ISSN: 0067-2904

Inspecting Hybrid Data Mining Approaches in Decision Support Systems for Humanities Texts Criticism

Baraa Hasan Hadi*, Tareef Kamil Mustafa
Department of Computers, College of Science, University of Baghdad, Baghdad, Iraq

Received:16/7/2020 Accepted: 7/1/2021

Abstract

The majority of systems dealing with natural language processing (NLP) and artificial intelligence (AI) can assist in making automated and automatically-supported decisions. However, these systems may face challenges and difficulties or find it confusing to identify the required information (characterization) for eliciting a decision by extracting or summarizing relevant information from large text documents or colossal content. When obtaining these documents online, for instance from social networking or social media, these sites undergo a remarkable increase in the textual content. The main objective of the present study is to conduct a survey and show the latest developments about the implementation of text-mining techniques in humanities when summarizing and eliciting automated decisions. This process relies on technological advancement and considers (1) the automated-decision support-techniques commonly used in humanities, (2) the performance evolution and the use of the stylometric approach in text-mining, and (3) the comparisons of the results of chunking text by using different attributes in Burrows' Delta method. This study also provides an overview of the efficiency of applying some selected data-mining (DM) methods with various text-mining techniques to support the critics' decision in artistry – one field of humanities. The automatic choice of criticism in this field was supported by a hybrid approach to these procedures.

Keywords: Automated Decision Support (ADS), Natural Language Processing (NLP), Text- Mining, Stylometry, Burrows’ Method, Hybrid Techniques.

التحقق من المناهج الهجينة لتنقيب البيانات في انظمة دعم قرارات نقد النصوص في العلوم الإنسانية

براء حسن هادي * ، طريف كامل مصطفى
قسم الحاسبات ، كلية العلوم ، جامعة بغداد ، بغداد ، العراق.

الخلاصة

من الممكن ان تساعد اغلب الأنظمة التي تتعامل مع معالجة اللغة الطبيعية (**NLP**) والذكاء الاصطناعي (**AI**)في اتخاذ القرارات ودعمها تلقائيًا. إلا أن هذه الأنظمة قد تواجه تحديات وصعوبات أو قد تكون مربكة في تحديد المعلومات المطلوبة (بمعنى آخر التوصيف) لاستنباط القرار عن طريق استخراج أو تلخيص المعلومات ذات الصلة من المستندات النصية الكبيرة أو المحتوى الضخم. عند الحصول على هذه المستندات عبر الإنترنت عن طريق الشبكات الاجتماعية ووسائل التواصل الاجتماعي، تعاني هذه المواقع من زيادة ملحوظة في المحتوى النصي. إن الهدف الأساسي من هذه الدراسة هو إجراء استعراض لأحدث التطورات

*Email: baraa.elshamery@gmail.com

حول تنفيذ تقنيات تعدين النصوص في العلوم الإنسانية لتلخيص واستنباط القرارات التلقائية بالاعتماد على التقدم التكنولوجي وفيما يتعلق بـ (1) تقنيات دعم القرار التلقائي المستخدم عادتاً في العلوم الإنسانية، (2) تطوير الأداء باستخدام نهج الأسلوب في تعدين النص، و (3) مقارنة النتائج من النص المتقطع باستخدام سمات مختلفة في طريقة دلتا بوروز. توفر هذه المقالة أيضاً نظرة عامة عن كفاءة بعض التطبيقات المنتقاة لطرق التنقيب عن البيانات (**DM**) مع تقنيات تعدين النصوص المتنوعة لدعم قرار النقاد في المجال الفني وذلك لكونه أحد مجالات العلوم الإنسانية. ولقد ساعد استخدام النهج الهجين لهذه الإجراءات على دعم انتقاء النقد تلقائيًا.

1. INTRODUCTION

Text mining is considered as one of the main focuses for the researchers due to the existing technological development and online availability of opinion data. This development represents the backbone that links various fields. For instance, in the fields of humanities, the political arena conducted studies about government issues using specific texts taken from one of the social media platforms (namely, Twitter Tweets), in which the opinions of the audience were identified and comprehended. Moreover, despite the difficulty of slang words, as well as the misspellings, the reactions and feelings of the users are analyzed to achieve the best insight into public opinion on any specific subject [1]. Another instance is from the field of law where crimes and their nature have been identified by different algorithms in some research studies [2] that used text-mining techniques and other algorithms; accordingly, this field considered the text extraction as a useful tool that may help in the detection of some crimes through posts written in Twitter. A third instance is from the fields of arts and culture where NLP methodologies have long been used to improve communication in (social) media, shopping, target ads, search engines, and online platforms. Moreover, for arts practitioners, textual analysis and NLP technologies form a pivotal part, since it contributes to indexing databases for archives and libraries. There are also some illustrative examples in the management of arts, such as searching into the opinion, artist biographies, and analysis of art criticism, as well as the distribution of cultural products [3]. On the other hand, with product development and elicitation of requirements, NLP starts with collecting user opinions from the social network service (SNS) – an online system. This system contains a lot of user opinions and personality relationships, like Weibo, Twitter, and Facebook, based on the statistical analysis to automatically achieve the classification of views [4]. Text-mining is, therefore, considered as a closely related approach to different fields of topics [5], as it is shown in Figure 1. Moreover, the process of extracting the texts has become somewhat complicated due to the incorrect practice of writing online by the users in different languages, in particular in the Arabic language, as explained elsewhere [6], which is due to its linguistic complexity.

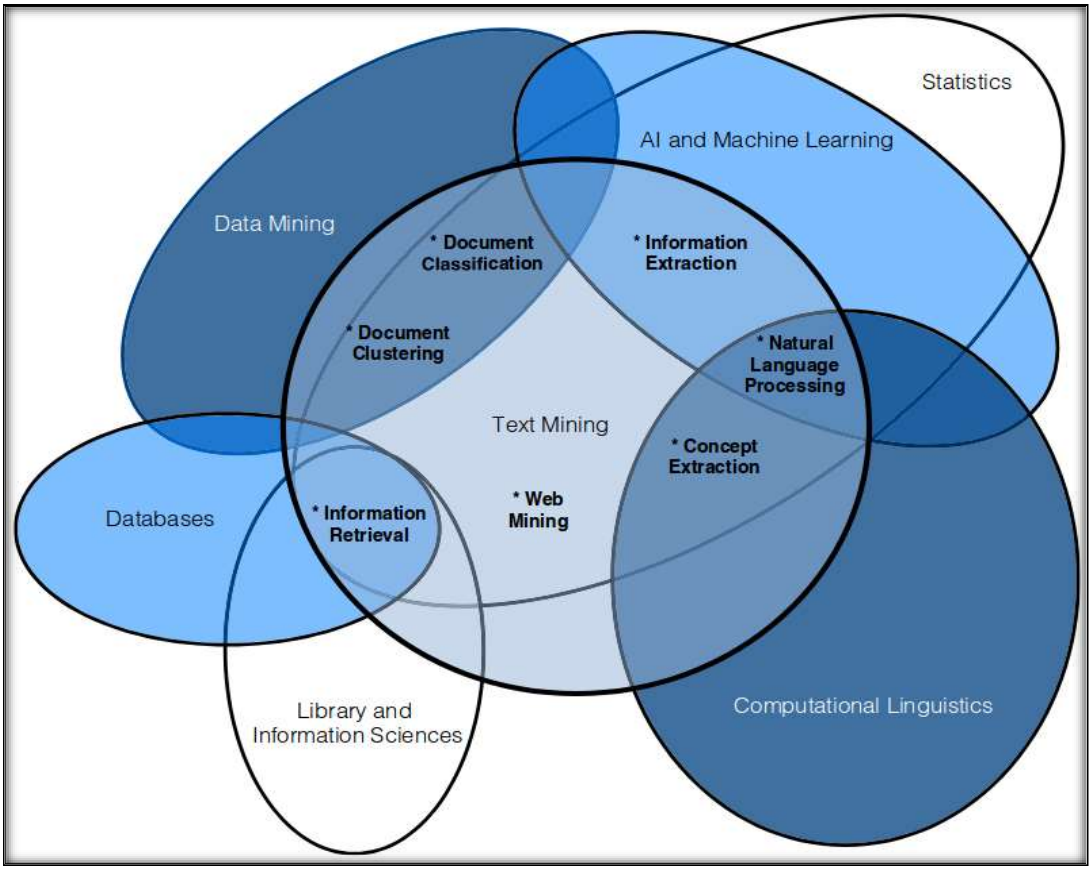


Figure 1-The extent of interconnectedness between text mining and various fields

To deal with these issues, this study suggests a hybrid approach that holds several data-mining methods, such as the Burrows' Method, stylometry approach, Bayesian theorem, and Apriori algorithm, to support specialists and reviewers who have expertise in humanities to be able to give their views in such regard.

The significance of this study is reflected through selecting the most effective attributes in the Burrows' Method, depending on the views written by consumers on artistry products that are available online (Quad attributes). As a consequence of this method and to overcome the increase in the chunking attributes, only the first top 300 attributes have been selected [6].

2. BACKGROUND AND LITERATURE REVIEW

This section presents a review of studies and methodologies related to text mining.

2.1 Digital Humanities with Text-Mining

Computer technology has a long history since the 1940s. It developed during the eighties with linguistics and subsequently used in stylometry approaches, also known as "Humanities Computing" [7]. Despite the differences in linguistic, as well as fundamental structures, between machine and human language while integrating research work into text mining tools, all of the discussions about how humanities research is arranged in computers have shown that machines are making unprecedented changes [[8]. However, many researchers represented the language initially as random numbers and then into binary numbers. On the same perspective, the vast digitization in the fields of literature and the cultural materials was considered as a new opportunity that accompanies contemporary scientific life to extract data and support the analysis of humanities [9]. Therefore, some uses of digitization have significantly devolved in several domains, especially in the fields of humanities, including literature, law, art history, philosophy, politics, and language studies – also known as the liberal arts [10]. In conclusion, text mining is defined as identifying new knowledge from previously unknown and unexpected information to obtain useful information and facts using the computer. Consequently, the methods and algorithms recently adopted have been devolved among the procedures of identifying, preprocessing, and mining of texts [11].

2.2 Automated Decision-Support Systems and Text Summaries

Traditional decision-support systems face serious challenges with summarizing texts due to the increase in text content in contemporary life. Text summarization is made by using many sentences and long paragraphs to identify essential phrases and turning them into a concise text without affecting the general meaning [12]. The unstructured knowledge is also considered necessary since it supports, to a high degree, the structured expertise in contemporary companies and highly-regarded financial markets. The investors need to decide through analysis, and it is usually the case that opinions are summarized out of the views of customers through the Internet. Since manual analysis and summarization are arduous, automatic support is required [13]. On the same perspective, a support system was used to automatically identify and summarize the sentences by removing intrusive expressions or phrases, with the assistance of human specialists in linguistic and grammatical knowledge, as demonstrated earlier [14]. Figure 2 illustrates some concepts of support-decision systems [15].

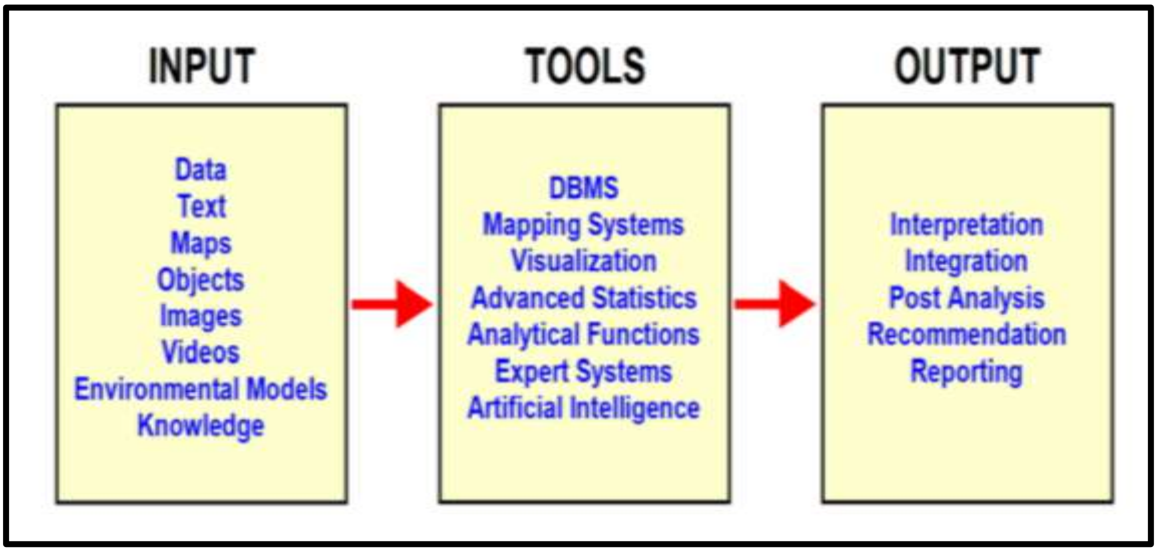


Figure 2-A diagram of the concepts of support-decision systems

Modern systems have recently used integrated approaches, including data mining, text mining, artificial intelligence, and machine learning algorithms. When extracting the required information, text documents are mostly obtained in massive and unorganized amounts. Therefore, it is difficult to manually extract such information because they cannot be void of errors during collection or implementation, which is something that is time- and effort-consuming. The customers can express this in natural language; therefore, automated programs have been developed by using phrases and meanings for the extracted information [16]. Moreover, the description of products and the text reviews of the Internet consumers help identify the implemented standard features, the matter that allows the manufacturers to understand the competitors and to develop products in this field [17].

2.3 Language independency using stylometry

Many old active authorship in literary studies had relied on the function words of language features by using stylometry to differentiate between, for instance, the dramatic works of Shakespeare and those of Fletcher [18].

Stylometry began to improve in the fields of humanities because most of the literature researchers have tendencies to art rather than to science. In other words, they can do better in English rather than in Maths [19]. However, stylometry was not restricted to linguistics, but it also has multiple applications in different fields, as in psycholinguistics, literary studies, forensics, sociolinguistics, as well as medical diagnosis [20]. Subsequently, stylometry has become commonly used in literature whereby the author can be identified. Accordingly, texts can be identified and attributed to their author through the content and the style of writing as, for instance, in research about novels and books that belong to Henry James [21], as well as studies in different languages such as Portuguese.

Researchers have managed to overcome the traditional uses by taking advantage of using natural language processing (NLP) tools. Natural language processing was used as a supporting tool for Automated Plagiarism Detection, as it is the case with Turnitin software, which contains a database of students’ research papers. Turnitin, which was found in 1996, works as a commercial service and is used to detect plagiarism. At the time of Internet use, stylometry is applied to messages sent over the Internet for author verification. Consequently, stylometry is considered as an approach that can be independently used without the need for a language expert [6].

3. METHODS AND TECHNIQUES

This section presents the methods and techniques that can be applied in the domain of humanities.

3.1 Burrows' Delta Method

In 2001, many researchers faced some problems with natural language processing tasks and machine learning methods. Therefore Winnow technique has been suggested for text chunking. Winnow is a regularized algorithm which is known for its effectiveness and durability to deal with irrelevant features, particularly when handling natural languages [22]. In 2002, analysis studies of authorship and stylistics by well-known authors, that were based on sequences of frequent words, were relatively accurate when differentiating whether the document had long or short sections in British and American novels or contemporary critical texts. However, the sequences of frequent words analysis sometimes failed, especially when personal pronouns are eliminated [23]. Burrows (2002) suggested a method that used a tool to measure the difference between two texts. The Delta measure between these texts was reformulated, as it is shown in equation 1 below, where X and Y are the n-dimensional vectors of the attributes words' frequencies at two different documents, and N is the number of attributes. Burrows’ Method was considered as a leading method for authorship attribution that was used to solve accuracy problems that emerged in automatized authorship attribution. It was then reformulated mathematically as the Burrows’ Delta Method to specify the result as a probability in an idiomatic expression that is known as the probability distribution [24].

Equation 1: Burrows’ Measurement

$$\begin{aligned}\sum_{i=1}^{n} |z(X_i) - z(Y_i)| &= \sum_{i=1}^{n} \left| \frac{X_i-\mu_i}{\sigma_i} - \frac{Y_i-\mu_i}{\sigma_i} \right| \\ &= \sum_{i=1}^{n} \left| \frac{(X_i-\mu_i)-(Y_i-\mu_i)}{\sigma_i} \right| \\ &= \sum_{i=1}^{n} \left| \frac{X_i-Y_i}{\sigma_i} \right|\end{aligned}$$

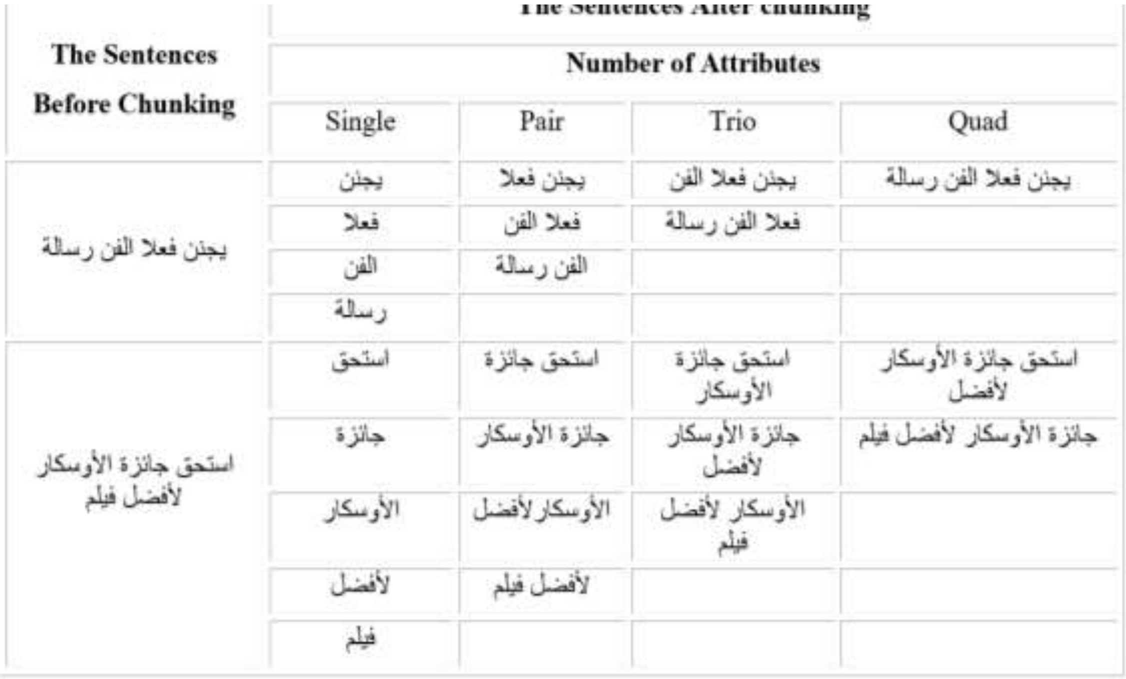
$$= w_1x_1 + w_2x_2 + w_3x_3 + \ldots + w_nx_n$$

The value of µi stands for the population mean of frequency attribute i with the only factor influencing the training set (σi), which stands for the standard deviation for attribute i.

Burrows's Delta probabilistic formulation was mathematically explanatory and effective. However, if it is assumed that the mean distribution (σi) is replaced by Yi, the formula will be similar to Laplace probability distribution. Over time, Linguistic Document Representation and the text itself were combined to attribute the automatic authorship to the alternative text of the traditional model, known as the Bag-of-Words model. Authorship was later tested in various frequency strata where one of these tests controlled variation in word frequency, excluding the commonly used words.

Burrows' Delta Method was improved in 2011 by choosing words for the word vector, the size of the similarity scale, the selected effect on the accuracy of text classification, and the optimal vector size of words which was between the sizes of 200–300 words. Accuracy is degraded beyond 300 words [25]. Eventually, the Burrows' Delta Method is considered the best and most prominent tool of measurement that is used to differentiate literary authorship attribution. The method was also applied to examine its effectiveness and suitability to find out and identify the author's style in the Arabic language. According to a previous study [26], the frequent and trio attributes are more dependable results and proved higher accuracy in the Burrows' Delta Method of up to 300 attributes. However, another study [6] was conducted and achieved further dependable results at the threshold of 300 with Quad attributes. Table 1 is an example illustrating the implementation of the modern Burrows' Method performance in Arabic by the use of various attributes with frequent sequences of words

Table 1- The manner of dealing with Arabic texts based on the Burrows Method



The Sentences Before Chunking	The Sentences After chunking			
	Number of Attributes			
	Single	Pair	Trio	Quad
يجنن فعلا الفن رسالة	يجنن	يجنن فعلا	يجنن فعلا الفن	يجنن فعلا الفن رسالة
	فعلا	فعلا الفن	فعلا الفن رسالة	
	الفن	الفن رسالة		
	رسالة			
استحق جائزة الأوسكار لأفضل فيلم	استحق	استحق جائزة	استحق جائزة الأوسكار	استحق جائزة الأوسكار لأفضل
	جائزة	جائزة الأوسكار	جائزة الأوسكار لأفضل	جائزة الأوسكار لأفضل فيلم
	الأوسكار	الأوسكار لأفضل	الأوسكار لأفضل فيلم	
	لأفضل	لأفضل فيلم		
	فيلم			

3.2 A Hybrid Approach Technique

Extracting and summarizing information from documents on the web is considered as a challenge because these documents are usually unorganized. Accordingly, many researchers have proposed hybrid approaches that involved various selected algorithms and methods for data-mining, as well as for text-mining techniques, as in the case described elsewhere [27]. This research implemented K-Means Clustering, Neural Network, and Text Mining to obtain a hybrid approach that gave an efficient result for the unorganized documents. Moreover, another research paper [28] has used several Data-Mining Algorithms like Association Rule, Naïve Bayesian, and Apriori to apply a new approach for analyzing information for healthcare services and informatics simultaneously. However, improving the hybrid approach was not restricted to the use of algorithms of only data mining and the hybrid machine-learning approach, as used in another work [29]. A third research paper [6] has identified the best feature from a set of elements using the machine-learning algorithm for customer comments and reviews across social networking sites, as shown in Figure 3.

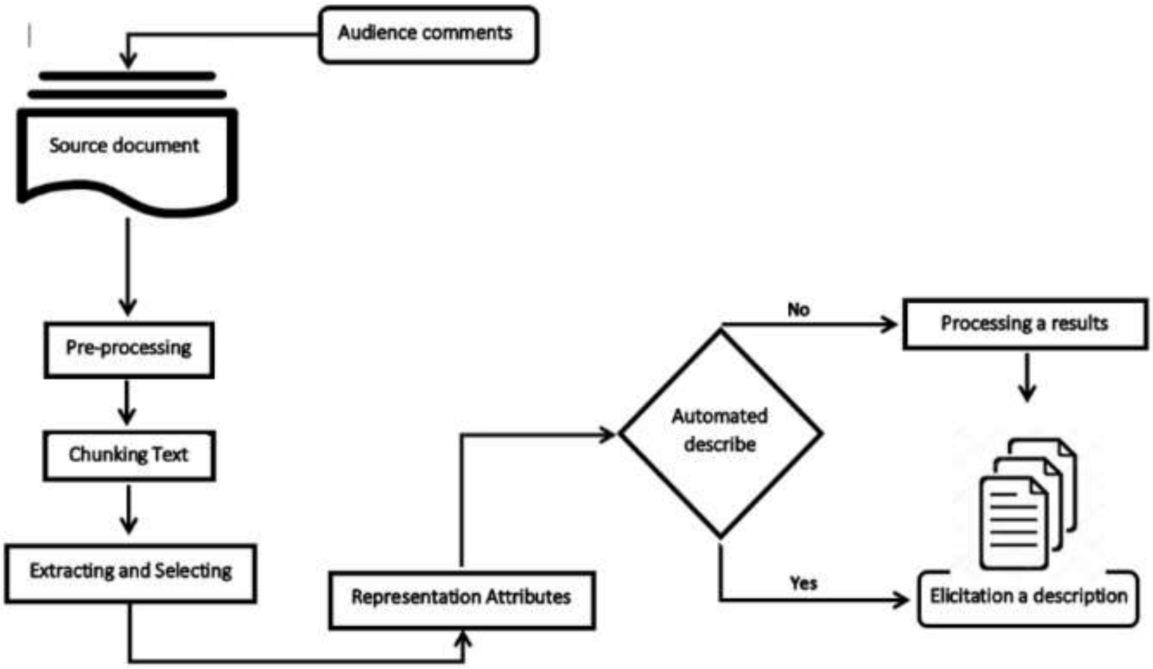


Figure 3-Simplistic framework for Hybrid Technique using reviews online

On the other hand, this hybrid technique was applied in the classification of data available online as negative, positive, or neutral feelings by investing individual search engines, such as Cuckoo Search, and using Artificial Intelligence where the result of the classified training-data was highly accurate [30].

4. CONCLUSIONS

1. Much research has demonstrated that each algorithm or approach has its advantages and drawbacks. A reader can notice that there is no preference for a technique or an algorithm over the other, whether in Text Mining or Data Mining. However, there is still a significant area for new and inventive analysis in this field. Therefore, the purpose of the present study was to acquire familiarity with comprehensive expertise in Automated Decision-Support Systems. With the summarization of this review and by relying on the adopted Hybrid Approach methods and algorithms, the inferences displayed that extensive research has been conducted in the fields of humanities. The findings of this study helped arriving at a number of conclusions; adding different attributes may improve Burrows' Method; selecting a specific threshold as the first top 300 attributes helps to avoid the gap that might raise many phrases from chunking the text; the efficiency of a specific method depends on its implementation; and suggesting the hybrid data-mining approach is to support the automated-decision system for text criticism.

To sum up, the results of this research are generally based on the Hybrid Approach Technique that included the Natural Language Processing techniques of Burrows' Delta Method, stylometry approach, selected data-mining algorithms, and Text-Mining, which are highly-improved by the Automated Decision-Support Systems.

REFERENCES

[1] Meduru, M., et al.,"Opinion mining using twitter feeds for political analysis". I*nternational Journal of Computer*, vol. 25, no. 1, pp. 116-123, 2017.

[2] Hissah, A.-S. and H. Al-Dossari, "Detecting and Classifying Crimes from Arabic Twitter Posts using Text Mining Techniques". *International Journal of Advanced Computer Science and Applications* (IJACSA), vol. 9, no. 10, 2018.

[3] Cieliebak, M., et al., "Natural Language Processing in Arts Management". 2019.

[4] Han, X., et al. "User requirements dynamic elicitation of complex products from social network service". in 2019 25th International Conference on Automation and Computing (ICAC). 2019. IEEE.

[5] Talib, R., et al., "Text mining: techniques, applications and issues. International Journal of Advanced Computer Science and Applications", vol. 7, no. 11, pp. 414-418, 2016.

[6] Hadi, B.H. and T.K. Mustafa, "Hot-Keys Elicitation using Hybrid Text-Mining Classification for Arabic Humanities Articles Criticism". *International Journal of Advance Science and Technology*, 2020.

[7] Kirschenbaum, M.G. "The remaking of reading: Data mining and the digital humanities. in The National Science Foundation symposium on next generation of data mining and cyber-enabled discovery for innovation", *Baltimore, MD*. 2007.

[8] Gold, M.K. and L.F. Klein, *Debates in the Digital Humanities 2016. U of Minnesota Press*, 2016.

[9] Liu, A., Where is cultural criticism in the digital humanities? 2012: eScholarship, University of California.

[10] Schreibman, S., R. Siemens, and J. Unsworth, "The digital humanities and humanities computing: An introduction". *A companion to digital humanities*, pp. xxiii-xxvii, 2004.

[11] Khashfeh, M., M.A. Mahmoud, and M.S. Ahmad. A Text Mining Algorithm Optimising the Determination of Relevant Studies. in 2018 International Symposium on Agent, Multi-Agent Systems and Robotics (ISAMSR). 2018. IEEE.

[12] Gupta, V. and G.S. Lehal, "A survey of text summarization extractive techniques". *Journal of emerging technologies in web intelligence*, vol. 2, no. 3, pp. 258-268, 2010.

[13] He, W., F.-K. Wang, and V. Akula, "Managing extracted knowledge from big social media data for business decision making". *Journal of Knowledge Management*, 2017.

[14] Jing, H., Sentence reduction for automatic text summarization. in Sixth Applied Natural Language Processing Conference. 2000.

[15] Booty, W., I. Wong, and C.S. Jao, "Case Studies of Canadian Environmental Decision Support Systems". *Decision Support Systems*, pp. 217-242, 2010.

[16] Hogenboom, F., et al., "A survey of event extraction methods from text for decision support systems". *Decision Support Systems*, vol. 85, pp. 12-22, 2016.

[17] Wang, W.M., et al., "Extracting and summarizing affective features and responses from online product descriptions and reviews: A Kansei text mining approach". *Engineering Applications of Artificial Intelligence*, vol. 73, pp. 149-162, 2018.

[18] Horton, T.B., "The Effectiveness of the Stylometry of Function words in Discriminating between Shakespeare and Fletcher". 1987.

[19] Holmes, D.I., "The evolution of stylometry in humanities scholarship". *Literary and linguistic computing*, vol. 13, no. 3, pp. 111-117, 1998.

[20] Smith, M., "Forensic stylometry: A theoretical basis for further developments of practical methods". *Journal of the Forensic Science Society*, vol. 29, no. 1, pp. 15-33, 1989.

[21] Hoover, D.L., "Corpus stylistics, stylometry, and the styles of Henry James". *Style*, vol. 41, no. 2, pp. 174-203, 2007.

[22] Zhang, T., F. Damerau, and D.E. Johnson. Text chunking using regularized winnow. in Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics. 2001.

[23] Hoover, D.L., "Frequent word sequences and statistical stylistics". *Literary and Linguistic Computing*, vol. 17, no. 2, pp. 157-180, 2002.

[24] Stein, S. and S. Argamon. A mathematical explanation of Burrows's Delta. in Proceedings of the Digital Humanities Conference. 2006. Citeseer.

[25] Smith, P.W. and W. Aldridge, "Improving authorship attribution: optimizing Burrows' Delta method". *Journal of Quantitative Linguistics*, vol. 18, no. 1, pp. 63-88, 2011.

[26] Abdul-Razzaq, A.A. and T.K. Mustafa, "Burrows-Delta method fitness for Arabic text authorship Stylometric detection". *IJCSMC*, vol. 3, pp. 69-78, 2014.

[27] Umamaheswari, R. and N. Shanthi, "An Efficient Hybrid Information Retrieval Approach for Unstructured Document Classification". *International Journal of Applied Engineering Research*, vol. 10, no. 24, pp. 44504-44508, 2015.

[28] Mustafa, T.K. and M.S. Abd, "Proposed approach for analysing general hygiene information using various data mining algorithms". *Iraqi Journal of Science*, vol. 58, no. 1B, pp. 337-344, 2017.

[29] Tripathy, A., A. Anand, and S.K. Rath, "Document-level sentiment classification using hybrid machine learning approach". *Knowledge and Information Systems*, vol. 53, no. 3, pp. 805-831, 2017.

[30] Kansal, V. and R. Kumar. A Hybrid Approach for Financial Sentiment Analysis Using Artificial Intelligence and Cuckoo Search. in 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS). 2019. IEEE.