applied sciences

MDPI

Article

Thematic Analysis: A Corpus-Based Method for Understanding Themes/Topics of a Corpus through a Classification Process Using Long Short-Term Memory (LSTM)

Yaser Altameemi [1,*] and Mohammed Altamimi [2]

[1] Department of English Language, College of Arts and Literature, University of Ha'il, Ha'il 81481, Saudi Arabia
[2] Department of Information and Computer Science, College of Computer Science and Engineering, University of Ha'il, Ha'il 81481, Saudi Arabia
* Correspondence: y.albakry@uoh.edu.sa

Abstract: Using advanced algorithms to conduct a thematic analysis reduces the time taken and increases the efficiency of the analysis. Long short-term memory (LSTM) is effective in the field of text classification and natural language processing (NLP). In this study, we adopt LSTM for text classification in order to perform a thematic analysis using concordance lines that are taken from a corpora of news articles. However, the statistical and quantitative analyses of corpus linguistics are not enough to fully identify the semantic shift of terms and concepts. Therefore, we suggest that a corpus should be classified from a linguistic theoretical perspective, as this would help to determine the level of the linguistic patterns that should be applied in the experiment of the classification process. We suggest investigating the concordance lines of the articles rather than only the relationship between collocates, as this has been a limitation for many studies. The findings of this research work highlight the effectiveness of the proposed methodology for the thematic analysis of media coverage, reaching 84% accuracy. This method provides a deeper thematic analysis than only applying the classification process through the collocational analysis.

Keywords: thematic analysis; Long Short-Term Memory (LSTM); corpus linguistics; concordance lines; collocations

check for updates

Citation: Altameemi, Y.; Altamimi, M. Thematic Analysis: A Corpus-Based Method for Understanding Themes/Topics of a Corpus through a Classification Process Using Long Short-Term Memory (LSTM). *Appl. Sci.* **2023**, *13*, 3308. https://doi.org/10.3390/app13053308

Academic Editor: Christos Bouras

Received: 25 January 2023
Revised: 27 February 2023
Accepted: 27 February 2023
Published: 5 March 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

There are diverse areas in which corpus linguistics (CL) has contributed specifically to the notion of language use, which is the norm of critical discourse analysis (CDA). This embedded relationship between CL and CDA has contributed to significant findings, specifically with the focus on building representative data and producing a representative result in order to generalize a social phenomenon. Recent studies have discussed the thematic analysis and what can be called the aboutness of a corpus, concentrating on collocational analysis. According to Baker ([1], p. 96), collocation is a "way of understanding meanings and associations between words which are otherwise difficult to ascertain from a small-scale analysis of a single text".

Previous studies highlight the importance of analyzing keywords and the collocations in a corpus to provide a full understanding of the themes and topics of a corpus. These methodological perspectives are important not only for providing themes and topics, but also for providing the semantic structure of a text. These studies examine deep analyses by investigating the changes in meaning through the collocational network, as argued by Brezina et al. [2].

However, statistical and quantitative analyses are not enough to fully identify the semantic shift of terms and concepts in a corpus, specifically for specialized corpora. Taking this into account, Altameemi [3] manually analyzed a sample of the concordance lines

for strong collocates, as the number of lines is considerable. However, he suggests the importance of applying an automatic thematic analysis to all of the concordance lines. Due to this limitation, the authors suggest the importance of developing a methodology that is based on the proposed algorithm to help linguists in analyzing all of the concordance lines of the strong collocates, in order to increase the representativeness when analyzing a corpus.

In this research, we test a proposed methodology that has not been applied in any previous studies, according to the knowledge of the researchers. This article aims to develop a methodology of applying automatic thematic analysis using deep learning models. In other words, we suggest that experts in computer science should consider the level of linguistic elements in discourse/text, as well as linguistic theories, to improve the proficiency of the algorithm's classification and categorization. Corpus linguists use tools such as LancsBox to apply a thematic analysis to collocations (relations between words) rather than a text's concordance lines. Therefore, this article will contribute to the field of corpus linguistics by annotating/classifying the extracted lines from keywords in context (KWIC). Previous studies have failed to apply thematic analysis for the concordance lines of the collocates. They only focus on thematic analysis based on the relationship between collocates. In order to provide a robust, thematic analysis of a corpus, we argue that corpus linguists need to analyze the lines of the strong collocates by taking into consideration their collocational relationships. The proposed methodology suggests applying the proposed algorithms to perform an automatic thematic analysis of the concordance lines in addition to the collocational analysis.

In the current paper, we try to provide a comprehensive methodology for thematic analysis by combining theoretical aspects in the linguistics field with the practical steps for conducting thematic analysis by applying deep learning models using LSTM. We did not simply classify the whole corpus, but, generally, we performed the following steps:

1. Pre-process and clean the dataset.
2. Identify strong collocates using the DeltaP measurement.
3. Extract lines from the concordance lines of the strong collocates.
4. Apply the algorithm for classifying the extracted KWIC.
5. Compare the automatic findings with the manual analysis of Altameemi's findings [3].

These steps are discussed in detail in Section four.

This paper starts with a discussion of the previous research that mainly focuses on the thematic analysis and the aboutness of a corpus. As this study is a combination of both linguistic and computer science perspectives, the literature review is presented from this perspective. Then we discuss the methodological aspects, including long short-term memory (LSTM) and the steps for automatic thematic analysis. In the analysis section, we discuss the findings based on the results of the experiment and the accuracy of the automatic thematic analysis. The results support the effectiveness and the high accuracy of the proposed methodology and the automatic analysis of the concordance lines.

2. Literature Review

2.1. Corpus Linguistics and Thematic Analysis

Corpus linguistics (CL) has contributed to various areas. In particular, it is typical of critical discourse analysis (CDA) as an ideal of language. This interrelation between CL and CDA has contributed to significant findings, especially with a focus on building representative data and making a prototypical result in order to allow for generalizations to social phenomena. For example, several studies have dealt with political discourse, language learning research, or self-representation discourse. In addition, the synchronic development of combining CL and CDA has become crucial from the perspective of understanding the context of the discourse regarding a specific issue. This point has been raised by Altameemi, who suggests the importance of following Wodak's [4] approach regarding the need for several levels of analysis that may help an analyst to understand the connection between discourse and context.

Recent studies have discussed the thematic analysis, and what can be called the aboutness of a corpus, by concentrating on collocational analysis. According to Baker ([1], p. 96), collocation is a "way of understanding meanings and associations between words which are otherwise difficult to ascertain from a small-scale analysis of a single text." Taking this into account, the collocation can be used for two central purposes: the lexical-syntactic analysis and the expanded consideration of the semantic field. In this section, we will discuss these two points.

Many studies have considered collocational analysis in the language production of first-language and second-language speakers [5], e.g., [6,7]. Other studies, e.g., [7], focus on analyzing the semantic relationship between words through their relationship and co-occurrence in a corpus. According to Evert [8], the frequency of occurrences, the exclusivity of collocates, and their directionality are important dimensions for the semantic relationship between words' co-occurrence (see Evert [8] for more details on the differences between the dimensions). Brezina et al. [2] assert the importance of analyzing collocation; they suggest that collocational analysis reveals the aboutness of a corpus to determine the topics and themes. Gablasova et al. [9] go further to measure and interpret the use of collocation in L1 and L2 production. They follow this approach as they argue that three dimensions have an impact on learners' knowledge, so these dimensions should be considered before the selection of the association measures for the collocation.

The analysis of collocation is also applied by various researchers to analyze how a corpus represents a topic with various ideas and themes. Neg and Tan [10] investigated the diverse representations of COVID-19 and link them to cultural values. They applied the corpus approach by using the mutual information score to measure the word associations in the corpora. Then they look at the media coverage, as well as the prevalence rate of COVID-19. Neg and Tan (2021) not only consider the collocation relationship, but they also connect their findings with the cultural values that were stated in advance before the collocation analysis. Another study conducted by Huang et al. [11] highlights the importance of analyzing online reviews using the polymerization topic sentiment model (PTSM) for sentiment thematic analysis. They argue that PTSM can reveal the impact of reviews on shaping products, as well as on the sales promotion. However, this study did not fully consider the analysis of keywords and the shifts in their meanings within the discourse. Therefore, analysts of discourse may need to go deeper into the analysis and consider how the meanings of concepts relate to various representations.

Analyzing the collocation is also used as a key tool for analyzing lexical–thematic analyses. Bondarchuk et al. [12] and Biber and Reppen [13] argue that analyzing the keywords quantitatively is effective for identifying the lexical structure of a corpus. Huang et al. [14] used the dependency SCOR-topic sentiment (DSTS) to predict online sales performance. Bondarchuk et al. [12] investigated the dominant themes in British weather news. By taking advantage of the capacity of WordSmith Tools 7.0 for lexical–thematic analysis, they highlight the frequently occurring words and categorize them into themes according to their lexical–thematic groups. In agreement with Sinclair [15], they suggest that lexical units and keywords are important elements that contribute to constructing the semantic framework.

The analysis of collocation is expanded not only for categorizing themes and topics of a corpus, but also for linking these topics to the actions that relate to self-presentation in social media. Thelwall et al. [16] applied a thematic analysis to identify the influence of gender identities on self-presentation on Twitter. They argue that many studies have not fully covered the personal disclosers that discuss non-binary users in addition to male and female users. In comparison with other studies, they deal with deep thematic analyses as they consider other possible themes and topics related to the identification of the self. They use thematic analysis by identifying word association [16], a method to uncover differences between sets of texts by identifying words that occur together with high frequency in a subset. Their study discussed the themes and gender differences in classified users. They found that each type of male, female, and nonbinary user had 33 themes based on the analysis of the collocations.

The studies of thematic analysis using collocation networks did not fully consider how collocates are used in various contexts by looking at the concordance lines. The previous studies considered the strong quantitative measurement of the relationship between words and lexis as a central approach for thematic analysis. By doing this, they discussed the general topics of a corpus depending on the mathematical relationship. This means that the previous studies did not investigate the semantic shifts of keywords and their meanings according to their situational context in a text. This limitation suggests the need for a more detailed methodology for providing deep classification of the same topics of a corpus.

2.2. Long Short-Term Memory Networks

Using human resources to handle text classification is a costly process. This problem can be solved by using advanced algorithms to handle the text-classification process. However, using advanced algorithms would require several processes, including data gathering, pre-processing/cleaning, annotation/tagging, and, finally, building a predictive model to facilitate classifying new instances. With the recent increase in the availability of data, this would reduce the time taken by humans to classify data, as well as increase the efficiency of the analysis.

The machine learning approach with neural networks, specifically long short-term memory (LSTM), is effective in the field of text classification and natural language processing (NLP). In this research, we adopt LSTM for text classification to perform a thematic analysis using concordance lines that are taken from news articles. We suggest the importance of investigating the concordance lines of the articles rather than only the topics of the article because multiple themes may occur within an article. For example, an article may start with a main idea, but various subthemes may appear clearly in other parts of the article. Thus, we focus on the semantic shift of concepts and ideas in the whole corpus of news articles.

Long short-term memory (LSTM) has excelled recently in the field of artificial intelligence and deep learning [17]. It is an advanced method of using the recurrent neural network (RNN). Unlike RNN, which is unable to learn long-term dependencies [18], LSTM keeps track of the input stream for a long period of time. LSTM efficiently improves performance by memorizing the relevant information, and it uses this information for prediction. This process allows for the retention of information passed to neurons, specifically with regard to long sentences [19].

LSTM has been used to perform various classification tasks in NLP. Ranjan et al. [20] presented document classification tasks using LSTM. The experiment was performed on 20 categories using 20 newsgroup datasets. The accuracy obtained from the classification was approximately 92% [20]. Furthermore, Andrade et al. [21] used LSTM for different purposes, such as detecting different types of malware. They looked at software hexadecimal codes instead of dealing with textual data. They classified malware to five different types and obtained an accuracy that reached 67.60%

In addition, LSTM was used to identify sentence subjectivity using political and ideological datasets. In Al Hamoud et al. [22], an experiment was performed to identify each sentence that referenced a specific issue as either as a pro or con. They achieved an accuracy rate approaching 96%. However, these studies did not provide a methodology for a thematic analysis of a specific topic, in contrast to the current study, which investigates themes and topics of the media coverage for the Syrian chemical attack.

Other studies applied a general thematic analysis, and the classification was only on a surface level; it was not a deep thematic classification, as is the case in the current study. For example, the dataset of Jelodar et al. [23] was classified into two classes, either positive or negative, and their model achieved 81.15% accuracy. Further, LSTM has been used in sentiment analysis, reaching 85% accuracy [24]. These studies do not contribute to providing detailed and complex thematic analyses for a corpus around a particular topic.

This literature review has shown that previous studies highlight the importance of analyzing keywords and the collocations in a corpus to provide a full understanding of

the themes and topics of a corpus. These methodological perspectives are important for providing themes and topics and outlining the semantic structure of a text. The literature suggests that LSTM is an effective tool for classifying the data into groups or themes. However, these studies should consider deep analysis by investigating the changes in the meaning through the collocational network, such as with a deeper method as proposed by Brezina et al. [2]. Altameemi [3] suggests that considering the themes of a corpus should not be limited to the collocation network; concordance lines should be analyzed to consider how collocates are used in their specific contexts. Thus, the statistical and quantitative analyses are not enough to fully present the semantic shift of terms and concepts in a corpus. Considering this, Altameemi [3] manually analyzed a sample of the concordance lines for the strong collocates. However, he suggests it might be good to apply an automatic thematic analysis of all the concordance lines. Due to these limitations, the current research suggests the importance of developing a methodology for using algorithms for the automatic analysis of all the concordance lines of the strong collocates, which will increase the representativeness of analyzing a corpus.

3. Methodology

In this study, we test a proposed methodology that has not been adopted by any previous studies, according to the best knowledge of the researchers. The proposed methodology suggests applying a combination of algorithms to carry out the automatic thematic analysis of the concordance lines. In this section, we highlight the steps that the researchers applied to conduct the automatic thematic analysis. Then we present the data that was used in this research. Next, we demonstrate the experiment conducted in this research, along with the obtained results.

3.1. Architectures of LSTM Units

There are several architectures of LSTM units. We can observe the basic architecture of an LSTM memory block in Figure 1. A typical architecture of an LSTM unit consists of three gates: a forget gate, an input gate, and an output gate, in addition to the memory cell, which is part of the LSTM unit. Information is added or removed to the cell state through those gates. The forget gates determines what relevant information needs to be kept or removed during training. Values are determined according to the sigmoid function, typically 0 or 1. When the value is close to zero, then the information is ignored. Otherwise, when the value is close to one, then the information is kept. The input gate controls which information needs to be saved in the memory. The output gate determines which information should be exposed to the memory cell. Similarly, both gates are controlled by the sigmoid function, which determines whether the information needs to be processed or not [23].

3.2. Datasets and Pre-Processing

The data of the research are adopted from Altameemi's thesis [3], but for a different purpose: testing a proposed methodology. The data is taken from four UK newspapers: The Telegraph, The Guardian, The Sun, and The Mirror. The dataset consists of articles related to covering the use of chemical weapons in Syria. The concordance lines of the corpus are taken from the occurrences of the strong collocates that are pre-processed with Altameemi's thesis. These lines are manually analyzed by Altameemi according to the themes (for more details about the thematic analysis of the concordance lines, see Altameemi [3]). The themes are extracted according to their original classification in Altameemi's project. Four themes are generated and represent most lines in the dataset. The themes used in the experiment are "Evaluation of the Syrian situation", "International stance", "Representation of the UK and USA together", and "Local context of the UK". A total of 1062 lines are extracted that represent the themes, as shown below in Table 1. The total number of lines and tokens for each theme are listed below.

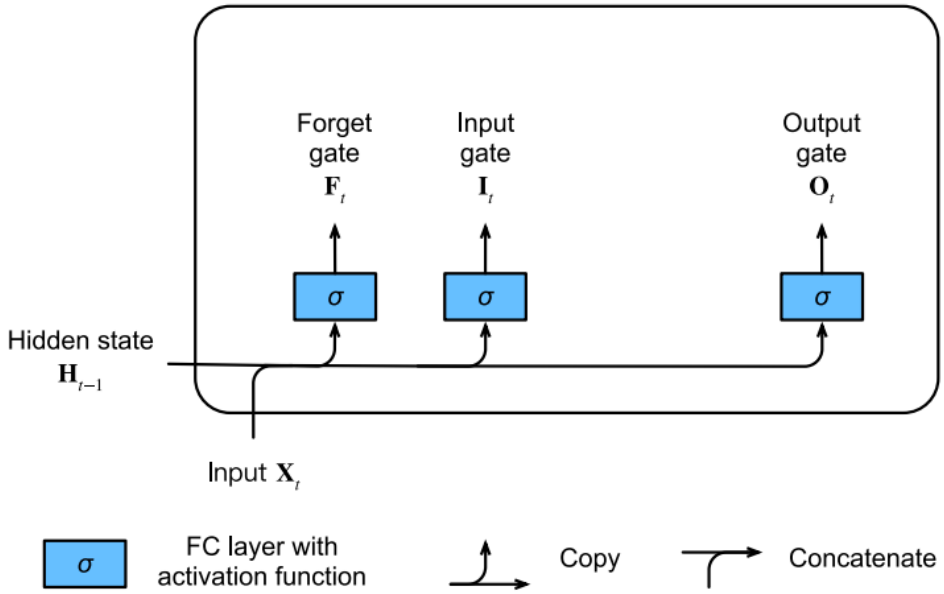


Figure 1. Architecture of the long short-term memory (LSTM) neural network (cited from [25]).

Table 1. Total number of concordance lines and tokens for each theme.

Themes	Number of Lines	Number of Tokens
Evaluation of the Syrian situation	250	8396.0
Representation of the UK and USA together	264	9230.0
Local context of the UK	250	9780.0
International stance	298	10,635.0

In the first theme, the lines focus on evaluating the Syrian situation with either the representation of the chemical attacks or the tragedy in Syria. The second theme focuses on the representation of the UK and USA together, specifically from the perspectives of supporting allies and their political orientations. The third theme represents the UK's political stance in relation to the Syrian chemical attack and the possibility of British military action. The last theme, "International stance", highlights the role of international agents such as the United Nations. Table 2 below shows a sample of these lines.

Table 2. Sample of thematic analysis from extracted KWIC.

Themes	Lines
Evaluation of the Syrian situation	It portrays ISIS as the main threat to Syrians, despite Assad killing at least six times more civilians.
Representation of the UK and USA together	Saturday August 24: David Cameron and USA President Barack Obama pledge a "serious response" to the Syria gas attack after a critical 30 min phone call between the pair at 4.30pm n which they decide to consider "all options".
Local context of the UK	Downing Street says, after a weekend of briefing that military intervention may be imminent.
International stance	A former UN appeal judge and leading international lawyer, said yesterday that Russia was wrong to insist that military intervention would hinge on UN consent.

In the current study, we suggest that analysts may take a random selection of concordance lines that might be representative of the results. Then, the rest of the concordance lines should be classified electronically according to the manual categorization of the se-

lected analyzed concordance lines. To test this method's validity, we used the same data that were used in Altameemi's thesis.

The lines are pre-processed with some processing steps. For example, numbers, special characters, stop words, and non-English letters are removed. Then, each line is converted into a sequence of tokens which serves as the input to the neural network. Each token represents a single word converted to a digit, and each digit is used for the training process. We chose the most frequent 5000 words. The chosen words represent the maximum features considered in the input layer of the neural network.

4. Experiments and Results

4.1. Experiments

The first layer in our model is an embedding layer. This layer receives the previously tokenized words using a dense vector representation. The order of the words processed is taken into consideration. The next layer is an LSTM layer with 100 neurons. A dropout layer is used for regulating the network and keeping it away, as much as possible, from any bias. The final dense layer is the output layer, which has four classes representing the four distinct themes in this experiment. All of the data processed for training are balanced among the four themes. The models are compiled using the Adam optimizer and sparse_categorical_crossentropy. The Adam optimizer is an ideal deep learning optimizer for handling classification problems [26,27]. For training and testing, we split our dataset into 80% training and 20% testing (849 lines for training and 213 lines for testing). The model was implemented in Keras, an open-source library that provides a high-level neural network API [28]. Table 3 shows the experimental configuration setting of our LSTM models.

Table 3. Details of long short-term memory (LSTM) model.

Layer	Output Shape	Number of Parameters
Embedding	(None, 520, 100)	5,000,000
Dropout layer	(None, 520, 100)	0
LSTM	(None, 100)	80,400
Dense	(None, 4)	404
Total parameters	5,080,804	
Trainable parameters:	5,080,804	
Non-trainable parameters	0	

Furthermore, in order to verify the effectiveness of choosing the LSTM model, we compared the results obtained with another deep learning model using a recurrent neural network (RNN). The experiment was performed using the same dataset and settings mentioned in Table 3. However, instead of using LSTM layers, we used the RNN layer (SimpleRNN), which handles information over time using a memory cell.

4.2. Evaluation Measures

After performing the experiment, we used several measures to evaluate our results. We calculated the confusion matrix for all the themes in the experiment. Each entry in the confusion matrix represents the following: the number of true positive (TP), which is the number of lines where the classification matches the correct theme; the false positive (FP) which represents the number of lines where the classification matches the incorrect theme as positive; true negative (TN), which is the number of cases where the classification matches the correct negative theme; and false negative (FN), which refers to the number of lines where the classification matches the incorrect negative theme.

We also evaluated the model to assess the performance using accuracy, precision, recall, and F1 score (Davis, 2006), based on the result of the confusion matrix.

Accuracy is the measurement of all lines that have been classified correctly. It consists of all the true positive and negative lines divided by all the lines being classified. It is defined as:

$$Accuracy = \frac{\#\,\mathrm{TP} + \#\,\mathrm{TN}}{\#\,\mathrm{TP} + \#\,\mathrm{FP} + \#\,\mathrm{FN} + \#\,\mathrm{TN}}.$$

Precision implies the number of selected lines that are correctly classified as positive. It consists of all the true positive lines divided by all the lines that are classified as positive. It is defined as:

$$Precision = \frac{\#\,\mathrm{TP}}{\#\,\mathrm{TP} + \#\,\mathrm{FP}}.$$

Recall indicates the proportion of actual classes that are correctly categorized as positive. It consists of all the true positive lines divided by all the lines that are actually positive. It is defined as:

$$Recall = \frac{\#\,\mathrm{TP}}{\#\,\mathrm{TP} + \#\,\mathrm{FN}}.$$

F1 score is the combination of both precision and recall. It is defined as:

$$F1 = 2 \times \frac{Recall \times Precision}{Recall + Precision}.$$

4.3. Results

In this section, we first present our classification results in all the measures mentioned above. Then we analyze the misclassification lines that were incorrectly classified by the modes. Lastly, we discuss our findings and limitations.

Table 4, below, presents the results of the thematic analysis of concordance lines using both long short-term memory (LSTM) and recurrent neural network (RNN). The result shows the high accuracy of using LSTM, reaching 84%, in comparison to the 79% accuracy achieved using RNN

Table 4. Results of using deep learning modes LSTM and RNN in thematic analysis of concordance lines.

Measures	LSTM	RNN
Precision	0.84	0.79
Recall	0.84	0.79
F1 score	0.83	0.78
Accuracy	0.84	0.79

Table 5, below, shows the results of the automatic thematic analysis of concordance lines. The results show a comparison between both LSTM and RNN for the four identified themes. We can clearly see that LSTM outperforms RNN within each theme, with regard to the accuracy of classifying the concordance lines. The lines are categorized into four themes: (1) Evaluation of the Syrian situation, (2) International stance, (3) Representation of the UK and USA together, and (4) Local context of the UK. The obtained accuracy was tested on unseen concordance lines.

Using the LSTM, the "Evaluation of the Syrian situation" theme achieved 100% recall without misclassifying any lines. This is the same result achieved by RNN. Next, the "Representation of the UK and USA together" theme achieved the second-highest recall of 94%, with only three lines having been misclassified. RNN achieved a very close result, reaching a recall of 93%. Furthermore, the "International stance" theme achieved a low recall of 69%, with 16 lines having been misclassified. RNN achieved the worst result, reaching a recall of 44%.

Table 5. Results of the text automatic thematic analysis over four themes using LSTM and RNN.

Themes	**Evaluation of the Syrian Situation**		**International Stance**		**Representation of the UK and USA Together**		**Local Context of the UK**	
Measures	**LSTM**	**RNN**	**LSTM**	**RNN**	**LSTM**	**RNN**	**LSTM**	**RNN**
Precision	0.85	0.74	0.72	0.77	0.88	0.94	0.89	0.72
Recall	1.00	1.00	0.69	0.44	0.94	0.93	0.71	0.79
F1 score	0.92	0.85	0.71	0.56	0.91	0.93	0.79	0.75

In terms of precision, the "Local context of the UK", "Representation of the UK and USA together", and "Evaluation of the Syrian situation" themes achieved high precision using LSTM, with results of 89%, 88%, and 85%, respectively. The "International stance" theme achieved a low precision of 72%, and 14 lines were classified as belonging to the "Local context of the UK" theme. Generally, RNN achieved a lower precision score than LSTM. For the themes of "International stance" and "Representation of the UK and USA together", the RNN obtained slightly better results (77% and 94%, respectively) than LSTM. However, LSTM obtained better results for the themes of "Evaluation of the Syrian situation" (74% in RNN) and "Local context of the UK" (72% in RNN). Considering the overall results in Tables 4 and 5, we can see how LSTM is more accurate, in general, for the thematic analysis than the RNN model.

The confusion matrix in Table 6 shows how many items match the correct categorization and how many lines match incorrect themes. Thus, we consider the percentage of the accuracy for each theme rather than only the total amount of accuracy. Table 4, below, shows that the categorization of the lines that relate to the first and third theme have the highest correct categorization among the themes. The fourth theme, "Local context of the UK", comes after the first and third themes in the accuracy of categorizing the lines. However, the "International stance" them has the lowest accuracy of categorization. In this section, we discuss how the lines are categorized in each theme and the instances of incorrect categorization, which highlights the effectiveness and issues of the proposed methodology.

Table 6. Confusion matrix for thematic analyses using LSTM model.

		Predicted Label			
		Evaluation of the Syrian Situation	International Stance	Representation of the UK and USA Together	Local Context of the UK
Truth Label	Evaluation of the Syrian situation	51	0	0	0
	International stance	6	36	5	5
	Representation of the UK and USA together	0	3	51	0
	Local context of the UK	3	11	2	40

As shown in the table above, the theme of evaluating the Syrian attack was the most accurate theme, which matches the exact categorization. All of the categorized lines meet the original classification. Thus, for this theme, there is no discussion of the incorrect instances.

The second most accurate theme was the "Representation of the UK and USA together" theme, which had three incorrect classifications. For example:

1. John Kerry administered a diplomatic slap in the face to Britain following David Cameron's withdrawal of military support for intervention in Syria.

In the experiment, Example 1 was categorized under the theme of "International stance", while it originally related to "Representation of the UK and USA together". Categorizing such examples is difficult, as this theme seems to be in the blurred area between

more than one theme (this point will be expanded below, with the consideration of the other themes). This example and the other lines in the corpus show the high accuracy of classifying the lines of this theme, even though the algorithm missed three instances.

A less-accurately categorized theme was "Local context of the UK". The algorithm missed the correct categorization of 16 lines. From the lexical perspective, regarding their semantic use, words such as "UK" and "Britain" were frequently used both for the representation of the local context and also from the perspective of the international stance. For example:

2. International community needs to ask: would Assad utilize chemical weapons stage bring potential western military intervention international community to investigate the reasons behind the use of chemical weapons.
3. Later came comments from the USA House speaker John Boehner backing a Syria war resolution, adding to the likelihood of Congress voting for USA action.

Examples 2 and 3 were classified as a part of the UK stance, while they were originally categorized in the theme of "International stance". Lexically, words such as "international", "community", "internationally", and "prohibited" were frequently used with the "International stance" theme. In Example 4, the algorithm considered terms such as "voting" and "resolution" as strongly connected to the local context, while the semantic function of the whole line suggested the connection of the example to the "International stance" theme.

The least accurate theme in the categorization was the "International stance" theme. Table 5, above, shows that the "International stance" theme had various incorrect classifications for the different themes. One reason might be that the "International stance" theme had strong terms reoccurring both in this theme and, at the same time, in other themes. The incorrect lines in the "International stance" theme have almost the same issues, for example:

4. He (General Sir Nick Houghton) revealed no decisions have been made on military involvement in Syria.
5. We have seen the unwinnable nature of the Afghan conflict. The terrible sores of the Balkan civil wars are still raw enough to remind us of what little effect our intervention had there.

Examples 4 and 5 were classified under the "International stance" theme, while they were originally categorized under the "Local context of the UK" theme. Terms such as "Syria", "Afghan", "Balkan", and "civil" strongly occurred in both themes. In the local context, they were used to refer to the social imagination of the country regarding the UK's experience in international interventions, such as in Iraq and Afghanistan. At the same time, these terms were also used to represent the role of the United Nations and other actors, such as the US. Thus, these findings are consistent with those of Altameemi (2020), in that some lines are in the blurred area between more than one theme.

5. Discussion and Conclusions

The main contribution of the article is in applying an automatic thematic analysis of concordance lines in a corpus using LSTM. The proposed methodology in this study is useful in various ways. First, the automatic categorization considers the aboutness of the corpus by the relationship between collocates and keywords, as well as their occurrences within the context of the whole text. The contribution this article provides is a deeper automatic thematic analysis than the previous studies, as discussed in the literature review. Linguists such as Brezina et al. [2] and Neg and Tan [10] focus only on the analysis of the aboutness of the corpus by quantitively analyzing the collocational network. On the other hand, researchers in the field of computer science, such as Gao et al. [19] and Ranjan et al. [20], apply classification and sentiment analysis of the whole text without looking at the semantic shifts of keywords in their concordance lines. The proposed method deals with how the collocates are used in their specific context through a thematic

analysis of the concordance lines. The categorization of the concordance lines determines which themes are given more concern than the others in the whole corpus.

Second, the findings of this paper help to reassert the original result of the manual analysis of the randomly selected lines by looking at some other lines in the automatic categorization. In this study, when we checked some of the automatically categorized lines, we found them in different categories from the ones they had been place in during the original categorization. However, in semantic use, the automatic categorization seems to be more accurate than the original, manual categorization. This automatic thematic analysis method fills the gap proposed by Altameemi [3] that the collocation analysis does not provide a deep interpretation of the aboutness of a corpus. Thus, this algorithm may help analysts to revisit their manual categorization of some lines.

Regarding the specific use of LSTM in thematic analyses, we have seen in the literature that this methodology is used for various purposes, such as sentiment analysis and document classification. Applying the method proposed in the current study will increase the efficiency of the result through a full analysis of the concordance lines, specifically for studies that investigate complex types of themes.

The findings of the research suggest the effectiveness of the proposed methodology for the thematic analysis of media coverage. This method is applied to provide a deep thematic analysis. In other words, the current proposed methodology divides the concordance lines into themes in specific orders of discourse and topics (e.g., themes in the UK press coverage for the Syrian chemical attack).

The proposed method of this research should take into account the following considerations. First, in applying this method, analysts need to look at the lines of the collocates and categorize a random selection of lines manually to train the algorithm before the automatic categorization. Further, analysts should apply the confusion matrix to test the algorithm before conducting the whole analysis. Then, the incorrect classifications should be reconsidered to improve the algorithm and validate the automatic analysis. For future research, it is recommended that the automatic categorizing should name the themes according to the most salient relationship between collocates/words rather than using the manual naming of themes in the manual analysis of the concordance lines.

Author Contributions: Conceptualization, Y.A. and M.A.; methodology, Y.A. and M.A.; software, M.A.; validation, Y.A. and M.A.; formal analysis, Y.A.; investigation, M.A.; resources, Y.A. and M.A.; data curation, M.A.; writing—original draft preparation, Y.A. and M.A.; writing—review and editing, Y.A.; visualization, M.A.; supervision, Y.A.; project administration, Y.A.; funding acquisition, Y.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research has been funded by the Deputy for Research and Innovation, Ministry of Education through the Initiative of Institutional Funding at University of Ha'il-Saudi Arabia through project number IFP-22 050.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: This research has been funded by the Deputy for Research and Innovation, Ministry of Education through the Initiative of Institutional Funding at University of Ha'il-Saudi Arabia through project number IFP-22 050.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Baker, P. *Using Corpora in Discourse Analysis*; Bloomsbury Academic: London, UK; New York, NY, USA, 2006.
2. Brezina, V.; Mcenery, T.; Wattam, S. Collocations in context A new perspective on collocation networks. *Int. J. Corpus Linguist.* **2015**, *202*, 139–173. [CrossRef]
3. Altameemi, Y. *Defining 'Intervention': A Comparative Study of UK Parliamentary Responses to the Syrian Crisis-ORCA*; University of Cardiff: Cardiff, UK, 2020.

4. Wodak, R. The semiotics of racism: A Critical Discourse-Historical Analysis. In *Discourse, of Course: An Overview of Research in Discourse Studies*; Renkema, J., Ed.; Benjamins: Amsterdam, The Netherlands, 2009; pp. 311–326.
5. Fernández, B.G.; Schmitt, N. How much collocation knowledge do L2 learners have?: The effects of frequency and amount of exposure. *Int. J. Appl. Linguist.* **2015**, *166*, 94–126.
6. Nesselhauf, N. Collocations in native and non-native speaker language. In *Collocations in a Learner Corpus*; John Benjamins Publishing Company: Amsterdam, The Netherlands, 2005; Volume 14, pp. 1–10.
7. Paquot, M.; Granger, S. Formulaic Language in Learner Corpora. *Annu. Rev. Appl. Linguist.* **2012**, *32*, 130–149. [CrossRef]
8. Evert, S. Corpora and collocations. In *Corpus Linguistics: An International Handbook*; Lüdeling, A., Kytö, M., Eds.; Mouton de Gruyter: Berlin, Germany, 2008; pp. 1212–1248.
9. Gablasova, D.; Brezina, V.; McEnery, T. Collocations in Corpus-Based Language Learning Research: Identifying, Comparing, and Interpreting the Evidence. *Lang. Learn.* **2017**, *67*, 155–179. [CrossRef]
10. Ng, R.; Tan, Y.W. Diversity of COVID-19 news media coverage across 17 countries: The influence of cultural values, government stringency and pandemic severity. *Int. J. Environ. Res. Public Health* **2021**, *18*, 11768. [CrossRef] [PubMed]
11. Huang, L.; Dou, Z.; Hu, Y.; Huang, R. Textual analysis for online reviews: A polymerization topic sentiment model. *IEEE Access* **2019**, *7*, 91940–91945. [CrossRef]
12. Bondarchuk, N.; Bekhta, I.; Melnychuk, O.; Matviienkiv, O. Keyword-based Study of Thematic Vocabulary in British Weather News. *CEUR Workshop Proc.* **2022**, *3171*, 451–460.
13. Biber, D.; Reppen, R. *The Cambridge Handbook of English Corpus Linguistics*; Cambridge University Press: Cambridge, UK, 2020.
14. Huang, L.; Dou, Z.; Hu, Y.; Huang, R. Online Sales Prediction: An Analysis with Dependency SCOR-Topic Sentiment Model. *IEEE Access* **2019**, *7*, 79791–79797. [CrossRef]
15. Sinclair, J. *Corpus, Concordance, Collocation*; Oxford University Press: Oxford, UK, 1991.
16. Thelwall, M.; Thelwall, S.; Fairclough, R. Male, Female, and Nonbinary Differences in UK Twitter Self-descriptions: A Fine-grained Systematic Exploration. *J. Data Inf. Sci.* **2021**, *6*, 1–27. [CrossRef]
17. Dou, Z.; Sun, Y.; Zhang, Y.; Wang, T.; Wu, C.; Fan, S. Regional Manufacturing Industry Demand Forecasting: A Deep Learning Approach. *Appl. Sci.* **2021**, *11*, 6199. [CrossRef]
18. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
19. Gao, M.; Shi, G.; Li, S. Online prediction of ship behavior with automatic identification system sensor data using bidirectional long short-term memory recurrent neural network. *Sensors* **2018**, *18*, 4211. [CrossRef] [PubMed]
20. Ranjan, M.N.M.; Ghorpade, Y.R.; Kanthale, G.R.; Ghorpade, A.R.; Dubey, A.S. Document classification using lstm neural network. *J. Data Min. Manag.* **2017**, *2*, 1–9.
21. Andrade, E.D.; Viterbo, J.; Vasconcelos, C.N.; Guérin, J.; Bernardini, F.C. A model based on LSTM neural networks to identify five different types of malware. *Procedia Comput. Sci.* **2019**, *159*, 182–191. [CrossRef]
22. Al Hamoud, A.; Hoenig, A.; Roy, K. Sentence subjectivity analysis of a political and ideological debate dataset using LSTM and BiLSTM with attention and GRU models. *J. King Saud Univ. Inf. Sci.* **2022**, *34*, 7974–7987. [CrossRef]
23. Jelodar, H.; Wang, Y.; Orji, R.; Huang, S. Deep Sentiment Classification and Topic Discovery on Novel Coronavirus or COVID-19 Online Discussions: NLP Using LSTM Recurrent Neural Network Approach. *IEEE J. Biomed. Health Inform.* **2020**, *24*, 2733–2742. [CrossRef] [PubMed]
24. Murthy, G.S.N.; Allu, S.R.; Andhavarapu, B.; Bagadi, M.; Belusonti, M. Text based sentiment analysis using LSTM. *Int. J. Eng. Res. Tech. Res.* **2020**, *9*, 299–303.
25. Zhang, A.; Lipton, Z.C.; Li, M.; Smola, A.J. Dive into deep learning. *arXiv* **2021**, arXiv:2106.11342.
26. Rao, A.; Spasojevic, N. Actionable and political text classification using word embeddings and LSTM. *arXiv* **2016**, arXiv:1607.02501.
27. Fatima, N. Enhancing Performance of a Deep Neural Network: A Comparative Analysis of Optimization Algorithms. *ADCAIJ Adv. Distrib. Comput. Artif. Intell. J.* **2020**, *9*, 79–90. [CrossRef]
28. Géron, A. *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow*; O'Reilly Media, Inc.: Sebastopol, CA, USA, 2022.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.