Sentiment Analysis of English Newspapers: A Corpus-based study

Dr. Muhammad Alrayes

College of Language Sciences, English Department at

King Saud University, Saudi Arabia

المستخلص:

كانت الأساليب النظرية والمنهجية المستخدمة في تحليل البيانات هي علم اللغة (CL) ونهج تحليل المشاعر كوسيلة لتحليل الخطاب الإعلامي للصحف وتمثيل المشاعر في المملكة العربية السعودية. كان استخدام التكنولوجيا في استخراج البيانات لهذه الدراسة أمرًا لا بد منه بسبب الكم الهائل من نصوص الشركات التي تم التحقيق فيها في البحث الحالي. تم إجراء المصدر الرئيسي لجمع البيانات وتحليلها من خلال محرك الرسم (SE). أيضًا، تم استخدام لغة البرمجة R أثناء عملية التحليل التي توفر العديد من المكتبات للترميز التي تسهل إجراء هذا النوع من الأبحاث بما في ذلك: Tidyverse و Tidytext و Syuzhet و Textstem و Ggplot2 و Redxl و Writexl. تم استخراج الصحف corpora من مجموعة بيانات SE بين الفترة 1993–2013. أخيرًا، تم تصنيف النتائج بطريقتين: إحداهما لكل صحيفة طوال الفترة الزمنية، والثانية كانت للبيانات الشخصية للصحف بأكملها معًا.

الكلمات المفتاحية: تحليل المشاعر ؛ وعلم اللغة ؛ CDA ؛ الخطاب الإعلامي ؛ التعلم الآلي ؛ رسم المحرك.

Abstract

The theoretical and methodological approaches deployed in analyzing the data were corpus linguistics (CL) and sentiment analysis approach as a way of analyzing newspapers media discourse and sentiment representations of Saudi Arabia. The use of technology in extracting data for this study was a must due the huge amount of corpora texts which were investigated in the current research. The main source of data collection and analysis was done through *sketch engine (SE).* Also, the R programming language was used during the analysis process which provides several libraries for coding that makes it easier to conduct this type of research including: Tidyverse, Tidytext, Syuzhet, Textstem, Ggplot2, Readxl and Writexl. The newspapers corpora were extracted from SE dataset between the period of 1993-2013. Finally, the findings were classified in two ways: one was for each newspaper for the entire time period, and the second was for the entire newspapers' corpora data together.

Keywords: Sentiment Analysis; Corpus Linguistics; CDA; Media Discourse; Machine Learning; Sketch Engine.

1. Introduction

This study overviews the media discourse of English-speaking newspapers primary from the Western countries about Saudi Arabia using corpus-based analytical approach. In today's world the contact between different cultures became undeniable reality; therefore, it is one of the major roles of scholars from different fields including linguistics to bridge such gaps. One of the best ways to approach such issues is through studying and analyzing discourse of the media about how each culture views the other. The reviewing of literature suggests a lack of up-to-date studies that tackle such issues of media discourse in the recent years. Thus, this study comes to analyze a more recent discourse of Western media within a linguistically oriented perspective using both corpus Linguistics CL and sentime`12nt analysis approach respectively. This study takes its importance from various dimensions including: the historical relation dimension between civilizations, the involvement of technological tools to analyze media discourse, and the linguistic analysis of sentiments of the selected corpora texts.

2. Literature Review

2.1. Previous studies

There have been a number of studies on the media discourse of different minorities, ethnicities and/or other nations or countries. One of the prominent examples was the representation of black ethnicity which has received the highest number of research studies compared to other groups. As far as Saudi Arabia is concerned being the main theme of investigations in this study; not much has been done about Saudi Arabia representation in the media discourse. However, as being part or the center of the Islamic world, we found it useful to include some studies that tackle the media discourse about

Islam/Muslims representation either as minority groups in the West or as other countries spread all over the world. Starting with the British context, Poole (1) investigates the representation of Islam in The Guardian and The Times newspapers during the period 1994-1996. Poole's results indicated that representation of Islam in those newspapers was frequently connected with themes such as: `Islam and Segregation of Women`, `Fundamentalism`, `Immigration`, and `Criminal Activities`. The results also stated the predominant discourse addressing those themes was highly negative. Islam was often associated with primitivism, homogeneity, restriction, and irrationality (1). Six years later, Poole in 2006 conducted another study of the same newspapers for the year of 2003 and the results showed a dramatic increase in the themes of 'terrorism and Islam' after the events of 9/11 which surpass all other themes in the previous study. Both of Poole's studies are of significant contributions to the field, however; both studies are outdated and focused only on representation of Muslims, so there is a need for investigating more recent media discourse of Arab countries particularly Saudi Arabia. A more recent study about the representation of Muslims in the British press was done by Paul Baker et al (2). This study was based on a 143 million-word corpus that includes 200,000 articles from 11 broadsheet and tabloid newspapers during the period 1998-2009. Similar to the previous study's results, this study's findings stated that Muslims are frequently represented in contexts of conflicts and associated with words such as: `terrorism` and `extremism`. It is very interesting to find out the word `terrorism` appeared in the data more than the word `Islam` even though it was not among the query words in the corpus compared to `Islam` being one of them. This observation supports the results of both studies by Poole above in that the representation of Islam was highly negative. This study gives a significant indication identifying the approximate position of the British press in representing of Muslims through both its long time period of data research and the huge amount data as well. Yet this study is

going to be far more focused on Saudi Arabia rather than the whole Muslims such as Poole's and Baker's studies.

In the American context, Nacos et al conducted a study about Muslims representation in the United States in (3). This study focuses on newspapers published in New York: New York Times, Daily News, and New York Post. The analysis includes 867 articles classified as news analysis, editorial, opinion columns, and straight reporting. The classification also includes whether the news topics were local or international, and representation classified as: negative, positive or neutral, and the time period for this research is between September 2000 to March 2002 (3). The findings showed that the number of stories published about Muslims after 9/11 in the four newspapers went up by eleven times compared to the six months before 9/11. However, the analysis indicated the representation of Muslims in the period following 9/11 was depicted in a more positive image (3). This study's result, in spite of its short period of data, is unexpected compared with previous studies in the British context.

These studies about representation of Islam and Muslims in the West give an indication about how the image of Islam in the West is being portrayed, and toward which specific trend it is being directed. Said (4) believed such discourse is not surprising during the medieval or colonial periods however it is astonishing to see similar discourse still exists until today even among academic writings. Jericho (5) drew similar conclusion about how strange it is for Western media to represent Islam and Muslims (which counted to be more than one fifth of the world's whole population) as an isolated culture with strange values, and is difficult to understand and/or negotiate with. Said (6) suggested that one should not rely solely on mass media as the main source of information, there are more reliable sources like books, lectures and academic articles available and have more

validity than the media. Said also believed that not all mass media sources stand on the same level, rather there is a variety of content and viewpoints among their production (6).

2.2 Sentiment polarity

According to Schulder *et al* (7), one of the best references' sentiment analysis particularly with regard to the survey on negation modeling is by Wiegand *et al*. (8). However, a closer look at the works that address polarity offers for instance the most complex general negation lexicon was published by Wilson *et al.* (9), moreover, the list extends to Szarvas *et al.* (10), Morante (11), and Choi and Wiebe (12) cited in Schulder *et al* (7).

On the use of corpora for sentiment analysis and polarity classification, The *Stanford Sentiment Treebank* (SST) (13) contains compositional polarity information for 11,855 sentences, as well as Schulder *et al*. (14), in which we bootstrap a lexicon of 980 English verbal shifters and evaluate their use for polarity classification cited in Schulder *et al* (7). A year later, Schulder *et al.* (15) created a lexicon of 1220 English verbal shifters and assigning shifter labels for individual word senses, whereas Wiegand, Loda, and Ruppenhofer (16) examine such sense-level information for shifting-specific word sense disambiguation cited in Schulder *et al* (7).

2.4. Media discourse Studies about Saudi Arabia

There are not many studies about Saudi Arabia in the Western media and as far as we are aware. At least, there is no study that has tackled this topic using sentiment analysis and corpus linguistics approach. Most of the materials about this topic are conference papers, opinion articles, television interviews, etc. including Keith Rowe (2003), Muna Osamah (2010), Hussain (2004) and Abdullah (2004). However, rather

than being not sophisticated enough to the level of academic standards, these writings address the topic from purely media perspective. Only two studies discussed the theme of the image of Saudi Arabia after the event of 9/11: Aldakkan (17) and Almutairi (18).

Starting with Aldakkan's study (17) of the portrayal of Saudi Arabia in Australian newspapers, he looked at the image of Saudi Arabia in the Daily Telegraph, an Australian newspaper which is produced every day except Sundays. Aldakkan began by looking at the frequency of mention four different countries (Saudi Arabia, Egypt, Israel and United States) in the most widely circulated Australian newspapers the Daily Telegraph. The focused time period was one year before and one year after the event of 9/11. The result showed an increase in the frequency of mention of Saudi Arabia post 9/11 compared to pre-9/11 period with a percentage increase of 300%, whereas the frequencies of the other countries were far lower, between 10%-20% difference (17). That was sufficient stimulus for the researcher to go on analyzing the discourse about Saudi Arabia in the English language newspapers. Another observation in the newspaper with regard to lexical choice is the term "Wahabi", the collocation of this term appeared more frequently in the year post 9/11 compared to the year before. That is, it was mentioned only once in the year before within a cultural and religious context whereas in the year post 9/11 use went to 7 times up, and 24 times up in the year of 2003 (17).

The second study which looks at the portrayal of Saudi Arabia after the events of 9/11 was by Almutairi (18). He investigated the Saudi Arabian image in three Israeli newspapers. These newspapers represent different political parties, and the targeted time period for this research was the year after the event of 9/11. Although the researcher followed a combined quantitative and qualitative method, the short amount of time the

study covered was a limitation. Choosing newspapers from different political parties made the study more representative, and that was clear in the data analysis part of the study. The researcher investigated various issues in the newspapers including, for instance, an analysis of the Saudi-American relations. The results for this topic showed that the newspapers had two analyses: the first one emphasized the friendly-strong Saudi-American relation, as well as the United States considered Saudi Arabia as a strategic ally especially in the Middle East region. The second emphasized the cold relation between the two countries particularly after the event of 9/11 (18).

Each perspective has its own evidences supporting their point of view. The first one, the healthy relation between the two countries, which was supported by several points such as Saudi Arabia's role in the Israeli-Palestinian conflict, which is supported and welcomed by the United States; the United States military bases in Saudi Arabia; and the fact that Saudi Arabia has a quarter reserves of the whole world's oil. The other side of the spectrum saw the Saudi-American relation as an unstable one, for the following reasons: the Saudi-US alliance is seen as not being steady and almost ended after the attack of 9/11, the constant impositions from the United States on the Saudi government to make changes in the school curricula the way they like, the majority of Saudi religious scholars hate the West, and other points (18).

The researcher investigated various topics in the newspapers which discuss Saudi Arabia in one way or another, and as he found out in the example above different opinions, he found similar trends with other topics as well in the way of sentiment representation; i.e. positive vs. negative (18). He concluded his research analysis stating that the image of Saudi Arabia in the Israeli newspapers was in the overall picture negative, taking into consideration the higher percentage of

negativity expressed against Saudi Arabia. However, this does not mean there were no general positive viewpoints about Saudi Arabia; rather, some of the positive examples include: the position of Saudi Arabia against extremism, the improvement of Saudi women's rights in society, and the Saudi Arabian role in the Israeli-Palestinian conflict. The general negative image of Saudi Arabia in Israeli newspapers was represented in the following: the constant financial support the Saudi government provided to terrorist Palestinian groups, economic corruption, the representation of the Saudi woman as a "victim", and the high focus on the Saudi-US dispute as a complicated and cultural based not just a simple one (18). The importance of this study is its breadth in the case of the number of subjects and areas of analysis it covered. That is, its qualitative and quantitative method helps the study observing the Saudi Arabian representation in the Israeli newspapers from fairly wide perspectives and describing both the negative and positive sides of the image

3.1. Statement of the Problem

Reviewing the literature above suggests a lack of work in the area of media discourse about Saudi Arabia particularly from corpus linguistic and discourse analysis; or more precisely sentiment analysis. The significance of such study is twofold: first its originality and novel as my review of literature suggests that there is no study ever has attempted an analysis of media discourse about Saudi Arabia from sentiment analysis perspective. Secondly, the fast-growing influence of media in the last decade on framing the perception of both ordinary people and decision makers as well.

3.2. Significance of the Study

It is hoped that this study will add to the field as it is, according to the researchers' limited knowledge, the first one considering its methodology. It is hoped that this study will increase awareness among people working in academia in the areas of Corpus linguistics and media discourse in general.

3.3. Research Questions

The following research questions have been revised and answered:

1. How is Saudi Arabia represented in the English-speaking media particularly from sentiment analysis perspective?

2. How can the findings of each newspaper on one hand and the whole data of the collective newspapers on the other hand be explained and compared to each other? How are they different from each other?

4. Methodology

4.1. Corpus linguistics (CL)

A corpus is described as a collection of naturally occurring linguistic features. According Hunston (19), a corpus refers to electronic collections of stored textual materials. The aim of creating most of existed corpora is to represent a variety of particular language, therefore these corpora deal with limited sample of texts rather than whole ones.

Corpus linguistics (CL) deals with specific software packages which analyze and calculate data in a language. The significance of such software comes from the ability to bring up specific comparative texts in a statistical format, for example focusing on word frequency, which will help researchers identify

linguistic patterns and trends in a given text. Some common concepts in corpus linguistics include collocations, concordance, and key words.

4.2. Critical Discourse Analysis

Critical discourse analysis as a field goes under the umbrella of discourse analysis in general. According to leading advocates in the field Wodak (20) and van Dijk (21), critical discourse analysis is different from other approaches to discourse analysis particularly in its main focus, i.e. CDA primary work is on the studying of prejudice, discriminatory, and anti-semitic discourse. According to Blommaert (22) critical discourse analysis is different from other types of discourse analysis in that it has been "groundbreaking in establishing the legitimacy of a linguistically oriented discourse analysis firmly anchored in social reality and with a deep interest in actual problems and forms of inequality in societies" (p.3). The term 'critical' is of great significance according to scholars in CDA. Fairclough (23) page:9 states that the term 'critical' indicates 'intervention, for example providing resources for those who may be disadvantaged through change'. Also, Wodak and Meyer (24) state that 'critical theories, thus CDA, want to produce and convey critical knowledge that enables human beings to emancipate themselves from forms of domination through self-reflection'.

Critical discourse analysis is known as being an interdisciplinary approach which investigates the way language in context reflects "power, injustice, abuse, and political-economic or cultural change in society" (25) p:357. According to Kress (26) p:85, CDA tries to show "the imbrications of linguistic-discursive practices with the wider socio-political structures of power and domination". CDA has focused its research approaches on tackling different issues varied between

unequal power relations and prejudice against minorities. It seems that CDA's main focus is on political texts, speeches, newspaper editorials etc. and studies about Islam and particularly Arab countries' representation are rare.

The goal of critical discourse analysis as it seems from its advocates' quotes above is to show unfair use of power and inequality through discourse as well as to link between discourse and society and more importantly societies' concerns and immediate life issues. In most of CDA's studies we could easily find a direct link between the linguistic phenomenon and its relation to people's life. Examples to this would include the representation of immigrants, minorities, ideology, racism, levels of power and classes in societies. According to Pennycook (27) p: 82, the goal of critical discourse analysis is to "look in discourse for manifestations of ideology … this approach makes sense for those coming from linguistic background, for whom language and discourse are elements that need to be related to larger concerns such as society and ideology"

4.3. Procedure

Firstly, we started with concatenating the text columns in order to have a resume text, then the selection of the columns which contain the searched data words in the newspapers' texts. Finally, classifying data according to each newspaper and storing it into excel files.

The second stage was reading the excel files one by one. Then, converting everything to lowercase, removing prepositions, conjunctions and punctuation characters. The reason behind this is to prevent the noise in the text, and to perform accurate analysis on this text. After that, we converted the data frame to one word per row, this process called tokenization. Furthermore, another process called lemmatization was applied which aims to get each word back to its root; e.g. (*is)*

gets back to (*be)*. As a result of the processing, we cleaned up the data to avoid any disturbance to the analysis process; such cases were excluded from the data frame to avoid troubling and keep the work on track. Lastly, we joined the words with a sentiments' dictionary and count the number of each category and did all possible visualizations of the results in the form of graphs and tables for all newspapers.

4.4. Dataset labeling

The dataset labeling process employed in this stage of data analysis was done in order to obtain clarity of data classifications, accuracy as well as being self-explanatory. Polarity classification was used as labels were classified into ten polarities: two main label records which are *positive* and *negative* records. The remaining eight label records include: trust, fear, anticipation, anger, sadness, joy, surprise and disgust as can be seen in the following table 1.

Table 1: Dataset sentiment labelling (ALL NEWSPAPERS)

Sentiment	**Frequency**
positive	24547
negative	17032
trust	16235
fear	11462
anticipation	10843

anger	8818
sadness	6577
joy	6015
surprise	4517
disgust	3581
Total	109627

After this process of classifying all newspapers data as one group, each newspaper was analyzed separately using the same labeling records mentioned above. The first newspaper was New York Times as the following Figure 1 states:

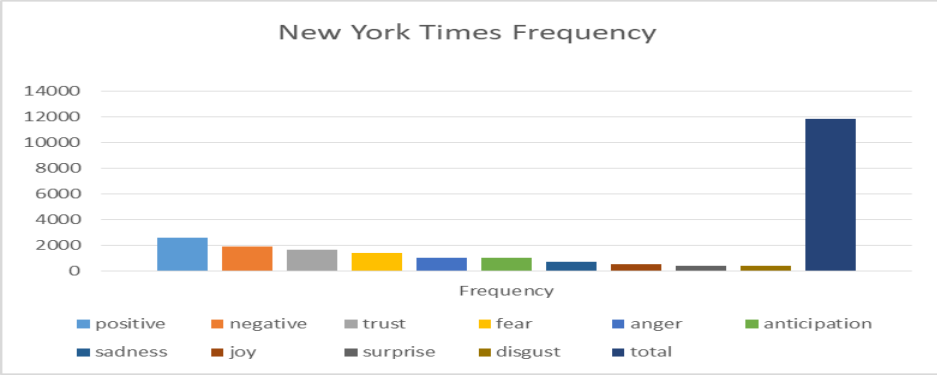


Figure 1 New York Times Sentiment Analysis frequency

The second newspaper was the Daily Mail as the following figure 2 states:

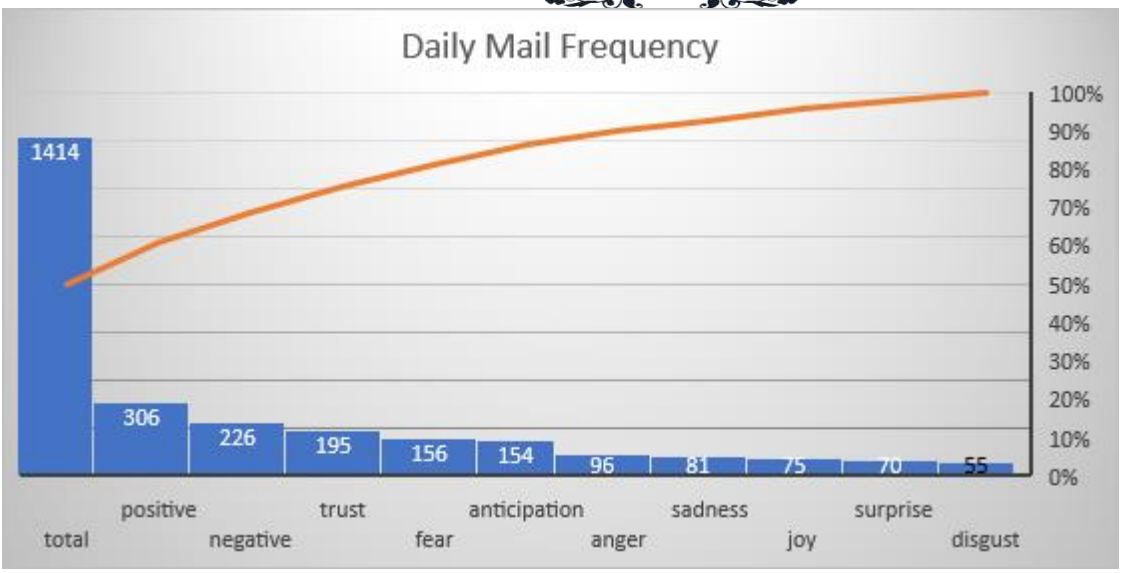


Figure 2 Daily Mail Sentiment Analysis frequency

5. Study Results and Discussion

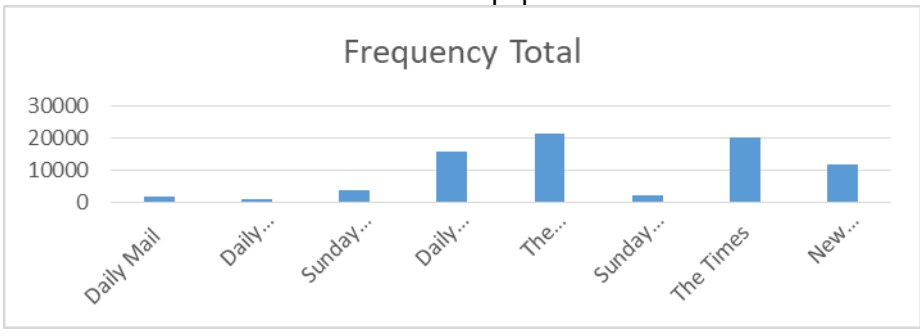
The results of this study, as table 2 below states, showed that all the newspapers selected on this corpus sentiment analysis have slightly more positive sentiment than negative throughout the use of the word *Saudi Arabia.* It is, however, true that there was a variation on the total amount of using the word *Saudi Arabia* from one newspaper to another. For example, the Guardian and the Times have talked about *Saudi Arabia* around 20000 times during the period between 1993-2013. On the other hand, Daily Mirror, Daily Mail, Sunday Telegraph and Sunday Times have used the word *Saudi Arabia* only less than 5000 times during the entire time period which tell us about the importance of such topic to each newspaper separately. Finally, Daily Telegraph and New York Times were in the middle range of mentioning the *Saudi Arabia* during the same period

Table 2: detailed findings

Newspapers				**Sentiment/Frequency**							
Daily Telegraph	Positive/ 3334	Negative/ 2518	Trust/ 2210	Anticipation/ 1559	Fear/ 1558	Anger/ 1400	Joy/ 948	Sadness/ 892	Surprise/ 610	Disgust/ 505	Total/ 15534
Daily Mirror	Positive/ 128	Negative/ 139	Trust/ 83	Anticipation/ 63	Fear/ 112	Anger/ 72	Joy/ 36	Sadness/ 69	Surprise/ 42	Disgust/ 34	Total/ 778
Sunday Times	Positive/ 760	Negative/ 571	Trust/ 459	Anticipation/ 314	Fear/ 372	Anger/ 293	Joy/ 202	Sadness/ 228	Surprise/ 170	Disgust/ 118	Total/ 3487
The Guardian	Positive/ 4333	Negative/ 3667	Trust/ 2887	Anticipation/ 1837	Fear/ 2549	Anger/ 1966	Joy/ 1031	Sadness/ 1372	Surprise/ 871	Disgust/ 823	Total/ 21336

Sunday Telegraph	Positive/ 367	Negative/ 309	Trust/ 236	Anticipation/ 144	Fear/ 239	Anger/ 177	Joy/ 86	Sadness/ 135	Surprise/ 84	Disgust/ 86	Total/ 1863
The Times	Positive/ 4472	Negative/ 3137	Trust/ 2867	Anticipation/ 1957	Fear/ 2124	Anger/ 1699	Joy/ 1119	Sadness/ 1241	Surprise/ 867	Disgust/ 665	Total/ 20148

The following figure shows the finding of the total sentiments towards the searched word 'Saudi Arabia' by the eight selected newspap



ers in Figure 3:

Figure 3 Total Sentiment Analysis frequency of each newspaper

6. Discussion of the Study

The results presented in Figure (3) indicated the level of the frequency of talking about Saudi Arabia on each newspaper as mentioned above. The overall level of sentiment analysis was positive according to the numbers on each table presented earlier; however, it is also true that there were instances where there was a level of negativity here and there. The

eight sentiments that were expressed and analyzed primarily on this study were fear, trust, anger, sadness, anticipation, surprise, joy and disgust.

7. Conclusion and Implication

The aims for this study were looking at and analyzing the sentiments of English language newspapers that addressed name of the country: 'Saudi Arabia'. The theoretical and methodological approaches deployed in analyzing the data were corpus linguistics (CL) and sentiment analysis approach as a way of analyzing newspapers media discourse and sentiment representations of Saudi Arabia. The main source of data collection and analysis was done through *sketch engine (SE).* The newspapers corpora were extracted from SE dataset between the period of 1993-2013. Finally, the findings after the analysis of the selected newspapers corpora have shown slightly more positive sentiment than negative throughout the use of the word *Saudi Arabia.* It is, however, true that there was a variation on the total amount of using the word *Saudi Arabia* from one newspaper to another. It is hoped that the study's findings have answered the research questions as well as could add a valuable contribution to the field in general.

Declaration of Competing Interest

No conflicts of interest

References

(1) Poole E, Sandford E. Reporting Islam: Media Representations of British Muslims. : Tauris; 2002.
(2) Baker P, Gabrielatos C, McEnery T. Discourse Analysis and Media Attitudes: The Representation of Islam in the British Press. : Cambridge University Press; 2013.
(3) Nacos BL, Torres-Reyna O. Muslim Americans in the News before and after 9-11. Department of Political Science. Columbia University 2002.
(4) Said E. 2003. Orientalism. 1978.
(5) Jerichow A. Islam in a Changing World. ; 2013.
(6) Said EW. Covering Islam: How the Media and the Experts Determine How We See the Rest of the World (London: Vintage). 1981.
(7) Schulder M, Wiegand M, Ruppenhofer J. Automatic generation of lexica for sentiment polarity shifters. Natural Language Engineering 2021 March;27(2):153-179.
(8) Morante R, Sporleder C. Proceedings of the Workshop on Negation and Speculation in Natural Language Processing. Proceedings of the Workshop on Negation and Speculation in Natural Language Processing 2010.

(9) Recognizing contextual polarity in phrase-level sentiment analysis. ; 2005.
(10) The BioScope corpus: Annotation for negation, uncertainty and their scope in biomedical texts. ; 2008.
(11) Descriptive analysis of negation cues in biomedical texts. ; 2010.
(12) +/-effectwordnet: Sense-level lexicon acquisition for opinion inference. ; 2014.
(13) Recursive deep models for semantic compositionality over a sentiment treebank. ; 2013.
(14) Schulder M, Wiegand M, Ruppenhofer J, Roth B. Towards Bootstrapping a Polarity Shifter Lexicon using Linguistic Features. Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers) 2017;1.
(15) Introducing a lexicon of verbal polarity shifters for english. ; 2019.
(16) Disambiguation of verbal shifters. ; 2019.
(17) The portrayal of Saudi Arabia in Australian newspapers. ; 2004.
(18) Investigating the Saudi Arabian Image in three Israeli Newspapers. ; 2004.
(19) Hunston S. Corpora and language teaching: Specific applications. Corpora in Applied Linguistics; 2012.
(20) Wodak R. The genesis of racist discourse in Austria since 1989. Texts and practices: Routledge; 2013. p. 115-136.
(21) Van Dijk TA. Elite discourse and racism. : Sage; 1993.
(22) Blommaert J. Discourse. Cambridge: Cambridge University Press; 2005.
(23) Fairclough N. Discourse and Social Change. Cambridge: Polity Press.; 1992.
(24) Wodak R, Meyer M. Methods of critical discourse studies. : Sage; 2015.
(25) Fairclough N, Mulderrig J, Wodak R. Critical discourse analysis. Discourse Studies: A Multidisciplinary Introduction; 2011.
(26) Kress G. Critical discourse analysis. Annual review of applied linguistics 1990;11:84-99.
(27) Pennycook A. Critical applied linguistics: A critical introduction. : Routledge; 2001.