Varia

ESSACHESS – Journal for Communication Studies

Critical Minds, Tolerant Voices: Media and Information Literacy as a Tool against Hate Speech

ESSACHESS –
Journal for Communication Studies
Volume 19 Issue 1(37), p. 175-198
© The Author(s) 2026
Reprints and Permission:
© ESSACHESS
https://www.essachess.com/
DOI: 10.21409/essachess.1775-352x

Cite : AL-ZOU'BI, R. (2026). Critical Minds, Tolerant Voices: Media and Information Literacy as a Tool against Hate Speech . *ESSACHESS*. 19(37): 175-198. https://doi.org/10.21409/FC4D-XH97

Reem AL-ZOU'BI
Associate Professor, Al al-Bayt University
JORDAN
e-mail: rzoubi@aabu.edu.jo

Abstract: The impact of media and information literacy (MIL) on university students' awareness of hate speech is investigated in this study. A quasi-experimental time-series design was employed. Data were collected from a convenience sample of 156 students from humanities and science disciplines at Al al-Bayt University in Jordan, through pre- and post-tests and qualitative responses. The results showed an improvement in hate speech awareness after the MIL intervention, with no statistically significant differences attributable to gender or academic specialization. Qualitative analysis revealed that students' definition of hate speech shifted from a superficial to a more critical one, with the speaker's intent emerging as a key indicator of hate speech in media contexts. The researcher recommends integrating MIL into higher education to promote critical awareness of hate speech in media environments.

Keywords: hate speech, media and information literacy, higher education, hate speech awareness, Jordan.

***

Article received December 1, 2025. Article accepted May 20, 2026.
Conflict of Interest: The author declares no conflict of interest.

Esprits critiques, voix tolérantes: la littératie médiatique et informationnelle comme outil de lutte contre les discours de haine

Résumé: L'impact de la littératie médiatique et informationnelle (LMI) sur la sensibilisation des étudiants universitaires aux discours de haine est étudié dans cette étude. Une étude quasi-expérimentale de séries temporelles a été menée auprès d'un échantillon de commodité de 156 étudiants en sciences humaines et en sciences exactes de l'Université Al al-Bayt en Jordanie, au moyen d'un prétest et de deux post-tests, ainsi que de réponses qualitatives. Les résultats ont montré une meilleure sensibilisation aux discours de haine à la suite de l'intervention de la LMI, sans différence statistiquement significative liée au genre ou à la spécialisation. L'analyse qualitative a révélé que la définition des discours de haine par les étudiants est passée d'une approche superficielle à une approche critique, l'intention de l'orateur apparaissant comme l'indicateur clé de la présence de discours de haine dans les médias. Le chercheur recommande d'intégrer la LMI dans l'enseignement supérieur afin de promouvoir la conscience critique des discours de haine dans les environnements médiatiques.

Mots-clés: discours de haine, littératie médiatique et informationnelle, enseignement supérieur, sensibilisation au discours de haine, Jordanie.

***

Introduction

Legal scholars and philosophers have given considerable thought to the hate speech phenomenon. However, most of this attention has centered on presenting and critically evaluating arguments for the prohibition of hate speech, while less attention has been given to developing a comprehensive definition.

According to Ben Jilali (2022), Chetty and Alathur (2018), and Elmoubarak (2021), hate speech is defined as hostile and malicious speech that motivates prejudice and targets a person or group of people based on actual or perceived innate characteristics. It expresses discriminatory, intimidating, dismissive, hostile, and/or prejudiced attitudes toward those characteristics, which include gender, race, religion, color, national origin, disability, or sexual orientation. Thus, hate speech aims to harm and dehumanize targeted groups, and to encourage hostility toward them.

Hate speech is a broad and problematic concept due to its flexible nature, which makes it open to interpretation and manipulation. It can manifest in various forms across different societies without necessarily leading to direct violence. In contrast to this conceptual breadth, the term "dangerous speech" has emerged as a more specific concept, used to denote patterns of discourse that explicitly or implicitly aim to incite or amplify violence, particularly when one group perpetrates violence against other groups (UNESCO, 2015).

Building on these definitions, the study defines hate speech as any verbal or nonverbal expression that is insulting or humiliating to an individual or group and

incites harm against them, based on their beliefs, races, religions, genders, nationalities, sexual identities, disabilities, or statuses as refugees.

Understanding hate speech requires more than just defining the phenomenon itself; it also includes understanding how individuals perceive this speech and how it operates within different social and media contexts (Fairclough, 1995; UNESCO, 2011).

Accordingly, awareness of hate speech can be conceived as a multidimensional capacity combining perception and critical evaluation. At a basic level, individuals must be able to identify speech that includes insults, discrimination, or incitement against individuals or groups. At a more advanced level, this awareness takes the form of evaluating such speech in light of contextual factors, such as the social and cultural context, and its potential impact on individuals and society (Potter, 2004; UNESCO, 2015).

More specifically, this awareness comprises three dimensions: cognitive, normative, and critical. The cognitive dimension refers to the ability to recognize and identify hate speech in media messages, which reflects the development of cognitive structures that allow individuals to process and interpret media content critically (Potter, 2004).

The normative dimension involves evaluating media discourse in relation to social and ethical norms such as respect, tolerance, and human dignity, which are central objectives of MIL (Bulger & Davison, 2018; UNESCO, 2015).

Finally, the critical dimension refers to the ability to analyze the communicative context of the message and the implications of the discourse, and the broader media environment in which it circulates, a process emphasized in critical media literacy scholarship (Buckingham, 2003; Kellner & Share, 2007; Obermaier et al., 2025).

Given the scope of this study, these dimensions are used as a conceptual framework to interpret the findings, which exclusively reflect the cognitive dimension, rather than as separate measured subscales.

1. Hate Speech: Forms, Consequences, and the Role of MIL

In this section, the forms, the impact of hate speech, and the educational role of MIL in addressing it will be reviewed.

1.1. *Forms of Hate Speech*

Research suggests hate speech is a complex way of communication that varies with context. Abu Sekin (2021) identified three major types of hate speech:

− The first and most dangerous type includes words that ought to be prohibited (e.g., direct calls for violence, genocide, harm to a group or society, destruction of property, or rejecting one's right to exist).

− The second type consists of words that can be restricted by global law if they threaten public order, such as intolerant comments or mocking that could lead to harm or conflict.

− The third type is not condemned by the law, but it is harmful to individuals; therefore, the best way to deal with this type is to encourage conversation and understanding.

While Abu Sekin (2021) categorizes hate speech according to its legal severity, other scholars focus on its thematic manifestations in digital environments. Moreover, Gracia-Calandín and Suárez-Montoya (2023) identified five key manifestations of hate speech in digital space: religious intolerance, cyber racism, political defamation, misogyny, and attacks on LGBTQ+ communities.

According to KAICIID (2019) and the United Nations (2019), hate speech can take several forms in the media. It often includes using language that insults, shames, or encourages violence against certain groups based on their race, religion, gender, or nationality. It may be expressed through writing and sharing of articles or posts, whether online or in print, that promote exclusion or disrespect.

As well as verbal expressions, hate speech can be expressed in visual and digital forms. The use of pictures, such as cartoons, gestures, symbols, and memes, can support stereotyping, which occasionally normalizes hate speech (He et al., 2025; Müller & Schwarz, 2021). In digital environments, more technologically mediated forms may also emerge, including AI-generated deepfakes or manipulated audiovisual content that can distort information and intensify hostility toward specific groups (Albladi et al., 2025). In addition, some forms of media violence, such as video games, encourage aggression against certain groups of people (Anderson et al., 2010).

Beyond mediated and symbolic forms of hate speech, the most direct and serious form occurs when individuals call for violence or crimes against others in public (Waldron, 2012). These different forms of hate speech urge us to think of the best strategies to mitigate them.

1.2. *The Harmful Impact of Hate Speech*

Findings of studies from the Council of Europe (2024) as well as by Kareem & Abdulrazzaq (2020) suggest that hate speech could cause prolonged mental and social harm that eventually leads to terror, seclusion, and undermining of community cohesion.

In this regard, Waldron (2012) clarifies that hate speech conveys two main harmful messages:

- The first is aimed at victims to dehumanize and exclude them,
- The second is aimed at the public to normalize unequal treatment.

When such messages persist, they intensify victims' suffering and embolden offenders, deepening social division (Online Hate Prevention Institute, 2014).

This also applies to the harmful effects of hate speech online. Digital platforms make hate speech spread more broadly and faster as they offer inexpensive and extensive communication channels, as noted by (Farrand, 2023; Laub, 2019; Müller & Schwarz, 2021). Some studies indicate that algorithms can promote extremist or divisive content, including hate speech, since such content results in higher levels of user interaction. These processes could enhance echo chambers and the dissemination of hate speech in online environments. Thus, to analyze the proliferation of hate speech in the online environment, it is important to know the function of algorithms in structuring and delivering digital content (Cinelli et al., 2021; Huszár et al., 2022; Sonni, 2025).

Consequently, victims of online hate speech feel more unsafe in their offline lives, unlike victims of other cybercrimes (Dreißigacker et al., 2024; Schäfer et al., 2024). Therefore, letting online hate speech go unchecked has turned it into a constant human rights problem: always circulating, and hard to see or stop in time. In extreme cases, its consequences can extend beyond psychological harm and are connected to serious risks, including suicide (UNESCO, 2015).

1.3. *Hate Speech and the Role of Media and Information Literacy*

It is challenging to combat hate speech on social media platforms; therefore, communities and individuals need to be equipped with the tools needed to manage this phenomenon to promote peaceful coexistence.

As per UNESCO (2023), MIL is a key strategy because it gives individuals the skills they need to identify, understand, and counter the harmful effects of hate speech.

According to Potter's theory (2004), media-literate individuals have a cognitive structure that allows them to process media content effectively and not just consume it. The more critically aware a person is, the more they are aware of the motivations of the media content. He also identifies two main components that shape how we deal with media content:

- Personal perspective (e.g., why we use media).
- Cognitive structures (e.g., what we know about media and reality).

So, when media-literate students deal with hate speech, this means that they move beyond irrational reactions, such as emotional outbursts, altered feelings, and a perceived offense, to a critical commentary on the communicative function and social context.

MIL can contribute to reducing hate speech, as well as promoting tolerance and acceptance (Ansari Jafari et al., 2024; UNESCO, 2011). It may inspire the fruitful partnership of educational institutions, media organizations, and civic actors in their efforts towards hate speech (Buckingham, 2003; Obermaier et al., 2025; Masterman, 2003).

Based on the reviewed literature, media and information literacy is widely considered one of the most important tools to fight hate speech in societies, because it makes individuals more critical, gives them the tools to use media properly, and contributes to more responsible and critical media engagement (Al-Ibrahim & Sabirah, 2022; Al-Zou'bi, 2025; Bulger & Davison, 2018; UNESCO, 2023).

2. Research Purpose and Questions

Fayyad (2023) shows that social platforms like Facebook and X now have a major and often harmful influence on public discourse. They have become channels for spreading hate-filled cultures, sectarianism, and discrimination. As more people use the internet, those who are racist, sexist, xenophobic, or anti-refugee have more space to spread hate. This highlights the importance of having educational strategies and interventions that can enable individuals to identify and respond critically to hate speech in digital environments.

Internationally, MIL was suggested as an educational tool to address harmful digital phenomena such as hate speech (Buckingham, 2003; Masterman, 2003; UNESCO, 2011), but such interventions have been studied mostly within Western education systems, which differ significantly from those in the Arab world, where hate speech is influenced by religious and tribal/sectarian tensions (Ben Jilali, 2022).

In addition, several existing international studies assess the outcomes of MIL by self-evaluation of competences and not through performance assessment, a tendency observed in the research of media literacy (Ansari Jafari et al., 2024; Obermaier et al., 2025). This methodological limitation is also evident in some Arab studies. For instance, the study by Al Serhan & Elareshi (2020) reported that 77% of Zarqa University students in Jordan claimed that they could identify hate speech, but this was based only on their responses to the questionnaires. This limits the extent to which their conclusions can be generalized.

While some Arab countries, such as Lebanon, have implemented programs like MDLAB to disseminate critical media literacy (MDLAB, 2022), the current study focuses specifically on examining the impact of MIL on the hate speech awareness. However, the political, cultural, and educational contexts vary across the Arab region, which may shape how MIL is received and how effectively it builds hate speech awareness across various settings.

Therefore, this study aims to provide empirical evidence on whether MIL can improve university students' awareness of hate speech through the following research questions:

1) To what extent does media and information literacy influence students' awareness of hate speech?

2) Are there statistically significant differences in the effect of media and information literacy on students' awareness of hate speech attributed to the variables of gender and academic specialization?

By addressing these questions, the study helps fill a gap in the literature, as empirical evidence on the MIL impact on hate speech awareness in educational contexts remains limited. It also adds empirical evidence from a non-Western context, extending global MIL debates on how hate speech is recognized and critically evaluated in diverse socio-cultural settings.

3. Methodology

This research used a quasi-experimental time-series design including one pre-test followed by two post-tests; as a result, causal inference is limited and should be drawn with caution. This design allows for measuring changes in the same individuals before and after an intervention, as well as assessing the continuity of this effect through multiple measurements. By doing so, it contributes to reducing the impact of individual differences and controlling external variables. It is also practical in educational settings where it is difficult to isolate control groups and is consistent with the nature of MIL programs, which are expected to produce short- to medium-term impacts (Shadish et al., 2002).

3.1. *Population and Sample*

The study population consisted of all 466 students registered in the MIL course at Al al-Bayt University for the 2023/2024 academic year. A sample of 156 students was selected from both the scientific and humanities departments. Out of these, 119 students (76.28%) were female, 37 students (23.72%) were male, 118 students (75.64%) were majoring in humanities, and 38 students (24.36%) were majoring in science.

The researcher taught all students in the course, so the convenience sampling method was used to choose participants. The researcher chose this method as it was easy to reach the participants, saved both time and expenses, and ensured that everyone in the sample got the educational intervention, which is essential for getting an accurate measure of the impact (Etikan et al., 2015; Creswell & Creswell, 2017). The final sample was composed of students who were present in all sessions and at each point in time completed all measures. Three tests were conducted to assess how MIL affected the students' perception of hate speech: one pre-test and two post-tests. The final number of participants who completed all three tests was 156.

3.2. *Study Tool*

While the theoretical framework conceptualizes hate speech awareness as multidimensional, the present tool exclusively captures the cognitive dimension through performance-based identification tasks, measuring students' identification of hate speech in media, rather than their actual behavioural responses in real-world contexts. Therefore, a three-section test was created: the first part asked the participants for personal information (e.g., gender and academic specialization) as prior research asserted that gender and academic specializations may influence how individuals critically engage with media content (Hayat et al., 2017; Kellner & Share, 2007), making these variables relevant for analysis.

The second part included a multiple-choice question with 10 items to differentiate media texts that contain hate speech. The third part had two questions (open-ended and closed-ended). The first of these is related to students' definition of hate speech. The second is related to the Rabat Plan of Action (OHCHR, 2012), which presents the criteria for defining hate speech; namely, the social and political context, the speaker's status and influence, the speaker's intention to incite, the content and style of the speech, its prevalence, and the potential for harm resulting from it.

Regarding the tool's language, the items relied on texts and linguistic expressions that constituted an essential part of the measurement process and respondents' understanding; they have been retained in their original language to maintain the integrity of the meaning and strengthen validity and reliability. For explanatory and informational purposes only, an unofficial English translation of the tool items is added after each item; however, the original version was the only one used in data collection. See Appendices (A, B, C).

3.2.1. Test Validity

To confirm the content validity, the test was reviewed by five experts (two in educational and three in MIL). The majority of their comments were related to the clarity, relevance, and appropriateness of items for the sample. Based on their recommendations, the researcher revised the test and omitted three items that were redundant or not in line with the study goals. These changes helped the final version of the test to measure what it was supposed to.

3.2.2. Internal Consistency and Reliability

Thirty students, enrolled in the media and information literacy course, were selected as an explanatory sample to examine the tool's internal consistency and reliability. First, the item-total correlation coefficients were calculated to examine the internal consistency. The values ranged from 0.63 to 0.93, all statistically significant ($p < 0.05$). This means that the test has adequate item-discrimination and supports internal consistency. See Table 1.

Table 1. Internal consistency of the test (N = 30)

STATEMENT	1	2	3	4	5	6	7	8	9	10
Correlation coefficient	0.72*	0.79*	0.93*	0.82*	0.701*	0.62*	0.69*	0.81*	0.63*	0.71*
Statistical significance	0.023	0.043	0.037	0.020	0.023	0.011	0.0208	0.046	0.015	0.0388

* $p < 0.05$

Second, the tool's reliability was calculated using Cronbach's alpha coefficient, which was 0.863. This value exceeds 0.70, which is acceptable and affirms the reliability of the test (Nunnally & Bernstein, 1994).

3.2.3. Calculation of Item Difficulty and Ease:

The ease of any item in the test was measured by calculating the mean of the correct or incorrect answers according to the following equation:

Ease coefficient = Number of correct answers ÷ (Number of correct + Number of incorrect) (Al-Sarraf, 2002). After calculating the ease coefficients for the test items, it was found that they ranged from 0.35 to 0.90, and hence fell within the acceptable range. The difficulty coefficients for the test items were also calculated, and these ranged from 0.10 to 0.65, which also fell within the acceptable range. See Table 2.

Table 2. Item difficulty and ease coefficients (N = 30)

STATEMENT	1	2	3	4	5	6	7	8	9	10
Ease coefficient	0.61	0.65	0.55	0.49	0.60	0.58	0.61	0.40	0.35	0.90
Difficulty coefficient	0.39	0.35	0.45	0.51	0.40	0.42	0.39	0.60	0.65	0.10

3.2.4. Quantitative Estimation of Student Performance on the Test

Student performance on the test was estimated based on one point for a correct answer and zero for an incorrect answer. The total final score for the test was 10 points.

3.2.5. Determining the Time Required for the Test:

The appropriate time for the test was calculated by applying it to the explanatory sample of 30 students studying MIL. The time taken by each student to answer the test questions was recorded. The time required to complete the test was calculated using the following equation by Al-Sayed (2017):

Z2=Z1×M2/M1

Z2=15.2×5/7.6

Z2=10 minutes

Where:

M2 = Expected average score= Number of questions/2.

M1 = Pilot average score (N = 30).

Z1 = Pilot completion time = (Time of first student + Time of last student)/2.

Z2 = Estimated time required to complete the test.

According to the equation's result, the test should take 10 minutes to be completed, not counting the time needed to give instructions.

3.2.6. The Final Form of the Test

There were ten items on the final test, each with two choices, and two questions: open-ended and closed-ended questions. The test was then prepared for use in the primary research experiment. All three tests followed the same statistical procedures as they were all structured the same way, had the same number of questions, required the same way of answering, and had the same format of questions.

3.3. *Participation*

This study conformed to the ethical guidelines for scientific research. Every student who took part in the research was informed of the nature of the study, its objectives, and implementation steps. They were free to participate. They were motivated to do so because the tests were part of their training and helped them improve their critical skills and the ability to distinguish hate speech. Students were told that they could leave at any time without consequences. They also wrote their names on the three tests to compare their performance and illustrate the progress they made in understanding hate speech in media content. Only the researcher could see all of the files and data, which were kept private.

3.4. Procedure

The design of the MIL intervention was theoretically based on media literacy theory and critical approaches (Buckingham, 2003; Kellner & Share, 2007; Potter, 2004).

The researcher designed the intervention's structure to pedagogically engage the three dimensions of hate speech awareness identified in the theoretical framework. First, the cognitive level was targeted with exercises in which learners had to recognize hate speech and non-hate speech in real-life samples. Second, students practiced the normative dimension by considering media messages from the

perspective of ethical guidelines, including respect, tolerance, and human dignity. Third, the critical level was addressed through a series of structured inquiries into the communicative situation of media messages, including the speaker's intent, influence, and the broader media environment in which the message circulates. These dimensions were not measured as separate constructs, and the study tool captures cognitive performance only.

To implement this framework, before introducing students to the hate speech unit in the MIL course, 156 of them took the pre-test, which was held on April 4, 2024, using Google Forms. See Appendix (A). The MIL intervention, which has 3 (90-minute) sessions on April 7, 9, and 11, 2024, introduced students to the concept of hate speech and its forms, using real-life examples from different social media platforms. Each session included extensive training on how to identify and critically analyze hate speech in media content. Two post-tests were given to assess the immediate and short-term effects of the intervention on their awareness of hate speech. The post-tests were distributed electronically to the same sample using Google Forms on April 14 and 16. See Appendices (B & C). The responses to the three tests were exported from Google Forms into Excel spreadsheets.

3.5. *Data Analysis*

In this section, the analysis begins with testing the normality of the data distribution, followed by the quantitative and qualitative analyses:

3.5.1. Data Distribution

The researcher used the Kolmogorov-Smirnov test to assess the normality of the data distribution. The results were as follows:

Table 3. Normality of data distribution for the sample

KOLMOGROV-SMIRNOV	PRE-TEST	FIRST POST-TEST	SECOND POST-TEST
N	156	156	156
Significance level	$p < .001$	$p < .001$	$p < .001$

As indicated in Table 3, the data distribution in all study tests was non-normal ($p < .001$) (Ghasemi & Zehediasl, 2012). As a result, non-parametric tests were used to answer the research questions. Specifically, the Friedman's test was applied to ascertain the difference between the mean ranks in all three tests regarding students' level of awareness of hate speech. Following this, the Mann-Whitney U test was conducted to ascertain the difference between the groups of males and females in the effect of MIL on students' awareness level of hate speech, and the difference between the groups of the scientific and humanities specializations in the impact of MIL on their awareness of hate speech.

3.5.2. Quantitative Analysis

Since the data did not follow a normal distribution (as confirmed by the Kolmogorov-Smirnov test), non-parametric tests were used for the main analysis. The Friedman test was used to assess differences across the three measurements (pre-test, first post-test, and second post-test), while the Mann-Whitney U test was used to examine differences based on gender and academic specialization. In addition, paired-samples t-tests and Cohen's d effect size were conducted as supplementary measures, with 95% CIs, to quantify the magnitude of pre–post change (Cohen, 2013).

3.5.3. Qualitative Analysis

In addition to the quantitative analysis, the responses to the closed-ended question were coded and analyzed using frequencies and percentages. In contrast, the responses to the open-ended question were analyzed qualitatively using qualitative content analysis. The researcher carefully read the responses multiple times and then grouped them into "conceptual clusters" based on semantic similarity. These clusters were considered initial codes and were later grouped into broader thematic categories. The frequency of each code and category was then calculated to provide a quantitative description of qualitative patterns in the responses. The categories were considered exhaustive, and each response was coded to the one category that most closely represented its essence. The coding proceeded until no new conceptual clusters could be formed from further responses. Trustworthiness of the analysis was enhanced by applying a stable coding scheme and recoding a random subset of responses (approximately 20%) after a two-week interval; discrepancies were resolved by refining category definitions.

To illustrate the coding process, the following anonymized responses from the pre-test to the open-ended question "*What is your definition of hate speech?"* are provided:

1- {*Student*}: *"It is a speech that harms others."*

2- {*Student*}*: "Hate speech is verbally attacking others".*

These two answers demonstrate students' initial understanding of the hate speech definition. They were grouped under one main theme: "perceived offensiveness".

4. Results

This section presents the results of the research, which aimed to identify the effect of MIL on students' awareness of hate speech:

4.1. *The Research Question Results*

To answer the first research question: "*To what extent does media and information literacy influence students' awareness of hate speech?"* The Friedman test was conducted to assess the difference in mean ranks of the research sample across the

pre-test, first post-test, and second post-test regarding their level of awareness of hate speech. To estimate the magnitude of the observed differences, the effect size, Kendall's W, was calculated. See Table 4.

Table 4. Friedman's test results for the difference in hate speech awareness across the sample's tests

MEASURE	MEAN RANKS	N	CHI-SQUARE	DF.	P	KENDALL's W
Pre-test	1.19					
First post-test	2.32					
Second post-test	2.48					
Friedman test (Effect Size)		156	192.627	2	< .001	0.62

Table 4 shows a statistically significant difference between the mean ranks of the students among the three tests. These differences go in favor of the first post-test and the second post-test, which show a positive role of MIL in sensitizing against hate speech. The mean ranks of the first post-test and second post-test (2.32, 2.48) were also higher than the mean rank of the pre-test (1.19). The effect size calculated using Kendall's W indicated a large effect (W = 0.62), suggesting a substantial change in awareness throughout the three stages of testing. Overall, the findings show that students' awareness was steadily improved in all three measurements, and the differences between the pre-test and the two post-tests were significant.

To further quantify the magnitude of pre–post change, paired-samples t-tests and Cohen's d effect size were conducted as supplementary measures. Although the data did not follow a normal distribution, with N = 156, the Central Limit Theorem supports the robustness of the t-test against violations of normality (Cohen, 2013). 95% CIs were also calculated to reflect the accuracy of the effect size estimates. See Table 5.

Table 5. Effect Size of the MIL intervention on hate speech awareness (Cohen's d, 95% CI)

COMPARISON	T	N	COHEN's D	95% CI
Pre-test vs first post-test	12.98	156	0.52	0.29 – 0.75
Pre-test vs second post-test	14.89	156	0.60	0.37 – 0.82

Table 5 indicates that the paired-samples t-tests revealed significant improvements from pre-test to first and second post-tests (t = 12.98 and t = 14.89, respectively). The values of Cohen's d for the first and second post-test were 0.52 and 0.60 respectively; indicating a medium-sized effect of the MIL intervention on hate speech awareness. The 95% CIs ranged from 0.29 to 0.75 and from 0.37 to 0.82, respectively, indicating that the effect size of the intervention was stable and could be considered meaningful. These findings reflect a moderate but stable level of increased awareness post-intervention.

To answer the second research question: "*Are there statistically significant differences in the effect of media and information literacy on students' awareness of hate speech attributed to the variables of gender and academic specialization?",* the non-parametric Mann-Whitney U test was conducted to test the difference between the mean ranks of males and females in the effect of MIL on their awareness of hate speech. The effect size (r) was calculated to estimate the magnitude of the difference between the two groups. See Table 6.

Table 6. Difference between the mean ranks of males and females regarding the MIL effect on students' awareness of hate speech

DEPENDENT VARIABLE	M.	F.	U	W	Z	SIG.	EFFECT SIZE (r)
Sum of Ranks	2921	9325	2060	2921	-1.351	0.177	0.11
Mean Ranks	71.24	81.09					

Table 6 indicates no statistically significant difference in the mean ranks between males and females in their level of awareness of hate speech (p = 0.177). In addition, the calculated effect size was small (r = 0.11), indicating no meaningful difference in the effect of MIL on the gender variable.

The researcher also conducted a Mann-Whitney U test to test the difference in mean ranks between the scientific and humanities departments in the effect of MIL on students' awareness of hate speech. In addition, the effect size (r) was calculated to estimate the magnitude of the difference between the two groups. See Table 7.

Table 7. Difference between the mean ranks of the scientific and humanities students regarding the MIL effect on students' awareness of hate speech

DEPENDENT VARIABLE	SCI.	HUM.	U	W	Z	SIG.	EFFECT SIZE (r)
Sum of Ranks	2838.5	9407	2018.5	2838.5	-1.38	0.168	0.11
Mean Ranks	70.96	81.7					

Table 7 reveals no statistically significant difference in the mean ranks between science students and humanities students in their level of awareness of hate speech (p = 0.168). The calculated effect size was small (r = 0.11), indicating no meaningful difference in the effect of MIL on the academic specialization variable.

4.2. *Results of the Open-Ended and Closed- Ended Questions*

Although the following results do not directly answer the research questions, they lay the groundwork for comparing levels of awareness and perceptions before and after the MIL intervention. Due to the qualitative nature of the coding, the percentages presented in the following section are approximate.

4.2.1. Open-Ended Question: Students' Definitions of Hate Speech

When students were asked to define hate speech, about 65-70% of the answers to the pre-test defined it as "offensive speech directed at an individual or group based on religion, race, or gender", emphasizing how it can break down social peace. About 15% of the students referred to hate speech as a direct incitement to harm or violence. The remaining (10–15%) were about general aspects, such as "discrimination between people" without going through additional details.

There was a clear change in the responses to the first post-test. The responses added new elements to the concept of hate speech, such as the speaker's intention and purpose (representing more than 30% of responses), the speaker's authority and influence over the audience (approximately 20%), and the repetition of the speech as a multiplier of the effect (approximately 10%). 40% of them kept perceiving hate speech as offensive language.

The second post-test showed that the sample enhanced their understanding of hate speech. More than 45% of students reported that hate speech was more than just insults but also featured the dissemination of distorted information, false facts, or lies that provoke hostile and negative attitudes towards individuals or groups. Approximately a quarter of students associated hate speech with the encouragement of violence, discrimination, and exclusion, to explain how the phenomenon manifests itself in offline contexts.

A small percentage, less than 20%, considered that the way things are said is a key factor in defining hate speech. A percentage of approximately (10%) kept their definitions simple and did not reveal any variation when stating what hate speech is.

Overall, the results indicate a progressive development in students' responses through the tests of the study, from a simplistic recognition of hate speech to the adoption of more comprehensive evaluation forms. Instead of defining hate speech as merely 'offensive language', the responses demonstrate the ability to associate it with discrimination, exclusion, and the propagation of harmful misinformation.

4.2.2. Closed-Ended Question: Students' Perspectives on Hate Speech Criteria

The next section presents students' perspectives on the criteria for detecting hate speech through the pre- and post-tests. See Tables (8, 9, and 10).

Table 8. Students' pre-test perspectives on criteria for identifying hate speech

CRITERION	APPROX. %	COUNT
Speaker's Intent: The speaker must have a clear intent or premeditated intention to spread hate speech.	19.9%	31
Speech Content: The extent to which the speech directly or indirectly advocates hatred.	28.8%	45
Speaker's Influence: The speaker's position and the power or influence they have over the audience.	23.1%	36
Persistence in repeating hate speech.	11.5%	18
Potential for harm.	12.8%	20
I do not know how to define hate speech.	3.8%	6
Total	100%	156

Table 9. Students' first post-test perspectives on criteria for identifying hate speech

CRITERION	APPROX. %	COUNT
Speaker's Intent: The speaker must have a clear intent or premeditated intention to spread hate speech.	62.2%	97
Speech Content: The extent to which the speech directly or indirectly advocates hatred.	21.2%	33
Speaker's Influence: The speaker's position and the power or influence they have over the audience.	8.3%	13
Persistence in repeating hate speech.	7.1%	11
Potential for harm.	0.6%	1
I do not know how to define hate speech.	0.6%	1
Total	100%	156

Table 10. Students' second post-test perspectives on criteria for identifying hate speech

CRITERION	APPROX. %	COUNT
Speaker's Intent: The speaker must have a clear intent or premeditated intention to spread hate speech.	73.7%	115
Speech Content: The extent to which the speech directly or indirectly advocates hatred.	15.4%	24
Speaker's Influence: The speaker's position and the power or influence they have over the audience.	7.1%	11
Persistence in repeating hate speech.	1.9%	3
Potential for harm.	1.3%	2
I still do not know how to define hate speech.	0.6%	1
Total	100%	156

Tables 8, 9, and 10 show the progress students made in identifying hate speech criteria across three phases: before training (pre-test), immediately after training (first post-test), and shortly after training (second post-test). The results can be interpreted as follows:

Pre-test (Table 8): Table 8 indicates that the criteria by which students identified hate speech were distributed as follows: speech content (28.8%), speaker's influence (23.1%), and speaker's intent (19.9%) were the most prominent elements. A percentage of (12.8%) focused on the potential for harm, and a percentage of 11.5% focused on the persistence in repeating hate speech. However, 3.8% stated that they did not know how to define hate speech.

First post-test (Table 9): Table 9 shows a change in students' awareness of hate speech indicators. In particular, the proportion of students relying on the speaker's intent criterion increased substantially from 19.9% in the pre-test to 62.2% in the first post-test, while the criteria of speaker's influence, persistence to repeat hate speech, and potential for harm accounted for 8.3%, 7.1%, and 0.6%, respectively. The proportion of students reporting uncertainty declined from 3.8% in the pre-test to 0.6% in the first post-test. This pattern may reflect the influence of the training sessions on students' understanding of hate speech identification.

Second post-test (Table 10): Table 10 shows that 73.7% of students selected the speaker's intent criterion. Other criteria, such as speech content, declined from 21.2% in the first post-test to 15.4% in the second post-test. References to other criteria, such as speaker's influence, persistence in repeating hate speech, and potential for harm, also declined to 7.1%, 1.9%, and 1.3%, respectively. Just 0.6% of students were still unable to detect hate speech in media content.

These results suggest that the effect of the MIL intervention was sustained, as the responses continued to focus primarily on the speaker's intent as the main criterion for identifying hate speech. The growing salience of the speaker's intent already indicated that the MIL intervention helped students distinguish offensive speeches from hate speech, suggesting a more critical evaluation of media messages.

5. Discussion

The results show that media and information literacy significantly improved students' awareness of the definitions and criteria of hate speech. The findings reflect cognitive awareness rather than actual behavioral responses. While most previous research has been conducted in Western contexts, this study extends international MIL research by demonstrating that the interpretation of hate speech is shaped by specific socio-cultural and media environments. This suggests that MIL operates as a context-dependent framework, in which the development of critical awareness varies according to local communicative conditions. This is consistent with media literacy theory (Potter, 2004) and aligns with findings reported in several international research by Buckingham (2003) and UNESCO (2011).

Gender and academic specialization characteristics did not significantly affect the results. This suggests that both male and female students benefited equally from MIL, and when used skillfully, these abilities may be cultivated in a variety of disciplines. This finding contrasts with that reported by Hayat et al. (2017) and Kellner & Share (2007), who found statistically significant differences in media consumption, attributed to gender and academic specializations.

The qualitative findings revealed that before the MIL's intervention, the responses reflected a simplified understanding of hate speech, and a few students were unable to do that at all. After the MIL intervention, the results of the post-tests revealed that the majority of the responses defined hate speech as the dissemination of misleading or hostile messages intended to provoke negative attitudes, discrimination, or violence against individuals or groups. This aligns with the findings of Ben Jilali (2022), Chetty and Alathur (2018), and Elmoubarak (2021).

After the MIL intervention, the results indicate a tendency to consider the speaker's intention as the main criterion for identifying hate speech, which aligns with the criteria outlined by the Rabat Plan of Action (OHCHR, 2012). However, relying solely on the speaker's intent illustrates the risk of reducing the identification of hate speech to a single criterion.

However, focusing on the speaker's intent alone may not fully capture how hate speech functions in digital environments, where its interpretation and impact are also shaped by algorithmic curation and amplification processes (Cinelli et al., 2021; Huszár et al., 2022; Sonni, 2025).

This is particularly problematic in digital environments where content spreads through algorithmic systems regardless of what the original speaker intended. For example, a hateful post may reach a wide audience not because its author planned for this, but because platform algorithms amplify highly engaging content, which can include polarizing or harmful messages. This suggests that while intent remains a legally recognized criterion of hate speech under the Rabat Plan of Action (OHCHR, 2012), MIL may need to go further by helping students understand how platforms shape what they see and why.

Notably, the stable decline in uncertainty, reflected in fewer ("I don't know") responses regarding the criteria for defining hate speech, indicates that MIL can provide analytical frameworks for understanding the complexity of digital communication.

This growing analytical confidence is further reflected in the pattern of students' responses. The study tool did not assess whether these cognitive changes translated into actual behavioral responses in real-world settings, nor did it operationalize the cognitive, normative, and critical dimensions as separate subscales. The findings indicate a shift from basic recognition of hate speech toward more evaluative and context-sensitive understanding, as responses increasingly incorporate considerations such as intent, context, and potential impact. This interpretation is consistent with the theoretical foundations of media literacy (Buckingham, 2003; Kellner & Share, 2007; Fairclough, 1995; Potter, 2004; UNESCO, 2015).

Although Buckingham (2003), Bulger and Davison (2018), and Gracia-Calandín and Suárez-Montoya (2023) argue that MIL by itself cannot address hate speech on digital platforms, the findings of this study suggest that MIL interventions can at least build the cognitive foundations students need to critically recognize hate speech. The findings demonstrate that a structured educational intervention can influence how students think about and approach the subject.

Future research should investigate if these cognitive changes in processing would also translate into changes in behavioral responses when students are exposed to hate speech in the "real world" of social media. This highlights the relevance of the study beyond its immediate context and supports its contribution to broader discussions in communication and media literacy research.

Conclusion

This study shows that structured MIL interventions can strengthen the cognitive foundations of hate speech awareness. Within the limits of a performance-based tool, the findings indicate that MIL has the potential to bring students from a basic recognition of hate speech toward a more critical contextual interpretation of it. It should be noted that, as Buckingham (2003) and Bulger & Davison (2018) argue, MIL promotes critical awareness rather than providing a comprehensive solution to hate speech.

The findings highlight the importance of combining textual interpretation with contextual awareness when evaluating potentially harmful speech. In this sense, this study contributes to global debates about the role of MIL in strengthening individuals' capacity to critically interpret hate speech in contemporary media environments, while also adding empirical evidence from a non-Western context, which is relatively underrepresented in the media and information literacy (MIL) literature.

Limitations & Recommendations

Despite the value of the results reached by this study, there are limitations:

- The implementation of this study in a Jordanian academic environment limits the generalizability of its results to other socio-cultural settings.

- The impact of the intervention was assessed over a short period, which restricts conclusions about the long-term sustainability of the observed effects.

- Conducting a quasi-experimental (time-series design) without a control group limits the strength of causal inferences, as changes in awareness cannot be attributed exclusively to the intervention.

- The dual role of the researcher in both implementing the intervention and analyzing the data may introduce potential bias.

- Rather than utilizing the tool as a multidimensional psychometric scale, the study used a performance-based test, so the results reflect students' cognitive ability to recognize hate speech.

Therefore, future studies should seek to overcome these limitations by involving more diverse population groups, including control groups where possible, designing multidimensional scales, conducting follow-up measurements over a longer period, and adopting dual-independent coding.

Funding and Acknowledgements

The researcher gratefully acknowledges Al al-Bayt University for its support in granting her a sabbatical leave of absence during the academic year 2025/2026, which made this research possible. The primary data is available from the researcher upon request. This work received no funding and is original, experimental, and well-documented. The researcher used Grammarly, QuillBot, and Wordtune for proofreading and translation.

References

Abu Sekin, H. (2021). Non-traditional security threats: hate speech as a model. *Arab Journal of Political Science,* 18(3), 28-48. Retrieved 7 September 2025, from http://search.mandumah.com/Record/1177059

Albladi, A., Islam, M., Das, A., Bigonah, M., Zhang, Z. & Jamshidi, F. (2025). Hate speech detection using large language models: a comprehensive review. *IEEE Access,* 13, 20871-20892.

Al-Ibrahim, D. & Sabirah, F. (2022). The role of media and digital literacy in confronting hate speech through social media. In *The First International Virtual Scientific Conference: Hate Speech and the Question of Tolerance in a Changing World: Practices, Implications, and Mechanisms of Confrontation*. Kasdi Merbah University, Algeria. Retrieved 15 March 2026, from https://www.researchgate.net/publication/374153029_dwr_altrbyt_alalamyt_walrqmyt_fy_mwajht_khtab_alkrahyt_br_wsayl_altwasl_alajtmay

Al Serhan, F. & Elareshi, M. (2020). New media and hate speech: a study of university students in Jordan*. Opción: Revista de Ciencias Humanas y Sociales*, *36*(26), 166-181. Retrieved 10 September 2025, from https://www.academia.edu/88202289/New_media_and_hate_speech_A_study_of_university_students_in_Jordan

Al-Sarraf, Q. (2002). *Measurement and evaluation in education and teaching.* Cairo: Dar Al-Kitab Al-Hadith.

Al-Sayed, F. (2017). *Statistical psychology and measurement of the human mind* (3rd ed.). Cairo: Dar Al-Fikr Al-Arabi. Retrieved 10 September 2025, from https://archive.org/details/khatabaya2007_yahoo_20170911_1649

Al-Zou'bi, R. (2025). Assessing the impact of media and information literacy on creative video content production. *Journal of Media Literacy Education,* 17(1), 106-120.

Anderson, C. A., Shibuya, A., Ihori, N., Swing, E. L., Bushman, B. J., Sakamoto, A., Rothstein, H. R. & Saleem, M. (2010). Violent video game effects on aggression, empathy, and prosocial behavior in eastern and western countries: a meta-analytic review. *Psychological Bulletin*, 136(2), 151-173.

Ansari Jafari, M. H., Rasouli, M. R. & Khajeh Noori, N. (2024). The role of critical thinking to improve the efficiency of media literacy by emphasizing the audience's cognitive system. *New Media Studies,* 10(37), 119-153.

Ben Jilali, M. A. (2022). Hate speech between concept and reality. *World Politics,* 6(1), 169-181. Retrieved 7 September 2025, from https://asjp.cerist.dz/en/article/191478

Buckingham, D. (2003). *Media education: literacy, learning and contemporary culture*. Cambridge: Polity Press.

Bulger, M. & Davison, P. (2018). The promises, challenges, and futures of media literacy. *Journal of Media Literacy Education*, 10(1), 1-21.

Chetty, N. & Alathur, S. (2018). Hate speech review in the context of online social networks. *Aggression and Violent Behavior*, 40, 108-118.

Cinelli, M., De Francisci Morales, G., Galeazzi, A., Quattrociocchi, W. & Starnini, M. (2021). The echo chamber effect on social media. *In Proceedings of the National Academy of Sciences*, 118(9).

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. 2nd ed. London: Routledge.

Council of Europe. (2024). *Preventing and combating hate speech in times of crisis.* Strasbourg: Council of Europe Publishing. Retrieved 5 September 2025, from https://unipd-centrodirittiumani.it/en/news/preventing-and-combating-hate-speech-in-times-of-crisis-the-council-of-europe-publishes-a-new-study

Creswell, J. W. & Creswell, J. D. (2017). *Research design: qualitative, quantitative, and mixed methods approaches*. 5th ed. Thousand Oaks, CA: SAGE Publications.

Dreißigacker, A., Müller, P. & Isenhardt, A. (2024). Online hate speech victimization: consequences for victims' feelings of insecurity. *Crime Science,* 13(4).

Elmoubarak, A. (2021). *Hate speech between legislation and prevention mechanisms*. Paper presented at Crimes of Discrimination and Hate Speech Conference, El Oued, Algeria. Retrieved 5 September 2025, from https://www.researchgate.net/publication/362253968

Etikan, I., Musa, S. A. & Alkassim, R. S. (2015). Comparison of convenience sampling and purposive sampling. *American Journal of Theoretical and Applied Statistics*, *5*(1), 1-4.

Fairclough, N. (1995). *Critical discourse analysis: The critical study of language*. London: Longman.

Farrand, B. (2023). Is this a hate speech? The difficulty in combating radicalization in coded communications on social media platforms. *European Journal on Criminal Policy and Research*, 29(3), 477–493.

Fayyad, H. (2023). *Sociology of hate speech: factors in the making of hatred and mechanisms of confrontation – an analytical study*. Believers without Borders. Retrieved 5 September 2025, from https://igit.me/2kkOT

Ghasemi, A. & Zehediasl, S. (2012). Normality tests for statistical analysis: a guide for non-statisticians. *International Journal of Endocrinology and Metabolism,* 10(2).

Gracia-Calandín, J. & Suárez-Montoya, L. (2023). The eradication of hate speech on social media: a systematic review. *Journal of Information, Communication and Ethics in Society*, 21(4), 406–421.

Hayat, T., Lesser, O. & Samuel-Azran, T. (2017). Gendered discourse patterns on online social networks: a social network analysis perspective*. Computers in Human Behavior***,** 77, 132–139.

He, J., Yu, C.-Y. & Zhang, W. (2025). *Hateful meme detection using multimodal fusion. In Proceedings of the International Conference on Artificial Intelligence and its Applications*, (pp. 1–9). New York: ACM.

Huszár, F., Ktena, S. I., O'Brien, C., Belli, L., Schlaikjer, A. & Hardt, M. (2022). Algorithmic amplification of politics on Twitter. *In Proceedings of the National Academy of Sciences*, 119(1).

International Dialogue Centre (KAICIID). (2019). *What is hate speech?* Retrieved 10 September 2025, from https://www.kaiciid.org/stories/news/what-hate-speech

Kareem, K. & Abdulrazzaq, M. H. (2020). *Hate speech and its role in undermining the social peace of the audience*. University of Wasit, College of Media. Retrieved 10 September 2025, from https://www.researchgate.net/publication/343281714_khtab_alkrahyt_wdwrh_fy_tqwyd_alslm_almjtmy_ldy_alj

Kellner, D. & Share, J. (2007). Critical media literacy, democracy, and the reconstruction of education. In D. Macedo & S. R. Steinberg (Eds.), *Media literacy: A reader* (pp. 3–23). Peter Lang. Retrieved 5 September 2025, from https://cxarchive.gseis.ucla.edu/xchange/critical-uses-of-media-and-technology/xpress/critical-media-literacy-democracy-and-the-reconstruction-of-education/

Laub, Z. (2019). *Hate speech on social media: global comparisons*. Retrieved 7 September 2025, from https://www.scribd.com/document/779139909/Laub-Z-2019-June-7-Hate-Speech-on-Social-Media-Global-Comparisons

Masterman, L. (2003). *Teaching the media*. London: Routledge.

Media and Digital Literacy Academy of Beirut (MDLAB). (2022). *About MDLAB. Lebanese American University*. Retrieved 10 September 2025, from https://mdlab.lau.edu.lb/about/

Müller, K. & Schwarz, C. (2021). Fanning the flames of hate: social media and hate crime. *Journal of the European Economic Association,* 19(4), 2131–2167.

Nunnally, J. C. & Bernstein, I. H. (1994). *Psychometric theory*. 3rd ed. New York: McGraw-Hill.

Obermaier, M., Schmid, U. K. & Rieger, D. (2025). Empowerment is key? How perceived political and critical digital media literacy explain direct and indirect bystander intervention in online hate speech. *Social Media + Society,* 11(1).

Office of the High Commissioner for Human Rights (OHCHR). (2012). *Rabat plan of action on the prohibition of advocacy of national, racial, or religious hatred that constitutes incitement to discrimination, hostility, or violence.* Retrieved 7 September 2025, from https://www.ohchr.org/en/documents/outcome-documents/rabat-plan-action

Online Hate Prevention Institute. (2014). *Press release: launch of online tool to combat hate*. Retrieved 5 September 2025, from https://ohpi.org.au/press-release-launch-of-online-tool-to-combat-hate/

Potter, W. J. (2004). *Theory of media literacy: A cognitive approach*. Thousand Oaks, CA: SAGE Publications.

Schäfer, S., Rebasso, I., Boyer, M. M. & Planitzer, A. M. (2024). Can we counteract hate? Effects of online hate speech and counter speech on the perception of social groups. *Communication Research,* 51(5), 553-579.

Shadish, W. R., Cook, T. D. & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.

Sonni, A. (2025). AI-based disinformation and hate speech amplification: analysis of Indonesia's digital media ecosystem. *Frontiers in Communication*, 10*,* 1-6.

United Nations Educational, Scientific and Cultural Organization (UNESCO). (2023). *Media and information literacy: an avenue for youth to combat hate speech, misinformation & disinformation.* Retrieved 5 September 2025, from https://www.unesco.org/en/articles/media-information-literacy-avenue-youth-combat-hate-speech-misinformation-disinformation

United Nations Educational, Scientific and Cultural Organization (UNESCO). (2011). *Media and information literacy curriculum for teachers.* Retrieved 7 September 2025, from https://www.unesco.org/en/articles/media-and-information-literacy-curriculum-teachers

United Nations Educational, Scientific and Cultural Organization (UNESCO). (2015). *Countering online hate speech*. Retrieved 5 September 2025, from https://unesdoc.unesco.org/ark:/48223/pf0000233231

United Nations. (2019). *Strategy and plan of action on hate speech*. Retrieved 5 September 2025, from https://www.un.org/en/genocideprevention/documents/UN%20Strategy%20and%20Plan%20of%20Action%20on%20Hate%20Speech%2018%20June%20SYNOPSIS.pdf

Waldron, J. (2012). *The harm in hate speech*. Cambridge, MA: Harvard University Press.