Furthermore, in case of a mismatch, it is possible to determine the shift value by using the char value instead of the mismatched character. However, the BMHS algorithm focuses on calculating the bad char function by considering the status of the next character to determine a right off-set. As a result, a major advantage is that the maximum shift is more than the pattern length by one [33]. However, a major issue is that fewer shifts are realized during a mismatch [34-36].

III. TEXT SIMILARITY MEASURES

There are four major categories of text similarities measures. The string-based similarity measure is one of the oldest methods. The main forms of string-based similarity comprise both character-based similarity and token-based similarity functions [36]. The corpus-based similarity follows semantic approaches. The methodology aids in the determination of similarity between two concepts in terms of information from the respective corpora, which is a collection of electronic, spoken, or written text. These methods enshrine some predefined set of sentences coupled with their translation in other dialects, with the intention of matching the input text within the corpus and the final translations [36]. The knowledge-based similarity measures are defined as a series of semantic measures enlisted for information adopted from semantic networks. The aim of such information is geared towards the identification of the extent of similarity within words. Knowledge-based similarity consists of semantic

relatedness and semantic similarity. Hybrid classification similarity measures do not form a distinct group. They are combinations of the previous approaches in an effort to attain their merits [36]. These approaches work on recursive approaches to tackle limitations of the other measures.

IV. LITERATURE REVIEW SUMMARY

The reviewed methods, their advantages and disadvantages are summarized in Table I. Some of the research studies that contributed to solving specific problems of text matching techniques are summarized in Table II.

TABLE I. TEXT MATCHING METHODS

Text matching techniques	Definition	Advantages	Disadvantages
Text mining	The process of analyzing and exploring large sets of unstructured data to identify and isolate topics, keywords, concepts, patterns, and other elements evident in data sets	It is possible to focus on text data instead of other more structured forms of data examined in data mining, but both processes involve first organizing and structuring the data in a particular fashion before subjecting them to qualitative and quantitative analysis	Depends on other methods.
Text clustering	It rests on the hypothesis that clustering is possible when relevant documents possess more similarities with each other compared to non-relevant documents	It is a trustworthy technique that is largely applied in analyzing and examining big sets of data akin to data mining	Less efficient approach
Word cloud	Word cloud or tag cloud is the visual representation of used words arranged according to their frequency	It is simple and the most common method for graphically presenting text data	Fails to contemplate the linguistic knowhow about text words and corresponding associations with a specific subject
Brute force algorithm	BD EPM that introduces a novel idea of comparing patterns by using the STW of a text string using a double pointer simultaneously during the searching phase	Efficient on memory use and time efficiency by increasing the length of a query configuration	Complex to use
Boyer-Moore Horspool algorithm	It increases the efficiency of intrusion detection systems	Easy to apply as it removes the concept of good suffix	Less shift is realized during a mismatch because of the character next to the last character, but not aligned or associated with the last pattern

TABLE II. EXAMPLES OF TEXT MATCHING TECHNIQUES

Ref	Text matching type	Problem	Approach used	Processes	Results	Year
[38]	String	Used for paralogism detection	String matching	N-gram with Dice's similarity coefficient	Generates a percentage of the similarity of the document, but the efficiency of processing time needs improvement	2019
[39]	Semantic	To detect intelligent plagiarism	Topic modeling	Semantic text alignment based on sentence-level topic modeling	Topic modeling is a potential solution for detecting intelligent plagiarism	2016
[40]	Semantic transformation	To solve the semantic gap in text matching	Semantic transformation model	Enhanced text matching method based on the Cycle GAN combined with the transformer network	The semantic matching is improved significantly	2020
[41]	Hybrid Knowledge Neural Network (NN)	Long text brings a big challenge to NN-based text matching approaches due to their complicated structures	Knowledge Enhanced Hybrid NN (KEHNN)	The three channels are processed by a convolutional NN to generate high level features for matching	Shows that KEHNN can significantly outperform the state-of-the-art matching models and particularly improve matching accuracy on pairs with long text	2018
[42]	Knowledge	To automatically answer medical questions online	Knowledge Abstraction Matching (KAM) method	The method consists of frequent segment N-gram mining, medical knowledge abstraction, medical segment matching and answer reretrieval	Generates more quality answers for Melody QA with a significant improvement of question coverage under acceptable accuracy	2019

V. DISCUSSION

The similarity between words can be expressed in lexical similarity terms (similar sequence of characters) or semantic similarity (same meaning) [37]. Lexical similarity is based on string-based tools or algorithms [37, 43]. String-based algorithms for lexical similarity matching fall into two categories [37]. The first category consists of character-based algorithms such as LCS, Damerau Levenshtein, Jaro, and N-gram [43, 44]. The second category consists of term-based algorithms, such as the Cosine similarity, Block Distance, Euclidean Distance, Jaccard similarity, and Matching Coefficient [45]. However, semantic similarity techniques are more rational in finding substantial relationships between texts

[37]. Some of the innovative string-based tools include SimMetrics [45], SimPack, FLAMINGO, and Alignment API [46]. These tools have one major drawback, which is that they do not provide semantic analysis of texts. Semantic analysis is based on Corpus-based and knowledge-based algorithms [43]. Corpus-based similarity matching techniques have gained widespread application. The most popular corpus-based technique is the Latent Semantic Analysis (LSA), which assumes that words with similar meaning occur in similar texts. Although LSA is easy to implement and offers good performance, it has inherent limitations for non-linear situations due to its linear approach and the probabilistic model may fail to match the observed data [47]. An alternative method is to use Latent Semantic Indexing, which overcomes synonymy [48]. Another technique based on the corpus approach is the Explicit Semantic Analysis (ESA), such as the Wikipedia-based technique, which represents texts in the form of high-dimensional vectors [49]. Although ESA has proven useful in handling different contexts, it was found that the algorithm exhibits unexpected behavior [49]. Knowledge-based approaches rely on information from semantic networks to determine the degree of similarity. The most popular knowledge-based tool is the WordNet, which consists of a lexical database of English words interlinked with conceptual relations [50]. Recently, there have been attempts to combine multiple approaches. The SimAll tool combines features of string, knowledge, and corpus similarity approaches [51]. The best tool for semantic analysis is SimAll because it combines lexical and semantic analysis at different levels of granularity [28].

VI. CONCLUSION

Some of the common text mining tools include text clustering, text visualization, and association rule extraction. Even though text clustering remains vital in browsing and navigation processes, various text clustering strategies have disadvantages such as space complexity and high computational time, reduced robustness, privacy risks, and inability to comprehend contextual elements of words. When the deep-learning vocabulary network is compared to other approaches such as the representative algorithms, the results demonstrate that feature vectors determined by using deep learning vocabulary network portray improved clustering performance. Rule mining underscores that various associations are related or occur among a group of data-sets in the database. In string searching, an algorithm is used to identify and isolate a location where a single or several strings and patters are found in a large body of textual data. Some of the most common algorithms include Brute Force, Boyer-Moore Horspool, Rabin-Karp, and Knuth-Morris-Pratt algorithms.

REFERENCES

[1] P. Kudi, A. Manekar, K. Daware, and T. Dhatrak, "Online Examination with short text matching," in *IEEE Global Conference on Wireless Computing Networking*, Lonavala, India, Dec. 2014, pp. 56–60, https://doi.org/10.1109/GCWCN.2014.7030847.

[2] R. Munoz, A. Montoyo, and E. Metais, *Natural Language Processing and Information Systems*. Alicante, Spain: Springer, 2011.

[3] M. Allahyari *et al.*, "A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques," Jul. 2017, Accessed: Dec. 26, 2020. [Online]. Available: http://arxiv.org/abs/1707.02919.

[4] K. B. Cohen and D. Demner-Fushman, *Biomedical natural language processing*. Amsterdam, Netherlands: John Benjamins Publishing Company, 2014.

[5] L. Xinwu, "A new text clustering algorithm based on improved k means," *Journal of Software*, vol. 7, no. 1, pp. 95–101, 2012.

[6] P. D. Asanka, "Finding similar files using text mining," in *8th International Conference on Computer Science Education*, Colombo, Sri Lanka, Apr. 2013, pp. 431–435, https://doi.org/10.1109/ICCSE.2013.6553950.

[7] T. Svadas and J. Jha, "Document Cluster Mining on Text Documents," *International Journal of Computer Science and Mobile Computing*, vol. 4, no. 6, pp. 778–782, Jun. 2015.

[8] M. J. Basha and K. P. Kaliyamurthie, "An Improved Similarity Matching based Clustering Framework for Short and Sentence Level Text," *International Journal of Electrical & Computer Engineering*, vol. 7, no. 1, pp. 551–558, 2017.

[9] M. Mateen, J. Wen, M. Hassan, and S. Song, "Text Clustering using Ensemble Clustering Technique," *International Journal of Advanced Computer Science and Applications*, vol. 9, no. 9, pp. 185–190, 2018, https://doi.org/10.14569/IJACSA.2018.090925.

[10] J. Yi, Y. Zhang, X. Zhao, and J. Wan, "A Novel Text Clustering Approach Using Deep-Learning Vocabulary Network," *Mathematical Problems in Engineering*, vol. 2017, Jan. 2017, Art. no. 8310934, https://doi.org/10.1155/2017/8310934.

[11] Y. Liu, M. Liu, and X. Wang, "Towards Semantically Sensitive Text Clustering: A Feature Space Modeling Technology Based on Dimension Extension," *PLOS ONE*, vol. 10, no. 3, pp. 1–18, 2015.

[12] D. Westergaard, H. H. Stærfeldt, C. Tonsberg, L. J. Jensen, and S. Brunak, "A comprehensive and quantitative comparison of text-mining in 15 million full-text articles versus their corresponding abstracts," *P L o S Computational Biology*, vol. 14, no. 2, 2018, Art. no. e1005962, https://doi.org/10.1371/journal.pcbi.1005962.

[13] S. A. Salloum, M. Al-Emran, A. A. Monem, and K. Shaalan, "Using Text Mining Techniques for Extracting Information from Research Articles," in *Intelligent Natural Language Processing: Trends and Applications*, K. Shaalan, A. E. Hassanien, and F. Tolba, Eds. Cham: Springer International Publishing, 2018, pp. 373–397.

[14] M. S. Bewoor and S. H. Patil, "Empirical Analysis of Single and Multi Document Summarization using Clustering Algorithms," *Engineering, Technology & Applied Science Research*, vol. 8, no. 1, pp. 2562–2567, Feb. 2018, https://doi.org/10.48084/etasr.1775.

[15] M. Kulkarni and S. Kulkarni, "Knowledge discovery in text mining using association rule extraction," *International Journal of Computer Applications*, vol. 143, no. 12, pp. 30–35, 2016.

[16] J. Manimaran and T. Velmurugan, "A survey of association rule mining in text applications," in *IEEE International Conference on Computational Intelligence and Computing Research*, Enathi, India, Dec. 2013, pp. 1–5, https://doi.org/10.1109/ICCIC.2013.6724258.

[17] A. A. Oliinyk and S. A. Subbotin, "A stochastic approach for association rule extraction," *Pattern Recognition and Image Analysis*, vol. 26, no. 2, pp. 419–426, Apr. 2016, https://doi.org/10.1134/S1054661816020139.

[18] S. Mahmood, M. Shahbaz, and A. Guergachi, "Negative and positive association rules mining from text using frequent and infrequent itemsets," *The Scientific World Journal*, vol. 2014, May 2014, Art. no. 973750, https://doi.org/10.1155/2014/973750.

[19] M. N. Moreno, S. Segrera, and V. F. López, "Association Rules: Problems, solutions and new applications," in *Actas del III Taller Nacional de Minería de Datos y Aprendizaje, TAMIDA2005*, 2005, pp. 317–323.

[20] R. Atenstaedt, "Word cloud analysis of the BJGP: 5 years on," *British Journal of General Practice*, vol. 67, no. 658, pp. 231–232, May 2017, https://doi.org/10.3399/bjgp17X690833.

[21] R. Atenstaedt, "Word cloud analysis of the BJGP," *British Journal of General Practice*, vol. 62, no. 596, pp. 148–148, Mar. 2012, https://doi.org/10.3399/bjgp12X630142.

[22] C. N. Hofer and G. Karagiannis, "Cloud computing services: taxonomy and comparison," *Journal of Internet Services and Applications*, vol. 2, no. 2, pp. 81–94, 2011, https://doi.org/10.1007/s13174-011-0027-x.

[23] M. A. Hearst, E. Pedersen, L. Patil, E. Lee, P. Laskowski, and S. Franconeri, "An Evaluation of Semantically Grouped Word Cloud Designs," *IEEE Transactions on Visualization and Computer Graphics*, vol. 26, no. 9, pp. 2748–2761, Sep. 2020, https://doi.org/10.1109/TVCG.2019.2904683.

[24] A. S. Tuchkova and P. P. Kondrasheva, "The term 'data mining'. tasks solved by data mining methods," *Trends in the Development of Science and Education*, vol. 5, no. 2, pp. 27–30, 2019.

[25] O. Filatova, "More Than a Word Cloud," *TESOL Journal*, vol. 7, no. 2, pp. 438–448, 2016, https://doi.org/10.1002/tesj.251.

[26] M. Nagao, "Special Issue: 'Collection of Best Annual Papers' Organized for the 20th Anniversary of the Association for Natural Language Processing," *Journal of Natural Language Processing*, vol. 21, no. 4, pp. 617–618, 2014, https://doi.org/10.5715/jnlp.21.617.

[27] S. Hakak, A. Kamsin, P. Shivakumara, M. Y. I. Idris, and G. A. Gilkar, "A new split based searching for exact pattern matching for natural texts," *PLOS ONE*, vol. 13, no. 7, 2018, Art. no. e0200912, https://doi.org/10.1371/journal.pone.0200912.

[28] M. Madhukar and S. Verma, "Hybrid Semantic Analysis of Tweets: A Case Study of Tweets on Girl-Child in India," *Engineering, Technology & Applied Science Research*, vol. 7, no. 5, pp. 2014–2016, Oct. 2017, https://doi.org/10.48084/etasr.1246.

[29] T. H. Nguyen, "A new approach to exact pattern matching," *Journal of Computer Science and Cybernetics*, vol. 35, no. 3, pp. 197–216, Aug. 2019, https://doi.org/10.15625/1813-9663/35/3/13620.

[30] C. C. Hoong and M. A. Ameedeen, "Boyer-Moore Horspool Algorithm Used in Content Management System of Data Fast Searching," *Advanced Science Letters*, vol. 23, no. 11, pp. 11387–11390, Nov. 2017, https://doi.org/10.1166/asl.2017.10289.

[31] S. Sharma and M. Dixit, "Single Digit Hash Boyer Moore Horspool Pattern Matching Algorithm for Intrusion Detection System," *International Journal of Future Generation Communication and Networking*, vol. 9, no. 9, pp. 169–180, 2016.

[32] Y. Jeong, N.-P. Tran, M. Lee, D. Nam, J.-S. Kim, and S. Hwang, "Parallelization and Performance Optimization of the Boyer-Moore Algorithm on GPU," *KIISE Transactions on Computing Practices*, vol. 21, no. 2, pp. 138–143, 2015, https://doi.org/10.5626/KTCP.2015.21.2.138.

[33] R. Janani and S. Vijayarani, "An efficient text pattern matching algorithm for retrieving information from desktop," *Indian Journal of Science and Technology*, vol. 9, no. 43, pp. 1–11, 2016.

[34] M. O. Kulekci, "Tara: An algorithm for fast searching of multiple patterns on text files," in *22nd international symposium on computer and information sciences*, Ankara, Turkey, Nov. 2007, pp. 1–6, https://doi.org/10.1109/ISCIS.2007.4456850.

[35] A. Weyer, "The Brute Force Algorithm," Ph.D. dissertation, Bowling Green State University, United States, 2019.

[36] P. Kuipers, "Empowerment in community-based rehabilitation and disability-inclusive development," *Disability, CBR & Inclusive Development*, vol. 24, no. 4, pp. 24–42, 2013.

[37] D. D. Prasetya, A. P. Wibawa, and T. Hirashima, "The performance of text similarity algorithms," *International Journal of Advances in Intelligent Informatics*, vol. 4, no. 1, pp. 63–69, Mar. 2018, https://doi.org/10.26555/ijain.v4i1.152.

[38] W. G. S. Parwita, I. G. A. A. D. Indradewi, and I. N. S. W. Wijaya, "String Matching based Plagiarism Detection for Document in Bahasa Indonesia," in *5th International Conference on New Media Studies*, Bali, Indonesia, Oct. 2019, pp. 54–58, https://doi.org/10.1109/CONMEDIA46929.2019.8981821.

[39] H. T. Le, L. N. Pham, D. D. Nguyen, S. V. Nguyen, and A. N. Nguyen, "Semantic text alignment based on topic modeling," in *IEEE RIVF International Conference on Computing Communication Technologies, Research, Innovation, and Vision for the Future*, Hanoi, Vietnam, Nov. 2016, pp. 67–72, https://doi.org/10.1109/RIVF.2016.7800271.

[40] S. Zhang, H. Tan, L. Chen, and B. Lv, "Enhanced Text Matching Based on Semantic Transformation," *IEEE Access*, vol. 8, pp. 30897–30904, 2020, https://doi.org/10.1109/ACCESS.2020.2973206.

[41] Y. Wu, W. Wu, Z. Li, and M. Zhou, "Knowledge Enhanced Hybrid Neural Network for Text Matching," Nov. 2016, Accessed: Dec. 26, 2020. [Online]. Available: http://arxiv.org/abs/1611.04684.

[42] J. Chen, J. Zhou, Z. Shi, B. Fan, and C. Luo, "Knowledge Abstraction Matching for Medical Question Answering," in *IEEE International Conference on Bioinformatics and Biomedicine*, San Diego, USA, Nov. 2019, pp. 342–347, https://doi.org/10.1109/BIBM47256.2019.8982973.

[43] M. M. Mironczuk and J. Protasiewicz, "A recent overview of the state-of-the-art elements of text classification," *Expert Systems with Applications*, vol. 106, pp. 36–54, Sep. 2018, https://doi.org/10.1016/j.eswa.2018.03.058.

[44] B. Liu, Y. Zhou, and W. Sun, "Character-level text classification via convolutional neural network and gated recurrent unit," *International Journal of Machine Learning and Cybernetics*, vol. 11, no. 8, pp. 1939–1949, Aug. 2020, https://doi.org/10.1007/s13042-020-01084-9.

[45] M. Oghbaie and M. Mohammadi Zanjireh, "Pairwise document similarity measure based on present term set," *Journal of Big Data*, vol. 5, no. 1, p. 52, Dec. 2018, https://doi.org/10.1186/s40537-018-0163-2.

[46] Z. Yousefi, H. Sotudeh, M. Mirzabeigi, S. M. Fakhrahmad, A. Nikseresht, and M. Mohammadi, "Investigating text power in predicting semantic similarity," *International Journal of Information Science and Management*, vol. 17, no. 1, p. 17, Jan. 2019.

[47] J. Guan, A. S. Levitan, and S. Goyal, "Text Mining Using Latent Semantic Analysis: An Illustration through Examination of 30 Years of Research at JIS," *Journal of Information Systems*, vol. 32, no. 1, pp. 67–86, Oct. 2016, https://doi.org/10.2308/isys-51625.

[48] K. Al-Sabahi, Z. Zhang, J. Long, and K. Alwesabi, "An Enhanced Latent Semantic Analysis Approach for Arabic Document Summarization," *Arabian Journal for Science and Engineering*, vol. 43, no. 12, pp. 8079–8094, Dec. 2018, https://doi.org/10.1007/s13369-018-3286-z.

[49] Z. Wu *et al.*, "An efficient Wikipedia semantic matching approach to text document classification," *Information Sciences*, vol. 393, pp. 15–28, Jul. 2017, https://doi.org/10.1016/j.ins.2017.02.009.

[50] K. Orkphol and W. Yang, "Word Sense Disambiguation Using Cosine Similarity Collaborates with Word2vec and WordNet," *Future Internet*, vol. 11, no. 5, May 2019, Art. no. 114, https://doi.org/10.3390/fi11050114.

[51] W. H. Gomaa and A. A. Fahmy, "Simall: A flexible tool for text similarity," in The Seventeenth Conference on Language Engineering ESOLEC, vol. 17, pp. 122–127, 2017.