2024, 21(2 Special Issue): 0662-0683
https://doi.org/10.21123/bsj.2024.9750
P-ISSN: 2078-8665 - E-ISSN: 2411-7986
Baghdad Science Journal

Prioritized Text Detergent: Comparing Two Judgment Scales of Analytic Hierarchy Process on Prioritizing Pre-Processing Techniques on Social Media Sentiment Analysis

Ummu Hani' Hair Zaki*[1], Roliana Ibrahim[1], Shahliza Abd Halim[2], Izyan Izzati Kamsani[1]

[1]Department of Applied Computing, Faculty of Computing, Universiti Teknologi Malaysia, Johor Bahru, Johor, Malaysia.
[2]Department of Software Engineering, Faculty of Computing, Universiti Teknologi Malaysia, Johor Bahru, Johor, Malaysia.
*Corresponding Author.
ICAC2023: The 4th International Conference on Applied Computing 2023.

Received 29/09/2023, Revised 10/02/2024, Accepted 12/02/2024, Published 25/02/2024

Abstract

Most companies use social media data for business. Sentiment analysis automatically gathers analyses and summarizes this type of data. Managing unstructured social media data is difficult. Noisy data is a challenge to sentiment analysis. Since over 50% of the sentiment analysis process is data pre-processing, processing big social media data is challenging too. If pre-processing is carried out correctly, data accuracy may improve. Also, sentiment analysis workflow is highly dependent. Because no pre-processing technique works well in all situations or with all data sources, choosing the most important ones is crucial. Prioritization is an excellent technique for choosing the most important ones. As one of many Multi-Criteria Decision Making (MCDM) methods, the Analytic Hierarchy Process (AHP) is preferred for handling complicated decision-making challenges using several criteria. The Consistency Ratio (CR) scores were used to examine pair-wise comparisons to evaluate the AHP. This study used two judgment scales to get the most consistent judgment. Firstly, the Saaty judgment scale (SS), then the Generalized Balanced Scale (GBS). It investigated whether two different AHP judgment scales would affect decision-making. The main criteria for prioritizing pre-processing techniques in sentiment analysis are Punctuation, Spelling, Number, and Context. These four criteria also contain sub-criteria. GBS pair-wise comparisons are closer to the CR value than SS, reducing the alternatives' weight ratios. This paper explains how AHP aids logical decision-making. Prioritizing pre-processing techniques with AHP can be a paradigm for other sentiment analysis stages. In short, this paper adds another contribution to the Big Data Analytics domain.

Keywords: Analytical Hierarchy Process, Data Pre-processing, Multi-Criteria Decision-Making, Pre-processing Technique, Prioritization, Sentiment Analysis, Social Media.

Introduction

Social media services like Twitter, Facebook, and YouTube generate enormous amounts of data [1,2]. Analyzing this data type is an approach most brands use to translate social media behavior into actionable business data. Sentiment analysis is an example of a social analytic method that

2024, 21(2 Special Issue): 0662-0683
https://doi.org/10.21123/bsj.2024.9750
P-ISSN: 2078-8665 - E-ISSN: 2411-7986
Baghdad Science Journal

automatically extracts, analyzes, and summarizes user-generated data [3]. However, social media data are often unstructured and difficult to manage. Noisy data can be a bottleneck that reduces the quality of the entire sentiment analysis pipeline. Examples of noises can be observed within social media data, including but not limited to the presence of slang, typographical errors, the repetition of characters within a word (resulting in elongation), intricate spelling mistakes, inadequately structured sentences, the combination of words, uncommon usage of acronyms, diverse form of word abbreviations, varying grammatical structures, and an overall informal linguistic style when compared with longer texts and standard documents.

In addition, processing massive amounts of social media data is an intense task [4]. That is why data pre-processing has become one of the significant phases in sentiment analysis [5]. Data pre-processing often involves more effort and time within the entire data analysis process, with more than 50% of the total effort [6]. Data pre-processing is the most critical process in maintaining data quality. It might reduce data accuracy if done incorrectly [7]. By removing noise from social media data, sentiment analysis and better decision-making based on unstructured data may begin [8].

However, most studies on machine learning do not include data pre-processing. As examined by [9], even in the studies where pre-processing was mentioned, only some parts of various techniques were presented. In another work, [10] proposed a quality model as noise filtering for multiple social media services data sources [11]. But the performance is not as good as expected. In its “Data Corruption” noise filtering, only 21.3% of the data items were fixed. It demonstrated that applying the “Data Corruption” noise filtering is a time and resource-intensive process that is not always scalable to large datasets. The performance of future learning algorithms will thus be undermined if they are presented with low-quality data.

Also, since sentiment analysis is a part of Big Data Analytics [12], the workflow is also highly dependent on various constraints [13,14]. Getting accurate results from sentiment analysis depends a lot on each phase of sentiment analysis, especially the data pre-processing phase. Because no single pre-processing technique works well in all situations or with all data sources [15,16], it is crucial to put the most important ones at the top of the list. The investigation of the ranking of pre-processing techniques concerning the degree of noise is a topic that has received limited attention in the research literature. This aspect holds significant importance in decision-making processes, particularly in critical scenarios such as disaster management.

To be exact, the degree of noise depends upon various parameters, including Punctuation, Spelling, Number, and Context [17]. For example, the Spelling noise posted on Twitter should be a higher priority than the Spelling noise posted on Facebook. It is due to Twitter generating the most data but with a short data sharing limit (280 characters) [18]; thus, users tend to do abbreviations, short forms, and misspell more. Another example is if a dataset from Twitter is collected containing Punctuation parameters, it should be prioritized with a higher-ranking score than Spelling, Number, and Context parameters.

So, the significance of the parameters or features varies as per the type of social media service. To achieve a systematic way when dealing with noise in social media data, [19] implemented 12 pre-processing techniques in a systematic sequence order. However, they were using only Twitter datasets. In addition, only a limited number of scholarly works have explored the utilization of various social media platforms. Therefore, this paper highlights the importance of analyzing a broader range of social media services that contain significant information. Utilizing a diverse range of social media services is a practical approach from a content-wise perspective [18].

To emphasize, as part of the highly dependent workflow of Big Data Analytics, plus with social media data generally tending to be highly noisy, the problem of prioritizing pre-processing techniques remains for practice and academic research. The decision involved in this problem is highly considering a diverse set of criteria that often may conflict. For example, when choosing a method of transportation, one might put speed, cost, and effect on the environment at the top of the list. MCDM helps find a balance and make choices based on good information when these criteria do not always match up. Therefore, MCDM offers a set of sophisticated techniques to help decision-makers choose the best option by considering different, sometimes conflicting, criteria [20].

2024, 21(2 Special Issue): 0662-0683
https://doi.org/10.21123/bsj.2024.9750
P-ISSN: 2078-8665 - E-ISSN: 2411-7986
Baghdad Science Journal

Related Works

Rational and accurate decision-making is one of the most critical processes for any organization [21]. Therefore, the need for better and faster decisions based on data rather than insightful choices is now more critical than ever. One of the most effective ways to settle on the perfect decision is through prioritization [22]. Prioritization techniques support the decision-making process. It can help determine what order to complete specific goals so it can be done more efficiently [22,23]. Combining MCDM with data mining, machine learning, and predictive analytics has been used in research [24–27]. MCDM techniques are still essential for making data-based decisions [21] with real-world problems with many criteria, objectives, and goals that often conflict.

Various MCDM methods, such as Elimination and Choice Translating Reality (ELECTRE), Preference Ranking Organization Method for Enrichment Evaluation (PROMETHEE), AHP, and Technique for Order Preference by Similarity to Ideal Solution (TOPSIS), have emerged in the recent past [28]. It has been empirically tested in various decision-making scenarios. In the majority of previous research works, the AHP has been favored as a methodological approach for addressing complex decision-making challenges that involve the consideration of multiple criteria [26,29,30]. The AHP method is known for being relatively simple, easily understandable and practical [26]. The ranking process will require a pretty long time if manually analyzed. Therefore, applying AHP resolves the issue [31].

The AHP method uses both mathematical and psychological observations. It considers the objectivity and the subjectivity of the people making the decisions [32]. In AHP, problems are broken down into an order of criteria and possible solutions (alternatives), and then each solution is given a weight. Generally speaking, the process of AHP is considered valid if the CR is less than 0.1 [33]. The AHP facilitates the ongoing enhancement, progression, monitoring, and evaluation of a given subject matter by repeating the pair-wise comparison procedure with domain-specific experts. AHP represents a notable advancement compared to most existing methods, as it effectively addresses a common limitation found in these methods - the lack of transparency in justifying the selection and weighing process for determining relative weights [34].

Much literature has been published on implementing MCDM through AHP in social media sentiment analysis. Most of the work related to sentiment analysis concentrates on recommending the most influential topics. For example, some studies were implemented on the most influential topics, such as social media accounts [24,35,36], vehicle consumer consumption behavior [37], and disaster vulnerability [29]. Other studies were conducted to identify the essential factors of fake social media accounts [25], the severity of urban issue complaints [26], and the organization's reputation [27]. Understandably, AHP's influence on decision-making is evident. However, most AHP studies have only used sentiment analysis to rank or choose related criteria. They used sentiment analysis to identify the list of essential criteria which will be used to evaluate, rank, or prioritize a particular topic.

Interestingly, there is one study by [38] where researchers have put forth a novel MCDM methodology wherein the AHP is utilized. This approach aims to determine the most appropriate sentiment analysis algorithm for specific business problems. It is achieved by considering various relevant criteria within a given context, making the decision-making process context aware. They are ranking sentiment analysis algorithms in different business cases. This model checks the consistency of the decision maker's evaluations (ranking of the sentiment analysis algorithm), ensuring a bias reduction in the decision. The CR is calculated to measure the consistency of the matrices. As mentioned, the value of this ratio must be less than or equal to 0.1 for the matrix to be reliable [26,38,39]. They use datasets about hotel reviews, movie reviews, and sentiment140.

However, there is less effort to prioritize techniques or methods, especially pre-processing techniques for social media sentiment analysis. AHP has been overlooked in sentiment analysis data pre-processing. The correct text data pre-processing techniques will improve decision-making [16]. Changes or failures in composing pre-processing techniques can affect social media sentiment analysis accuracy. It needs a new method to get rid of the noise in social media data using multiple social media services as data sources. A sequence of prioritized pre-processing techniques that results in

2024, 21(2 Special Issue): 0662-0683
https://doi.org/10.21123/bsj.2024.9750
P-ISSN: 2078-8665 - E-ISSN: 2411-7986
Baghdad Science Journal

the best sentiment classifier performance is recommended. Study by [6] surveyed data pre-processing for data mining and supported this view.

With attention to the weight obtained during the pair-wise comparison process in AHP, the weights assigned to criteria or alternatives are commonly represented as a priority vector. Several parameters are related to determining and understanding weights assigned to the criteria or alternatives. Those parameters are known as the weight bound, the weight ratio, the weight uncertainty, and the weight dispersion. These parameters help people make decisions by assigning judgment values to how essential or preferred different criteria of a hierarchical system are.

The weight bound refers to the upper and lower limits or boundaries set for the weights assigned to criteria or alternatives in AHP. The weights assigned to each criterion should fall within this range. It ensures that no criterion is overly dominant or negligible. The weight bound helps maintain a balanced evaluation. It helps ensure that the weights are not excessively biased or extreme. After pair-wise comparisons have been made, the weight ratio will be gained. These weight ratios reflect the relative importance she assigns to each criterion. The weight ratio represents the relative importance or priority of one criterion or alternative compared to another. It is determined through pair-wise comparisons, where decision-makers assess the relative significance between elements. The weight ratio quantifies the degree of preference or importance assigned to different elements with each other.

Decision-making involves uncertainty. It means that sometimes, the decision maker might not be entirely sure about the exact weight ratios assigned to the criteria. Weight uncertainty reflects the level of uncertainty or lack of confidence associated with the assigned weights in AHP. It acknowledges that decision-makers may not have precise or perfect knowledge regarding the relative importance of criteria or alternatives. Weight uncertainty captures the degree of doubt or ambiguity in the assigned weights. The weight dispersion can be observed after determining the weights for each criterion. Weight dispersion refers to the variability or spread of the assigned weights across criteria or alternatives in AHP. It measures how dispersed, concentrated, or distributed the weights are within the decision hierarchy. A higher weight dispersion indicates a more significant difference in importance or priority among the elements, while a lower weight dispersion suggests more similarity in their relative priorities.

By considering these concepts in the AHP scale, decision-makers can make more informed decisions. Many examples can be found within the existing body of literature wherein various judgment scales have been employed and subsequently compared to determine the optimal solution. A study conducted by [39] has shed light on the influence of 11 distinct scales on the resultant priorities, thereby facilitating the identification of a suitable scale for projects employing the AHP. The findings indicate that implementing the GBS enhances weight dispersion and uncertainty compared to the original AHP scale, the SS. The ABS overcomes the problem of a change in the maximum weight depending on the Number of decision criteria.

Materials and Methods

This study aims to prioritize text pre-processing techniques for social media sentiment analysis using the AHP technique. AHP defines the prioritization criteria through a priority assessment of all possible criteria pairs. It uses a pair-wise comparison matrix. AHP exploits pair-wise comparison to determine how one of the criteria is more important than the other. The AHP method has been designed to be fault-tolerant, meaning it can handle errors or failures without significantly impacting its overall performance. AHP also incorporates a consistency check, a mechanism used to ensure that the decision-making process remains reliable and free from inconsistencies. The priorities from this analysis are relative, as they are determined based on a ratio scale. This scale enables practical criteria assessment, facilitating a more comprehensive evaluation process. This work adopts the AHP technique using AHP-OS [40]. Fig. 1 shows the brief process of this study.

2024, 21(2 Special Issue): 0662-0683
https://doi.org/10.21123/bsj.2024.9750
P-ISSN: 2078-8665 - E-ISSN: 2411-7986
Baghdad Science Journal

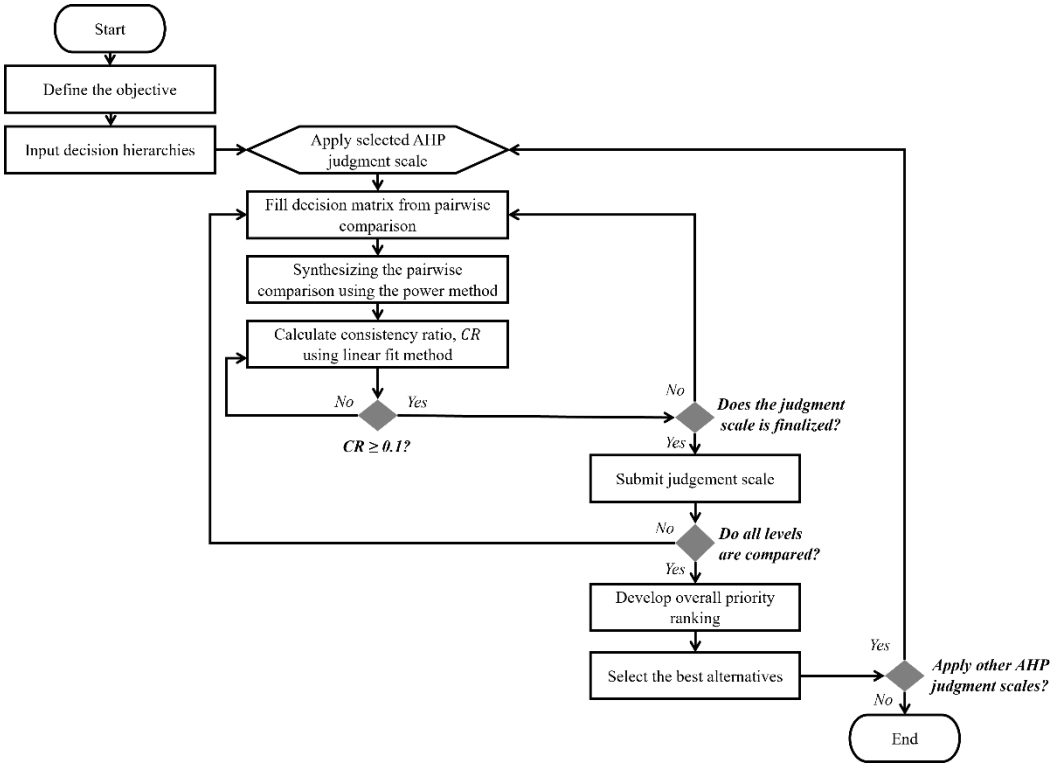

Figure 1. The prioritized text detergent process

Step 1: Define Objective

A case study for this research is about social media data pre-processing in sentiment analysis [17]. After reviewing research related to pre-processing techniques for social media data, there are 14 techniques to be included. The implementation sequence of techniques is also arranged based on statistical analysis. The sequence of techniques is determined according to the level of noise. As tabulated in Table 1, the degree of social media data noise depends upon several parameters, like Punctuation, Spelling, Number, and Context. There is a direct correlation between the degree value and data de-noising urgency. Specifically, as the degree value decreases, de-noising data becomes increasingly urgent. Ensuring the accuracy and consistency of textual data is imperative to pre-process the corpus. This step is crucial as it allows for a cleaner and more streamlined dataset, facilitating subsequent analyses.

Once the Punctuation has been removed, the next step involves enhancing the Spelling of the text by standardizing terms. This process ensures the corpus adheres to established linguistic conventions, enabling more effective and accurate research outcomes. Subsequently, the Number should be eliminated as it lacks substantial influence on the sentiment analysis process. It is imperative to execute the Context parameter to facilitate the understanding of the dataset for the analyst. From a detailed perspective, when dealing with different data sources (social media services), it is essential to prioritize the urgent and critical pre-processing techniques over the other. It proved that selecting pre-processing techniques for multiple social media services is vital. It can be completed by using the prioritization technique, i.e., AHP.

2024, 21(2 Special Issue): 0662-0683
https://doi.org/10.21123/bsj.2024.9750
P-ISSN: 2078-8665 - E-ISSN: 2411-7986
Baghdad Science Journal

Table 1. The noise mapped with respective pre-processing techniques

Value	Degree	Noise	Pre-processing techniques
1	Punctuation	URLs	Remove URLs
		Hashtag symbol	Remove hashtag
		User mention	Remove user mention
		Emoticon	Replace emoticon
		Word contraction	Replace contraction
		Punctuation characters	Remove punctuation characters
2	Spelling	Extended spelling character	Data elongation
		Slang and acronyms	Expand slang and acronyms
		Misspelling word	Spelling correction
3	Number	Number	Remove number
4	Context	Uppercase text	Lowercase
		Stop words	Remove stop words
		Word derivation & word inflection	Lemmatization
		Word derivation & word inflection	Stemming

Step 2: Input Hierarchy

The authors reviewed and reanalyzed the problem hierarchically. The criteria are gained through this paper's literature review process and experimentation work [17]. The problem is decomposed into levels, where Level 0 is about the goal or objective which addresses the problem. After that, Level 1 will consist of the main criteria. It is followed by Level 2, sub-criteria (on which the next level will depend). Fig. 3 shows the decision hierarchy of this work. Initially, the decision's primary aim or ultimate objective can be identified at the highest level of the hierarchical structure. The primary objective of this application is to identify and determine the optimal pre-processing techniques for conducting sentiment analysis on social media data.

The second level relates to the primary criteria that influence the process of data pre-processing in sentiment analysis for social media. The main criteria can be classified into four parameters: Punctuation (P), Spelling (S), number (N), and Context (C). The sub-criteria are represented at the third level of the hierarchy. Six sub-criteria affect the punctuation pre-processing such as removing URLs (RU), removing hashtags (RH), removing user mention (RUM), replacing emoticons (RE), replacing contraction (RC), and removing punctuation characters (RPC). Data elongation (DE), expand slang and acronyms (ESAA), and spelling correction (SC) are sub-criteria that affect in term of Spelling. As for the Number, sub-criteria remove the Number (RN). While lowercase (LOW), remove stop words (RSW), lemmatization (LEM), and stemming (STE) are sub-criteria affecting in terms of Context parameters, respectively.

Finally, the social media service alternatives as data sources for social media sentiment analysis are identified at the lowest level of the hierarchy. These are the decision options, as shown in Fig. 2. There are more social media services with important information that needs to be analyzed, but few pieces of work use more than one social media service [11]. As examined by [18], the most significant and related dataset is produced by Facebook. In terms of user engagement, it has been observed that social media users exhibit a preference for the YouTube platform. It is reasonable to argue that employing a diverse range of social media services can be advantageous in terms of content wise. Various sources in the field of sentiment analysis on social media services indicate that incorporating multiple services can be a beneficial strategy in terms of the importance, relevance, and level of user interaction with the content.

2024, 21(2 Special Issue): 0662-0683
https://doi.org/10.21123/bsj.2024.9750
P-ISSN: 2078-8665 - E-ISSN: 2411-7986
Baghdad Science Journal

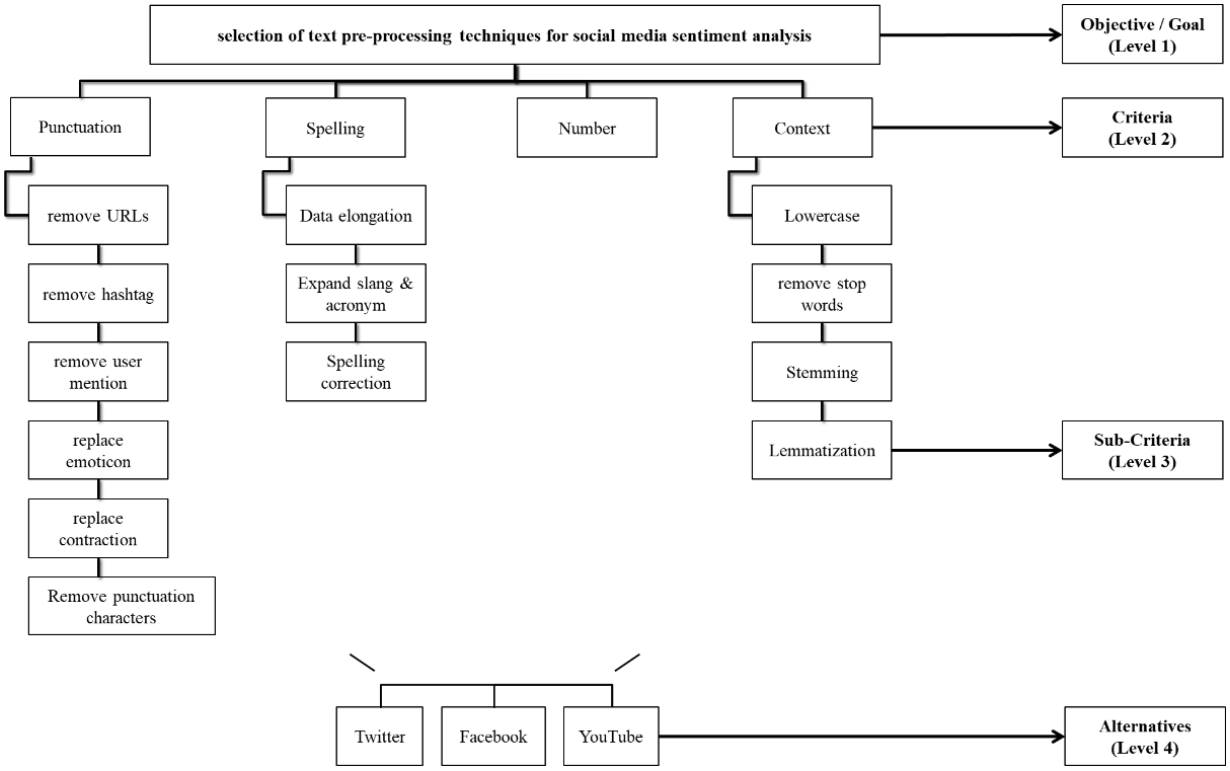

Figure 2. A hierarchy model for the selection of text pre-processing techniques

Step 3: Apply the Judgment Scale

In this section, the SS introduced by [33] is compared with the GBS [39]. The SS is constructed as the equation (1), where x as the value on the integer judgment scale for pair-wise comparisons from 1 to 9 while c as the ratio used as entry into the decision matrix.

$$c = x \quad (1)$$

$$c = \frac{9 + (n-1)x}{9 + n - x} \quad (2)$$

The GBS is described with x as the value on the integer judgment scale for pair-wise comparisons from 1 to 9, c as the ratio used as entry into the decision matrix, and n as the Number of criteria. The detail of the calculation can be seen in equation (2). The GBS improves weight dispersion and uncertainty compared to the fundamental AHP scale. A further advantage is that their weight uncertainty is constant over the whole judgment range from 1 to 9, and the uncertainty does not exceed 5% for up to ten criteria. Practical projects indicate an improvement of the CR for the GBS.

Step 4: Fill the Decision Matrix from the Pair-wise Comparison

After constructing the decision hierarchy model, a pair-wise comparison between each criterion and possibly sub-criterion must be made. The assessment of each criterion should be conducted by referring to the scales specified in Table 2. The present study demonstrates the rigorous assessment of individual criteria by utilizing the judgment scale, which is subsequently employed to develop the matrix. The relative importance of each pair of criteria elements was assessed using a numerical scale ranging from 1 to 9. A higher value on this scale indicates that the chosen criteria element is more significant than the other criteria element being compared. The markings in Table 3 until Table 6 explain the assignment of values for every criterion in each matrix that applies SS. While the markings in Table 7 until Table 10 explain the assignment of values for every criterion in each matrix that applies GBS.

2024, 21(2 Special Issue): 0662-0683
https://doi.org/10.21123/bsj.2024.9750
P-ISSN: 2078-8665 - E-ISSN: 2411-7986
Baghdad Science Journal

Table 2. Fundamental AHP judgment scale

Judgment value	Judgment description
1	Equally important
3	Moderately more important
5	Strongly more important
7	Very strongly more important
9	Extremely more important
2, 4, 6, 8	Intermediate importance between two adjacent judgment scales
1/2, 1/3, 1/4, 1/5, 1/6, 1/7, 1/8, 1/9	Reciprocals; the values for inverse comparison

Table 3. Pair-wise comparison of criteria for the overall goal using SS

matrix	P	S	N	C
P	1	4	5	7
S	1/4	1	5	4
N	1/5	1/5	1	2
C	1/7	1/4	1/2	1
Sum (col)	**1.5929**	**5.4500**	**11.5000**	**14.0000**

Table 4. Pair-wise comparison of criteria for sub-criteria Punctuation using SS

matrix	RU	RH	RUM	RE	RC	RPC
RU	1	3	4	5	6	9
RH	1/3	1	2	3	4	7
RUM	1/4	1/2	1	3	5	7
RE	1/5	1/3	1/3	1	2	6
RC	1/6	1/4	1/5	1/2	1	2
RPC	1/9	1/7	1/7	1/6	1/2	1
Sum (col)	**2.0611**	**5.2262**	**7.7662**	**12.6667**	**18.5000**	**32.0000**

Table 5. Pair-wise comparison of criteria for sub-criteria Spelling using SS

matrix	DE	ESAA	SC
DE	1	3	5
ESAA	1/3	1	2
SC	1/5	1/2	1
Sum (col)	**1.5333**	**4.5000**	**8.0000**

Table 6. Pair-wise comparison of criteria for sub-criteria Context using SS

matrix	LOW	RSW	LEM	STE
LOW	1	4	5	6
RSW	1/4	1	4	5
LEM	1/5	1/4	1	2
STE	1/6	1/5	1/2	1
Sum (col)	**1.6167**	**5.4500**	**10.5000**	**14.0000**

Table 7. Pair-wise comparison of criteria for the overall goal using GBS

matrix	P	S	N	C
P	1	2 1/3	3	5
S	3/7	1	3	2 1/3
N	1/3	1/3	1	1 1/3
C	1/5	3/7	3/4	1
Sum (col)	1.9619	4.0952	7.7333	9.6970

Table 8. Pair-wise comparison of criteria for sub-criteria Punctuation using GBS

matrix	RU	RH	RUM	RE	RC	RPC
RU	1	2	2 5/8	3 2/5	4 1/3	9
RH	1/2	1	1 1/2	2	2 5/8	5 1/2
RUM	3/8	2/3	1	2	3 2/5	5 1/2
RE	2/7	1/2	1/2	1	1 1/2	4 1/3
RC	1/4	3/8	2/7	2/3	1	1 1/2
RPC	1/9	1/5	1/5	1/4	2/3	1
Sum (col)	**2.5115**	**4.7445**	**6.074**	**9.315**	**13.515**	**26.795**

Table 9. Pair-wise comparison of criteria for sub-criteria Spelling using GBS

matrix	DE	ESAA	SC
DE	1	1 2/3	2 5/7
ESAA	3/5	1	1 1/3
SC	3/8	3/4	1
Sum (col)	**1.9684**	**3.4359**	**5.0143**

Table 10. Pair-wise comparison of criteria for sub-criteria Context using GBS

matrix	LOW	RSW	LEM	STE
LOW	1	2 1/3	3	3 6/7
RSW	3/7	1	2 1/3	3
LEM	1/3	3/7	1	1 1/3
STE	1/4	1/3	3/4	1
Sum (col)	**2.0212**	**4.0952**	**7.0667**	**9.2208**

2024, 21(2 Special Issue): 0662-0683
https://doi.org/10.21123/bsj.2024.9750
P-ISSN: 2078-8665 - E-ISSN: 2411-7986
Baghdad Science Journal

Step 5: Synthesizing the Pair-wise Comparison

After all pair-wise comparisons, the pair-wise comparison matrix is created. A mathematical formula calculates the weights of each criterion or alternative using this matrix. Weights can be used to rank alternatives by importance. This section calculates the priority vector using the pair-wise comparison from the previous section. The priority vector is also called the normalized principal eigenvector. The principal eigenvector is associated with the largest or the dominant eigenvalue [41]. To find the eigenvalues and eigenvectors along with the dominant one, λ of A, can be applied using the power method [42].

The power method is a numerical method commonly adopted with AHP. It is relatively simple to implement, making it an attractive choice for quickly obtaining an approximation of the dominant eigenvalue and its corresponding eigenvector. It is an iterative algorithm that can determine the dominant eigenvalue of a matrix A [43]. To get a priority vector using the power method, firstly must have an $n \times n$ of matrix A. The matrix will have n eigenvalues $\lambda_1, \lambda_2, \lambda_3, \cdots \lambda_n$ and its corresponding eigenvectors $v_1, v_2, v_3, \cdots v_n$. The initial guess for the eigenvector, usually $v_{k=0} = [1 \cdots 1]^T$, and normalize it by assigning $v_0 = \frac{v}{m_{k+1}}$. The T is a transpose matrix which means switching its rows with its columns. The dominant eigenvalue (the largest magnitude), λ_1, where $|\lambda_1| > |\lambda_2| > \cdots |\lambda_n|$ and its corresponding eigenvector, v_1, can be obtained by the power method, such as described in the equation (3), where m_{k+1} is the maximum absolute value of Av^k.

$$v^{(k+1)} = \frac{1}{m_{k+1}} Av^{(k)}, k = 0, 1, 2, \cdots, \quad 3$$

The power method will repeat with a starting eigenvector, v^0 until all the values of v^k are the same. Then, the process will stop. If v_k denotes the last vector calculated in this process, then the absolute value of the dominant eigenvalue is $\lambda_1 = m_{k+1}$, and its corresponding eigenvector is $v_1 = v^{k+1}$. The summary results for this calculation are shown in the section below. Table 11 until 14 shows the synthesized matrix using SS that produces a priority vector for criteria and sub-criteria. While Table 15 until Table 18 shows the synthesized matrix using GBS that produces a priority vector for criteria and sub-criteria.

Table 11. Synthesized SS matrix for the overall goal

matrix	P	S	N	C	Priority vector
P	1	4	5	7	0.5907
S	1/4	1	5	4	0.2593
N	1/5	1/5	1	2	0.0899
C	1/7	1/4	1/2	1	0.0601
Sum (col)	**1.5929**	**5.4500**	**11.5000**	**14.0000**	**1.0000**

Table 12. Synthesized SS matrix for Punctuation

matrix	RU	RH	RUM	RE	RC	RPC	Priority vector
RU	1	3	4	5	6	9	0.4410
RH	1/3	1	2	3	4	7	0.2174
RUM	1/4	1/2	1	3	5	7	0.1746
RE	1/5	1/3	1/3	1	2	6	0.0905
RC	1/6	1/4	1/5	1/2	1	2	0.0494
RPC	1/9	1/7	1/7	1/6	1/2	1	0.0272
Sum (col)	**2.0611**	**5.2262**	**7.6762**	**12.6667**	**18.5000**	**32.0000**	**1.0000**

Table 13. Synthesized SS matrix for Spelling

matrix	DE	ESAA	SC	Priority vector
DE	1	3	5	0.6483
ESAA	1/3	1	2	0.2296
SC	1/5	1/2	1	0.1220
Sum (col)	**1.5333**	**4.5000**	**8.0000**	**1.0000**

Table 14. Synthesized SS matrix for Context

matrix	LOW	RSW	LEM	STE	Priority vector
LOW	1	4	5	6	0.5849
RSW	1/4	1	4	5	0.2591
LEM	1/5	1/4	1	2	0.0948
STE	1/6	1/5	1/2	1	0.0613
Sum (col)	**1.6167**	**5.4500**	**10.5000**	**14.0000**	**1.0000**

2024, 21(2 Special Issue): 0662-0683
https://doi.org/10.21123/bsj.2024.9750
P-ISSN: 2078-8665 - E-ISSN: 2411-7986
Baghdad Science Journal

Table 15. Synthesized GBS matrix for the overall goal

matrix	P	S	N	C	Priority vector
P	1	2 1/3	3	5	0.4978
S	3/7	1	3	2 1/3	0.2723
N	1/3	1/3	1	1 1/3	0.1287
C	1/5	3/7	3/4	1	0.1012
Sum (col)	**1.9619**	**4.0952**	**7.7333**	**9.6970**	**1.0000**

Table 16. Synthesized GBS matrix for Punctuation

matrix, A	RU	RH	RUM	RE	RC	RPC	Priority vector
RU	1	2	2 5/8	3 2/5	4 1/3	9	0.3808
RH	1/2	1	1 1/2	2	2 5/8	5 1/2	0.2118
RUM	3/8	2/3	1	2	3 2/5	5 1/2	0.1878
RE	2/7	1/2	1/2	1	1 1/2	4 1/3	0.1128
RC	1/4	3/8	2/7	2/3	1	1 1/2	0.0697
RPC	1/9	1/5	1/5	1/4	2/3	1	0.0371
Sum (col)	**2.515**	**4.745**	**6.074**	**9.315**	**13.515**	**26.795**	**1.0000**

Table 17. Synthesized GBS matrix for Spelling

matrix	DE	ESAA	SC	Priority vector
DE	1	1 2/3	2 5/7	0.5118
ESAA	3/5	1	1 1/3	0.2849
SC	3/8	3/4	1	0.2033
Sum (col)	**1.9684**	**3.4359**	**5.0143**	**1.0000**

Table 18. Synthesized GBS matrix for Context

matrix	LOW	RSW	LEM	STE	Priority vector
LOW	1	2 1/3	3	3 6/7	0.4806
RSW	3/7	1	2 1/3	3	0.2766
LEM	1/3	3/7	1	1 1/3	0.1383
STE	1/4	1/3	3/4	1	0.1045
Sum (col)	**2.0212**	**4.0952**	**7.0667**	**9.2208**	**1.0000**

Step 6: Calculate consistency ratio, CR

The CR in AHP is a measure used to evaluate the consistency of a decision-makers judgments or preferences in a pair-wise comparison matrix. In other words, CR calculation is to eliminate inconsistency in the judgments. It is the advantage of using AHP, whereby the consistency of the decisions can be revealed. To calculate the CR, AHP uses a mathematical concept called the eigenvalue. The matrix is transformed into a vector representing the dominant eigenvector, which indicates the relative weights of the criteria or alternatives.

Then, the CR is computed by comparing the consistency of the matrix with a randomly generated matrix - the original CR calculation by [33]. If the CR is low (usually less than 0.1), the decision-maker's judgments are consistent and reliable [44]. A higher CR indicates a higher level of inconsistency in the judgments, suggesting that the decision-maker should review and revise their judgments to ensure a more reliable decision.

Instead of using the equation [33] to calculate CR, this work are using the linear fit proposed by [45] to CR. The linear fit uses a consistency index which is simpler than Saaty's, λ1 and a very simple criterion for accepting or rejecting matrices [45]. The linear fit can be calculated using equation (4). In summary, the CR in AHP helps assess the reliability of a decision-maker's judgments by comparing the consistency of their pair-wise comparison matrix with a random matrix. It provides an indication of the internal consistency of the judgments and helps ensure more robust decision-making.

$$CR = \frac{\lambda - \mathrm{n}}{2.7699 \cdot n - 4.3513 - n} \quad (4)$$

Table 19 shows the consistency result for the matrix $A_{ss\_overall\_goal}$. The values of the dominant eigenvalue, λ_1 is derived equation using the power method.

2024, 21(2 Special Issue): 0662-0683
https://doi.org/10.21123/bsj.2024.9750
P-ISSN: 2078-8665 - E-ISSN: 2411-7986
Baghdad Science Journal

Then, the value of CR is calculated based on equation (4). As the value of CR is less than 0.1, the judgments are acceptable. It indicates that the weight created from the AHP process is reliable. If CR > 0.1, the judgment matrix is inconsistent. Judgments should be reviewed and improved to obtain a consistent matrix.

Table 19. The SS consistency test for the overall goal

matrix	P	S	N	C	Priority vector	n	λ_1	CR	%
P	1	4	5	7	0.5907	6	6.3157	0.0504	5.04%
S	1/4	1	5	4	0.2593	3	3.0037	0.0039	0.4%
N	1/5	1/5	1	2	0.0899	-	-	-	-
C	1/7	1/4	1/2	1	0.0601	4	4.2099	0.0769	7.7%

Table 20. The consistency test for SS matrix Punctuation

matrix	RU	RH	RUM	RE	RC	RPC	Priority vector
RU	1	3	4	5	6	9	0.4410
RH	1/3	1	2	3	4	7	0.2174
RUM	1/4	1/2	1	3	5	7	0.1746
RE	1/5	1/3	1/3	1	2	6	0.0905
RC	1/6	1/4	1/5	1/2	1	2	0.0494
RPC	1/9	1/7	1/7	1/6	1/2	1	0.0272
						λ_1	**6.3157**
						CR	**0.0503**

Proceeding to the subsequent level of criteria (sub-criteria and alternatives) is imperative to maintain the consistency test as an ongoing procedure. Tables 20 through 22 in the present study serve to conduct a consistency test for the sub-criteria and alternatives under investigation. Based on the analysis, it can be concluded that the values of the CR for all sub-criteria and alternatives are below the threshold of 0.1. It indicates that the judgments made in this study are deemed acceptable. Conducting a comprehensive analysis is imperative to evaluate and compare each criterion and thoroughly assess and compare each alternative. All alternatives are compared by evaluating them against a set of 14 criteria, as outlined in Table 23.

Table 21. The consistency test for SS matrix Spelling

matrix	DE	ESAA	SC	Priority vector
DE	1	3	5	0.6483
ESAA	1/3	1	2	0.2296
SC	1/5	1/2	1	0.1220
			λ_1	**3.0037**
			CR	**0.0038606**

Table 22. The consistency test for SS matrix Context

matrix	LOW	RSW	LEM	STE	Priority vector
LOW	1	4	5	6	0.5849
RSW	1/4	1	4	5	0.2591
LEM	1/5	1/4	1	2	0.0948
STE	1/6	1/5	1/2	1	0.0613
				λ_1	**4.2099**
				CR	**0.0769343**

2024, 21(2 Special Issue): 0662-0683
https://doi.org/10.21123/bsj.2024.9750
P-ISSN: 2078-8665 - E-ISSN: 2411-7986
Baghdad Science Journal

Table 23. The SS consistency test for the alternatives

PRIORITY VECTOR/ EIGENVECTOR													
GOAL													
	P						S			N	C		
	RU	RH	RUM	RE	RC	RPC	DE	ESAA	SC		LOW	RSW	LEM	STE
Twitter	0.333	0.5	0.778	0.333	0.648	0.625	0.615	0.682	0.297	0.333	0.333	0.333	0.333	0.333
Facebook	0.333	0.25	0.111	0.333	0.122	0.137	0.117	0.082	0.163	0.333	0.333	0.333	0.333	0.333
YouTube	0.333	0.25	0.111	0.333	0.23	0.238	0.267	0.236	0.54	0.333	0.333	0.333	0.333	0.333
CONSISTENCY TEST														
λ_1	3.000	3	3	3	3.004	3.018	3.067	3.002	3.009	3	3	3	3	3
n	3	3	3	3	3	3	3	3	3	3	3	3	3	3
CR	0	0	0	0	0.004	0.019	0.070	0.002	0.009	0	0	0	0	0
%	0%	0%	0%	0%	0.4%	1.9%	7.0%	0.2%	1%	0%	0%	0%	0%	0%

The consistency outcome for the matrix $A_{gbs\_overall\_goal}$ is displayed in Table 24. The power method is used to determine the values of the dominant eigenvalue, λ_1. The value of CR is then determined using equation (4). The judgments are acceptable because the value of CR is less than 0.1. It suggests that the weight produced by the AHP technique is trustworthy. The judgment matrix is incoherent if CR exceeds 0.1. To create a consistent matrix, judgments need to be reviewed and refined.

Table 24. The GBS consistency test for the overall goal

matrix	P	S	N	C	Priority vector	n	λ_1	CR	%
P	1	2 1/3	3	5	0.4978	6	**6.0900**	**0.0144**	**1.44%**
S	3/7	1	3	2 1/3	0.2723	3	**3.0060**	**0.0063**	**0.63%**
N	1/3	1/3	1	1 1/3	0.1287	-	-	-	-
C	1/5	3/7	3/4	1	0.1012	4	**4.0450**	**0.0165**	**1.65%**

The consistency test is carried out for the sub-criteria and alternatives that comprise the following criteria level. The consistency test for the sub-criteria and alternatives is represented by the elements in Tables 25 through Table 27. The judgments are acceptable because the value of CR for all sub-criteria and alternatives is less than 0.1. It is required to compare each alternative in addition to each criterion. All alternatives are analyzed using the 14 criteria listed in Table 28 because there are 14 criteria.

Table 25. The consistency test for GBS matrix Punctuation

matrix	RU	RH	RUM	RE	RC	RPC	Priority vector
RU	1	2	2 5/8	3 2/5	4 1/3	9	0.3808
RH	1/2	1	1 1/2	2	2 5/8	5 1/2	0.2118
RUM	3/8	2/3	1	2	3 2/5	5 1/2	0.1878
RE	2/7	1/2	1/2	1	1 1/2	4 1/3	0.1128
RC	1/4	3/8	2/7	2/3	1	1 1/2	0.0697
RPC	1/9	1/5	1/5	1/4	2/3	1	0.0371
						λ_1	**6.0900**
						CR	**0.0144**

2024, 21(2 Special Issue): 0662-0683
https://doi.org/10.21123/bsj.2024.9750
P-ISSN: 2078-8665 - E-ISSN: 2411-7986
Baghdad Science Journal

Table 26. The consistency test for GBS matrix Spelling

matrix	DE	ESAA	SC	Priority vector
DE	1	1 2/3	2 5/7	0.5118
ESAA	3/5	1	1 1/3	0.2849
SC	3/8	3/4	1	0.2033
			λ_1	**3.006**
			CR	**0.0063**

Table 27. The consistency test for GBS matrix Context

matrix	LOW	RSW	LEM	STE	Priority vector
LOW	1	2 1/3	3	3 6/7	0.4806
RSW	3/7	1	2 1/3	3	0.2766
LEM	1/3	3/7	1	1 1/3	0.1383
STE	1/4	1/3	3/4	1	0.1045
				λ_1	**4.045**
				CR	**0.0165**

Table 28. The GBS consistency test for the alternatives

	PRIORITY VECTOR/ EIGENVECTOR													
	GOAL													
	P						S			N	C			
	RU	RH	RUM	RE	RC	RPC	DE	ESAA	SC		LOW	RSW	LEM	STE
Twitter	0.33	0.394	0.696	0.33	0.512	0.483	0.479	0.599	0.326	0.33	0.33	0.33	0.33	0.33
Facebook	0.33	0.303	0.152	0.33	0.204	0.226	0.206	0.126	0.252	0.33	0.33	0.33	0.33	0.33
YouTube	0.33	0.303	0.152	0.33	0.285	0.292	0.315	0.276	0.422	0.33	0.33	0.33	0.33	0.33
	CONSISTENCY TEST													
λ_1	3	3.001	3.008	3	3.008	3	3.008	3.082	3.002	3	3	3	3	3
n	3	3	3	3	3	3	3	3	3	3	3	3	3	3
CR	0	0.0010	0.0083	0	0.0083	0	0.0083	0.0856	0.0021	0	0	0	0	0
%	0%	0.1%	0.8%	0%	0.8%	1.9%	0.8%	8.6%	0.2%	0%	0%	0%	0%	0%

Step 7: Develop Overall Polarity Ranking

The overall polarity ranking refers to determining the relative importance or preference of different elements or criteria in decision-making. AHP allows decision-makers to compare the importance of various factors and make informed decisions based on their relative priorities. After the consistency calculation for all levels is completed, further calculation of the overall priority vector to prioritize pre-processing techniques must be performed. The overall polarity rating can be calculated by considering the relative weights allocated to each criterion once the priorities of all criteria have been established and validated for consistency. The criterion with the highest weight is considered the most important, while the one with the lowest is considered the least important. The brief findings will be discussed in the Results and Discussions section.

Step 8: Select the best alternatives

The final stage of the decision-making process involves choosing the optimal alternative, considering the established criteria, as well as the priority vector and priority vector assigned to each alternative. Evaluate alternatives by assessing each alternative against each criterion. Assign judgment score value to the alternatives based on their performance or suitability for each criterion. The alternative score can be calculated by multiplying the weight of each criterion by the score of the corresponding alternative for that criterion. Sum up these values for each alternative to calculate an overall score. Lastly, select the best alternative by comparing the overall scores of alternatives. The

2024, 21(2 Special Issue): 0662-0683
https://doi.org/10.21123/bsj.2024.9750
P-ISSN: 2078-8665 - E-ISSN: 2411-7986
Baghdad Science Journal

alternative with the highest overall score is considered the best choice based on the criteria and assigned weights. The brief findings will be discussed in the Results and Discussions section.

Results and Discussion

The result in Table 29 presents all priority vectors for criteria, sub-criteria, and alternatives. It also shows the overall priority vector of the alternatives for the criteria. The elements in Table 30 represent the overall priority vector for three social media services alternatives to the sub-criteria. Determining the overall priority vector involves the multiplication of the priority vector associated with the alternatives with the vector representing the priority of the sub-criteria. An illustrative demonstration of the comprehensive priority calculation is presented below:

$$(0.3330 \times 0.4410) + (0.5000 \times 0.2174) + (0.7780 \times 0.1746) + (0.3330 \times 0.0905) + (0.6480 \times 0.0494) + (0.6250 \times 0.0272) = 0.4705$$

Table 29. All priority vectors for criteria, sub-criteria, and alternatives using SS

Level 1	Level 2		Level 3		Level 4		
Goal	**Criteria**		**Sub-criteria**		**Twitter**	**Facebook**	**YouTube**
prioritized text detergent model (selection of text pre-processing techniques for social media sentiment analysis)	**P**	0.5907	**RU**	0.441	0.3330	0.3330	0.3330
			RH	0.2174	0.5000	0.2500	0.2500
			RUM	0.1746	0.7780	0.1110	0.1110
			RE	0.0905	0.3330	0.3330	0.3330
			RC	0.0494	0.6480	0.1220	0.2300
			RPC	0.0272	0.6250	0.1370	0.2380
	S	0.2593	**DE**	0.6483	0.6150	0.1170	0.2670
			ESAA	0.2296	0.6820	0.0820	0.2360
			SC	0.122	0.2970	0.1630	0.5400
	N	0.0899			0.3330	0.3330	0.3330
	C	0.0601	**LOW**	0.5849	0.3330	0.3330	0.3330
			RSW	0.2591	0.3330	0.3330	0.3330
			LEM	0.0948	0.3330	0.3330	0.3330
			STE	0.0613	0.3330	0.3330	0.3330

Table 30. Overall priority vectors sub-criteria for the criteria using SS

	OVERALL PRIORITY VECTOR			
	P	**S**	**N**	**C**
Twitter	0.4705	0.5915	0.3330	0.3330
Facebook	0.2605	0.1146	0.3330	0.3330
YouTube	0.2686	0.2932	0.3330	0.3330

The data presented in Table 31 represents the prioritization of the alternatives to the criteria, as determined by the overall priority vector. The derivation of the overall priority vector involves the multiplication of the priority vector for the alternatives with the priority vector of the criteria. An illustrative demonstration of the comprehensive priority calculation is provided below:

$$(0.4705 \times 0.5907) + (0.5915 \times 0.2593) + (0.3330 \times 0.0899) + (0.3330 \times 0.0601) = 0.4813$$

2024, 21(2 Special Issue): 0662-0683
https://doi.org/10.21123/bsj.2024.9750
P-ISSN: 2078-8665 - E-ISSN: 2411-7986
Baghdad Science Journal

Table 31. Overall priority vector for the alternatives for the criteria using SS

	OVERALL PRIORITY VECTOR				OVERALL PRIORITY
	P	**S**	**N**	**C**	
	0.5907	0.2593	0.0899	0.0601	
Twitter	0.4705	0.5915	0.3330	0.3330	0.4813
Facebook	0.2605	0.1146	0.3330	0.3330	0.2335
YouTube	0.2686	0.2932	0.3330	0.3330	0.2846

To continue further, by using GBS, further calculation of the overall priority vector to select the best alternatives or criteria has also already been performed. The result in Table 32 presents the priority vectors for criteria, sub-criteria, and alternatives. It also shows the overall priority vector of the alternatives for the criteria. The elements in Table 33 represent the overall priority vector for three social media services alternatives to the sub-criteria. The overall priority vector can be determined by multiplying the priority vector associated with the alternatives with the vector representing the priority of the sub-criteria. An illustrative demonstration of the comprehensive priority calculation is presented below:

$$(0.3330 \times 0.3808) + (0.3939 \times 0.2118) + (0.6961 \times 0.1878) + (0.3330 \times 0.1128) + (0.5117 \times 0.0697) + (0.4827 \times 0.0371) = 0.4321$$

The elements in Table 34 show the overall priority vector of the alternatives to the criteria. The overall priority vector can be obtained by multiplying the priority vector for the alternatives by the priority vector of the criteria. An example of the overall priority calculation is as follows:

$$(0.4321 \times 0.4978) + (0.4822 \times 0.2723) + (0.3330 \times 0.1287) + (0.3330 \times 0.1012) = 0.4229$$

Table 32. All priority vectors for criteria, sub-criteria, and alternatives using GBS

Level 1	**Level 2**		**Level 3**		**Level 4**		
Goal	**Criteria**		**Sub-criteria**		**Twitter**	**Facebook**	**YouTube**
prioritized text detergent model (selection of text pre-processing techniques for social media sentiment analysis)	**P**	0.4978	**RU**	0.3808	0.3330	0.3330	0.3330
			RH	0.2118	0.3939	0.3031	0.3031
			RUM	0.1878	0.6961	0.1519	0.1519
			RE	0.1128	0.3330	0.3330	0.3330
			RC	0.0697	0.5117	0.2035	0.2848
			RPC	0.0371	0.4827	0.2257	0.2916
	S	0.2723	**DE**	0.5118	0.4793	0.2062	0.3146
			ESAA	0.2849	0.5987	0.1257	0.2756
			SC	0.2033	0.3262	0.2519	0.4219
	N		0.1287		0.3330	0.3330	0.3330
	C	0.1012	**LOW**	0.4806	0.3330	0.3330	0.3330
			RSW	0.2766	0.3330	0.3330	0.3330
			LEM	0.1383	0.3330	0.3330	0.3330
			STE	0.1045	0.3330	0.3330	0.3330

2024, 21(2 Special Issue): 0662-0683
https://doi.org/10.21123/bsj.2024.9750
P-ISSN: 2078-8665 - E-ISSN: 2411-7986
Baghdad Science Journal

Table 33. Overall priority vectors sub-criteria for the criteria using GBS

	OVERALL PRIORITY VECTOR			
	P	**S**	**N**	**C**
Twitter	0.4321	0.4822	0.3330	0.3330
Facebook	0.2797	0.1925	0.3330	0.3330
YouTube	0.2878	0.3253	0.3330	0.3330

Table 34. Overall priority vector for the alternatives for the criteria GBS

	OVERALL PRIORITY VECTOR				OVERALL PRIORITY
	P	**S**	**N**	**C**	
	0.4978	0.2723	0.1287	0.1012	
Twitter	0.4321	0.4822	0.3330	0.3330	0.4229
Facebook	0.2797	0.1925	0.3330	0.3330	0.2682
YouTube	0.2878	0.3253	0.3330	0.3330	0.3084

The contribution of the findings in this work is remarkable. While researching the previous social media sentiment analysis studies, it was observed that a wide variety of criteria (pre-processing techniques) were used in de-noising unstructured social media data. In the first step, while defining the objective of the AHP, a list of criteria suitable for the pre-processing stage in social media sentiment analysis is checked. The pre-processing techniques list helps data analyst to steer the data pre-processing stage. Considering whole analytic hierarchy processes, as referred to in Fig. 5, it mapped perfectly with the degree of noise as tabulated in Table 1. It is understood that Punctuation is the most essential criterion in the pre-processing stage of sentiment analysis. It is followed by Spelling, then Number, and lastly, Context criteria. Based on Fig. 2, the sub-criteria are also constructed hierarchically.

In addition, several AHP studies conducted with the purpose of the determination of the most consistent judgment scale were discovered. In this study, it is essential to note that two distinct judgment scales were utilized for accuracy. The initial element in the series under consideration is denoted as the SS, which exclusively comprises whole numbers. The second judgment scale, the GBS, encompassed a combination of integer and decimal numbers. This study aimed to investigate whether the utilization of two distinct AHP scales would result in any significant differences in the decision-making procedure.

As shown in Fig. 3, the distribution of the overall priority vector, i.e., weights achieved by SS and GBS, varies considerably. The result shows that Twitter was the most important and urgent social media service to be pre-processed in sentiment analysis. Facebook was the least important. The observed variation in weight distribution across scales (Twitter-YouTube-Facebook) can be attributed to the inconsistency gained in the pair-wise comparisons.

The computation of CR values was conducted to assess the degree of consistency in the pair-wise comparisons performed. According to the observations made in Fig. 4, it can be inferred that the relative differences observed in the decision matrix during the sensitivity analysis are significantly significant. It suggests the situation may be overstated, potentially drifting from a realistic scenario. According to the findings

2024, 21(2 Special Issue): 0662-0683
https://doi.org/10.21123/bsj.2024.9750
P-ISSN: 2078-8665 - E-ISSN: 2411-7986
Baghdad Science Journal

presented in Fig. 4, it can be observed that the scale of GBS exhibits a higher level of consistency in its results when compared to the SS. The analysis reveals that the results obtained from the SS method do not exhibit a perfect and consistent pair-wise comparison.

However, it is noteworthy that most CR values, except for the Context criteria results, demonstrate either perfect or nearly perfect consistent pair-wise comparisons when employing the GBS method. The mean CR value achieved by the SS method was found to be 0.0437, whereas the mean CR value obtained through the GBS method was observed to be 0.0372. The findings in Fig. 4 suggest that the pairwise comparisons conducted using the GBS demonstrate a higher degree of closeness to the consistent margin than those made with the SS.

The observed phenomenon of lower CR values obtained from the GBS compared to the SS does not provide strong proof supporting the superiority of the GBS as a preferred option. The observed outcome aligns with expectations, as the values within the GBS scale tend to exhibit a higher degree of closeness to one another. This characteristic is the reason behind its designation as a balanced scale. Therefore, it can be inferred that there is a tendency for individuals to attain improved levels of consistency in their performance. The objectives of this study encompass enhancing consistency and fostering more representative judgments.

As depicted in Fig. 5, the utilization of the GBS has resulted in a noteworthy reduction in the weight ratios of the alternatives, as evidenced by the pair-wise comparison matrix. The relative ratios of the Punctuation and Spelling criteria with SS and GBS are calculated as follows: the ratio for SS is 2.2781 (0.5907 divided by 0.2593), while the ratio for GBS is 1.8281 (0.4978 divided by 0.2723). Despite being relatively minor, the observed differences in the weight ratios do not affect the ranking order. However, it is essential to note that the AHP exclusively considers the relative ratios. Consequently, the comparison highlights a scenario where different scales yield substantially different outcomes. For more information, examine Fig. 6, 7, and 8 for the priority vector for each sub-criteria, such as Punctuation, Spelling, and Context.

The significance of weight dispersion lies in its ability to enhance the application of pre-processing techniques for sentiment analysis on social media data. By understanding and utilizing these weights, data analysts can effectively evaluate and categorize social media data based on specific criteria, thereby improving the accuracy and efficiency of sentiment analysis. During this investigation, it was noted that there were specific subtle differences when comparing two AHP judgment scales. However, it is essential to highlight that despite these variances, both methodologies resulted in the same rankings. Using integers in the SS system offers a heightened convenience for decision-makers when interpreting judgments and conducting transactions.

Selecting the correct text pre-processing techniques when analyzing sentiment across social media services is crucial. Implementing appropriate evaluation and decision tools should be considered at the pre-processing step, which involves many complex decision-making tasks. Several limitations were identified during this study, including the restricted availability of experts to determine the individual weights for the criteria. In light of the limitations, it is advisable to employ software solutions that compute numerous alternatives, such as AHP-OS [40]. This method not only provides support and validation for decisions made but also empowers decision-makers to rationalize their choices and simulate potential outcomes. The present study's analysis proves the significance of the AHP in facilitating rational decision-making. The utilization of the AHP technique during the pre-processing phase can be regarded as an exemplary illustration or a guidebook for implementing the method in the following stages of the sentiment analysis process. From a particular perspective, the article contributes to the broader field of Big Data Analytics.

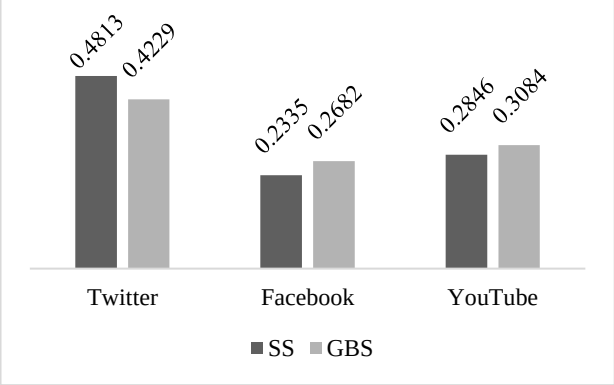

Figure 3. Overall priority vector

2024, 21(2 Special Issue): 0662-0683
https://doi.org/10.21123/bsj.2024.9750
P-ISSN: 2078-8665 - E-ISSN: 2411-7986
Baghdad Science Journal

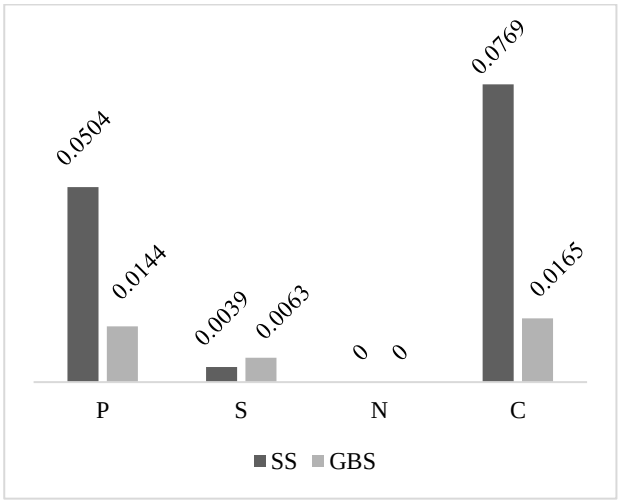

Figure 4. CR for overall goal

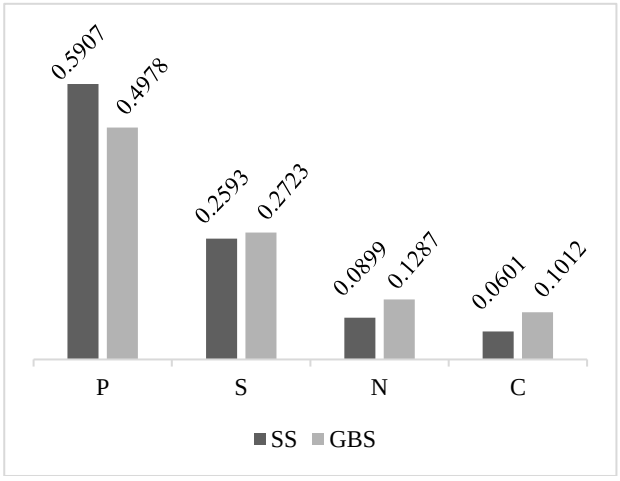

Figure 5. Priority vector by criteria

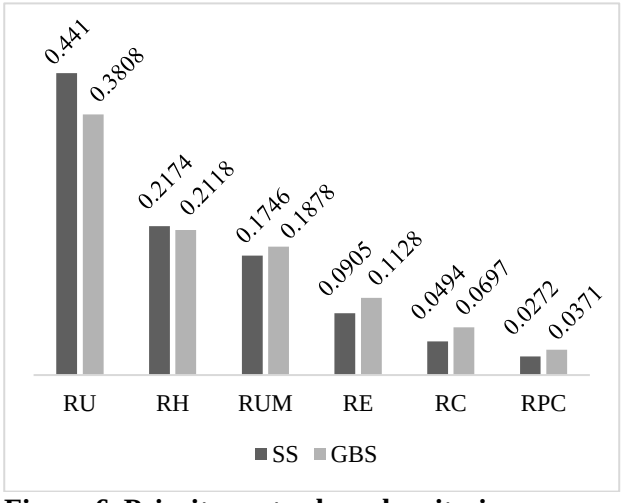

Figure 6. Priority vector by sub-criteria Punctuation

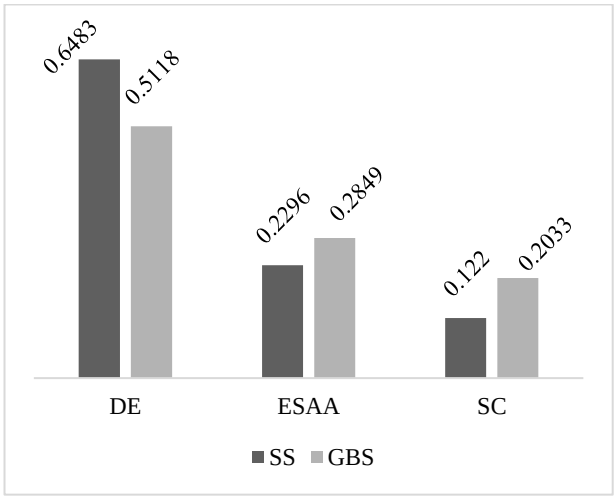

Figure 7. Priority vector by sub-criteria Spelling

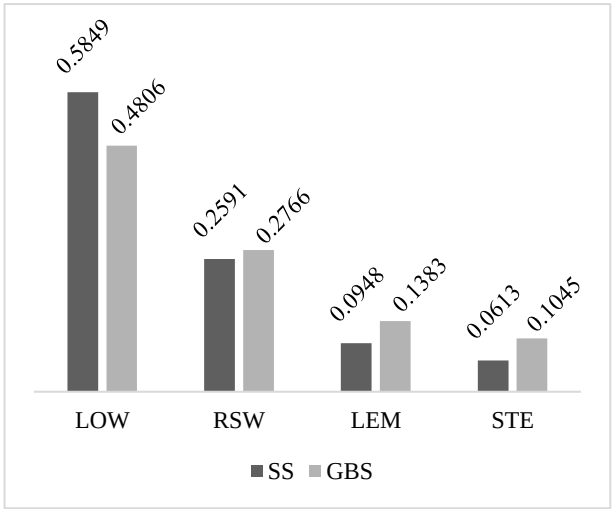

Figure 8. Priority vector by sub-criteria Context

2024, 21(2 Special Issue): 0662-0683
https://doi.org/10.21123/bsj.2024.9750
P-ISSN: 2078-8665 - E-ISSN: 2411-7986
Baghdad Science Journal

Conclusion

The contribution of this paper is aimed at using the AHP to assess essential pre-processing techniques and select the optimal alternative. AHP is among the pioneer of MCDM methods. The method derives ratio scales from paired comparisons. The ratio scales are derived from the principal eigenvectors, and the consistency index is derived from the principal eigenvalue. This study applied SS and GBS for judgment in this study. Both judgment scales exhibit high reliability, indicating their suitability for use in various research contexts. Furthermore, the criteria employed in the pre-processing technique case study can serve as a valuable reference for future investigations seeking to achieve similar objectives within different stages of sentiment analysis. The present study prioritized text pre-processing techniques for sentiment analysis on social media data. Multiple social media services, including Twitter, Facebook, and YouTube, are considered potential alternatives in this study. This study can also shed light on the literature on choosing the essential text pre-processing techniques on other social media services. Interesting results can be obtained if different MCDM techniques are used, especially in selecting text pre-processing techniques for sentiment analysis over other social media services.

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

U. H., H. Z. designed the study, performed the computations, analyzed the data, and discussed the results of the paper. R. I., S. A., and I. I. K., supervised and revised the findings of this work, and discussed the results and contributed to the final manuscript.

References

1. Mutasher WG, Aljuboori AF. New and Existing Approaches Reviewing of Big Data Analysis with Hadoop Tools. Baghdad Sci J. 2022;19(4):887–98. https://doi.org/ 10.21123/bsj.2022.19.4.0887
2. Al-Bakri NF, Yonan JF, Sadiq AT, Abid AS. Tourism companies assessment via social media using sentiment analysis. Baghdad Sci J. 2022;19(2):422–9. https://doi.org/10.21123/BSJ.2022.19.2.0422
3. Singh NK, Tomar DS, Sangaiah AK. Sentiment analysis: a review and comparative analysis over social media. J Ambient Intell Humaniz Comput. 2020;11(1):97–117. https://doi.org/ 10.1007/s12652-018-0862-8
4. Shehab N, Badawy M, Arafat H. Big Data Analytics Concepts, Technologies Challenges, and Opportunities. In: Intelligent Transport Systems and Its Challenges. Springer Cham. 2019. https://doi.org/ 10.1007/978-3-030-31129-2_9
5. Triguero I, García-Gil D, Maillo J, Luengo J, García S, Herrera F. Transforming big data into smart data: An insight on the use of the k-nearest neighbors algorithm to obtain quality data. Wiley Interdiscip Rev Data Min Knowl Discov. 2019;9(2):1–24. https://doi.org/10.1002/widm.1289
6. Ramírez-Gallego S, Krawczyk B, García S, Woźniak M, Herrera F. A survey on data preprocessing for data stream mining: Current status and future directions. Neurocomputing. 2017;239:39–57. https://doi.org/10.1016/j.neucom.2017.01.078
7. Murshed BAH, Abawajy J, Mallappa S, Saif MAN, Al-Ghuribi SM, Ghanem FA. Enhancing Big Social Media Data Quality for Use in Short-Text Topic Modeling. IEEE Access. 2022;10(October):105328–51. https://doi.org/10.1109/ACCESS.2022.3211396
8. Yan X, Li Y, Fan W. Identifying domain relevant user generated content through noise reduction: a test in a Chinese stock discussion forum. Inf Discov

Deliv. 2017;45(4):181–93. https://doi.org/10.1108/IDD-04-2017-0043
9. Woo HS, Kim JM, Lee WG. Validation of Text Data Preprocessing Using a Neural Network Model. Math Probl Eng. 2020;2020. https://doi.org/10.1155/2020/1958149
10. Ali K. Sentiment Analysis as a Service. RMIT University. 2019.
11. Ali K, Dong H, Bouguettaya A, Erradi A, Hadjidj R. Sentiment analysis as a service: a social media based sentiment analysis framework. In: 2017 IEEE International Conference on Web Services (ICWS). 2017. p. 660–7. https://doi.org/10.1109/ICWS.2017.79
12. Saggi MK, Jain S. A survey towards an integration of big data analytics to big insights for value-creation. Inf Process Manag. 2018;54(5):758–90. https://doi.org/10.1016/j.ipm.2018.01.010
13. Alaoui I El, Gahi Y, Messoussi R. Full consideration of big data characteristics in sentiment analysis context. In: 2019 IEEE 4th International Conference on Cloud Computing and Big Data Analytics, ICCCBDA 2019. IEEE; 2019. p. 126–30. https://doi.org/10.1109/ICCCBDA.2019.8725728
14. Siriweera THAS, Paik I, Kumara BTGS. Constraint-Driven Dynamic Workflow for Automation of Big Data Analytics Based on GraphPlan. In: IEEE 24th International Conference on Web Services, ICWS 2017. 2017. p. 357–64. https://doi.org/10.1109/ICWS.2017.120
15. Melo PF, Dalip DH, Junior MM, Gonçalves MA, Benevenuto F. 10SENT: A stable sentiment analysis method based on the combination of off-the-shelf approaches. J Assoc Inf Sci Technol. 2019;70(3):242–55. https://doi.org/10.1002/asi.24117
16. Pradha S, Halgamuge MN, Tran Quoc Vinh N. Effective text data preprocessing technique for sentiment analysis in social media data. In: Proceedings of 2019 11th International Conference on Knowledge and Systems Engineering, KSE 2019. IEEE; 2019. https://doi.org/10.1109/KSE.2019.8919368
17. Hair Zaki UH, Ibrahim R, Abd Halim S, Kamsani II. Text Detergent: The Systematic Combination of Text Pre-processing Techniques for Social Media Sentiment Analysis. LNDECT. 2022. p. 50–61. https://doi.org/10.1007/978-3-030-98741-1_5
18. Hair Zaki UH, Ibrahim R, Abd Halim S. A Social Media Services Analysis. Int J Adv Trends Comput Sci Eng. 2019;8(1.6):69–75. https://doi.org/10.30534/ijatcse/2019/1181.62019
19. Naseem U, Razzak I, Eklund PW. A survey of pre-processing techniques to improve short-text quality: a case study on hate speech detection on twitter. Multimed Tools Appl. 2020. https://doi.org/10.1007/s11042-020-10082-6
20. Baykasoğlu A, Gölcük İ. A dynamic multiple attribute decision making model with learning of fuzzy cognitive maps. Comput Ind Eng. 2019;135(April):1063–76. https://doi.org/10.1016/j.cie.2019.06.032
21. Yalcin AS, Kilic HS, Delen D. The use of multi-criteria decision-making methods in business analytics: A comprehensive literature review. Technol Forecast Soc Change. 2022;174(September 2021):121193. https://doi.org/10.1016/j.techfore.2021.121193
22. Tufail H, Qasim I, Masood MF, Tanvir S, Butt WH. Towards the selection of Optimum Requirements Prioritization Technique: A Comparative Analysis. In: International Conference on Information Management (ICIM). IEEE; 2019. p. 227–31. https://doi.org/10.1109/INFOMAN.2019.8714709
23. Sufian M, Khan Z, Rehman S, Haider Butt W. A systematic literature review: Software requirements prioritization techniques. Proc - 2018 Int Conf Front Inf Technol FIT 2018. 2019;35–40. https://doi.org/10.1109/FIT.2018.00014
24. Tüzemen A. Which YouTuber Should Be Followed? A Comparison Based Delphi-AHP-TOPSIS. Int J Contemp Econ Adm Sci. 2020;X(2). https://doi.org/10.5281/zenodo.4430009
25. Afify EA, Eldin AS, Khedr AE. Facebook Profile Credibility Detection using Machine and Deep Learning Techniques based on User's Sentiment Response on Status Message. Int J Adv Comput Sci Appl. 2020;11(12):622–37. https://doi.org/10.14569/IJACSA.2020.0111273
26. Yenkar PP, Sawarkar SD. A novel ensemble approach based on MCC and MCDM methods for prioritizing tweets mentioning urban issues in smart city. Kybernetes. 2022. https://doi.org/10.1108/K-08-2021-0785
27. Al-Yazidi SA, Berri J, Hassan MM. Novel hybrid model for organizations' reputation in online social networks. J King Saud Univ - Comput Inf Sci. 2022;34(8):5305–17. https://doi.org/10.1016/j.jksuci.2022.01.006
28. A. M, Gandhi GM. Framework for Social Media Analytics based on Multi-Criteria Decision Making (MCDM) model. Multimed Tools Appl. 2020. https://doi.org/10.1007/s11042-019-7470-2
29. Wu Z, Shen Y, Wang H. Assessing urban areas' vulnerability to flood disaster based on text data: A case study in Zhengzhou City. Sustain. 2019;11(17). https://doi.org/10.3390/su11174548
30. Ye Y, Zhao Y, Shang J, Zhang L. A hybrid IT framework for identifying high-quality physicians using big data analytics. Int J Inf Manage. 2019;47(August 2018):65–75. https://doi.org/10.1016/j.ijinfomgt.2019.01.005
31. Saifullah S. Fuzzy-AHP approach using Normalized Decision Matrix on Tourism Trend Ranking based-on Social Media. J Inform.

2019;13(2):16. https://doi.org/10.26555/jifo.v13i2.a15268
32. Kaur R, Singh S, Kumar H. AuthCom: Authorship verification and compromised account detection in online social networks using AHP-TOPSIS embedded profiling based technique. Expert Syst Appl. 2018;113:397–414. https://doi.org/10.1016/j.eswa.2018.07.011
33. Saaty TL. Decision Making with the Analytic Hierarchy Process. Int J Serv Sci. 2008.https://doi.org/10.1504/IJSSCI.2008.017590
34. Adenle YA, Chan EHW, Sun Y, Chau CK. Modifiable campus-wide appraisal model (MOCAM) for sustainability in higher education institutions. Sustain. 2020;12(17). https://doi.org/10.3390/SU12176821
35. Sailunaz K, Alhajj R. Emotion and sentiment analysis from Twitter text. J Comput Sci. 2019;36:101003. https://doi.org/10.1016/j.jocs.2019.05.009
36. Lamirán-Palomares JM, Baviera T, Baviera-Puig A. Sports influencers on twitter. Analysis and comparative study of track cycling world cups 2016 and 2018. Soc Sci. 2020;9(10):1–23. https://doi.org/10.3390/socsci9100169
37. Zhou F, Lim MK, He Y, Pratap S. What attracts vehicle consumers' buying: A Saaty scale-based VIKOR (SSC-VIKOR) approach from after-sales textual perspective.Ind Manag Data Syst. 2020;120(1):57–78. https://doi.org/ 10.1108/IMDS-01-2019-0034
38. Bueno I, Carrasco RA, Ureña R, Herrera-Viedma E. A business context aware decision-making approach for selecting the most appropriate sentiment analysis technique in e-marketing situations. Inf Sci (Ny). 2022;589:300–20. https://doi.org/10.1016/j.ins.2021.12.080
39. Klaus DG. Comparison of Judgment Scales of the Analytical Hierarchy Process - A New Approach. Int J Inf Technol Decis Mak. 2019;18(2):445–63. https://doi.org/10.1142/S0219622019500044
40. Goepel K. Implementation of an Online software tool for the Analytic Hierarchy Process (AHP-OS). Int J Anal Hierarchy Process. 2018;10(3):469–87. https://doi.org/10.13033/ijahp.v10i3.590
41. Zhang S, Kindlmann G. Diffusion tensor MRI visualization. In: Visualization Handbook. Elsevier Inc.; 2005. p. 327–40. https://doi.org/10.1016/B978-012387582-2/50018-6
42. Larson R, Falvo DC. Power method for approximation eigenvalues. In: Elementary Linear Algebra. Boston, New York: Houghton Mifflin Harcourt Publishing Company. 2009; p. 550–8. Chapter 10
43. Ford W. The Algebraic Eigenvalue Problem. In: Ford WBTNLA with A, editor. Numerical Linear Algebra with Applications using MATLAB. Boston: Academic Press; 2015. p. 379–438. https://doi.org/10.1016/B978-0-12-394435-1.00018-1
44. Badeel R, Subramaniam SK, Muhammed A, Hanapi ZM. A Multicriteria Decision-Making Framework for Access Point Selection in Hybrid LiFi/WiFi Networks Using Integrated AHP–VIKOR Technique. Sensors. 2023;23(3). https://doi.org/10.3390/s23031312
45. Alonso JA, Lamata MT. Consistency in the Analytic Hierarchy Process: A new approach. Int J Uncertainty, Fuzziness Knowledge-Based Syst. 2006;14(4):445–59. https://doi.org/10.1142/S0218488506004114

2024, 21(2 Special Issue): 0662-0683
https://doi.org/10.21123/bsj.2024.9750
P-ISSN: 2078-8665 - E-ISSN: 2411-7986
Baghdad Science Journal

منظف النصوص ذو الأولوية: مقارنة مقياسين للحكم لعملية التسلسل الهرمي التحليلي بشأن إعطاء الأولوية لتقنيات ما قبل المعالجة في تحليل المشاعر على وسائل التواصل الاجتماعي

اوم هانئ هير زكي[1]، روليانا إبراهيم[1]، شهليزا عبدالحاليم[2]، ازين عزتي كمسني[1]

[1]قسم الحوسبة التطبيقية، كلية الحوسبة، جامعة تكنولوجي، ماليزيا، جوهور باهرو، جوهور، ماليزيا.
[2]قسم هندسة البرمجيات، كلية الحوسبة، جامعة تكنولوجي، ماليزيا، جوهور باهرو، جوهور، ماليزيا.

الخلاصة

تستخدم معظم الشركات بيانات وسائل التواصل الاجتماعي للأعمال. يقوم تحليل المشاعر تلقائيًا بجمع التحليلات وتلخيص هذا النوع من البيانات. من الصعب إدارة بيانات وسائل التواصل الاجتماعي غير المنظمة. حيث تمثل البيانات المزعجة او غير متناسقة تحديًا لتحليل المشاعر. نظرًا لأن أكثر من 50٪ من عملية تحليل المشاعر هي معالجة مسبقة للبيانات، فإن معالجة بيانات الوسائط الاجتماعية الضخمة تمثل تحديًا أيضًا. إذا تم إجراء المعالجة المسبقة بشكل صحيح، فقد تتحسن دقة البيانات. أيضًا، يعتمد سير العمل في تحليل المشاعر بشكل كبير. نظرًا لعدم وجود تقنية معالجة مسبقة تعمل بشكل جيد في جميع المواقف أو مع جميع مصادر البيانات، فإن اختيار أهم المصادر أمر بالغ الأهمية. تحديد الأولويات هو أسلوب ممتاز لاختيار الأكثر أهمية.باعتبارها إحدى طرق اتخاذ القرار متعدد المعايير (MCDM)، تُفضل عملية التحليل الهرمي (AHP) التعامل مع تحديات صنع القرار المعقدة باستخدام عدة معايير. تم استخدام درجات نسبة الاتساق (CR) لفحص المقارنات الزوجية لتقييم AHP. استخدمت هذه الدراسة مقياسين للحصول على الحكم الأكثر اتساقًا. أولاً، مقياس حكم Saaty (SS)، ثم المقياس المتوازن المعمم (GBS). لقد تم التحقق فيما إذا كان هناك مقياسان مختلفان لحكم AHP سيؤثران على صنع القرار. المعايير الرئيسية لتحديد أولوية تقنيات المعالجة المسبقة في تحليل المشاعر هي علامات الترقيم والإملاء والرقم والسياق. تحتوى هذه المعايير الأربعة أيضا معايير فرعية. تكون مقارنات الزوجية GBS أقرب إلى قيمة CR من SS، مما يقلل من نسب وزن البدائل.تشرح هذه الورقة كيف يساعد AHP في اتخاذ القرار المنطقي. يمكن أن يكون تحديد أولوية تقنيات المعالجة المسبقة باستخدام AHP نموذجًا لمراحل تحليل المشاعر الأخرى. باختصار، تضيف هذه الورقة مساهمة أخرى إلى مجال تحليلات البيانات الضخمة.

الكلمات المفتاحية: صنع القرار متعدد المعايير، تحديد الأولويات، عملية التسلسل الهرمي التحليلي، وسائل التواصل الاجتماعي، تحليل المشاعر، المعالجة المسبقة للبيانات، تقنية ما قبل المعالجة.