*data*

MDPI

*Data Descriptor*

# Sentiment Analysis of Multilingual Dataset of Bahraini Dialects, Arabic, and English

**Thuraya Omran [1,*], Baraa Sharef [2], Crina Grosan [3] and Yongmin Li [1]**

[1] Department of Computer Science, Brunel University London, Uxbridge UB8 3PH, UK; yongmin.li@brunel.ac.uk
[2] Department of Information Technology, College of Information Technology, Ahlia University, Manama P.O. Box 10878, Bahrain; bsharif@ahlia.edu.bh
[3] Division of Applied Technologies for Clinical Care, King's College London, London WC2R 2LS, UK; crina.grosan@kcl.ac.uk
* Correspondence: thuraya.omran@brunel.ac.uk

**Abstract:** Sentiment analysis is an application of natural language processing (NLP) that requires a machine learning algorithm and a dataset. In some cases, the dataset availability is scarce, particularly with Arabic dialects, precisely the Bahraini ones, which necessitates using an approach such as translation, where a rich source language is exploited to create the target language dataset. In this study, a dataset of Amazon product reviews in Bahraini dialects is presented. This dataset was generated using two cascading stages of translation—a machine translation followed by a manual one. Machine translation was applied using Google Translate to translate English Amazon product reviews into Standard Arabic. In contrast, the manual approach was applied to translate the resulting Arabic reviews into Bahraini ones by qualified native speakers utilizing constructed customized forms. The resulting parallel dataset of English, Standard Arabic, and Bahraini dialects is called English_Modern Standard Arabic_Bahraini Dialects product reviews for sentiment analysis "E_MSA_BDs-PR-SA". The dataset is balanced, composed of 2500 positive and 2500 negative reviews. The sentiment analysis process was implemented using a stacked LSTM deep learning model. The Bahraini dialect product dataset can be utilized in the transfer learning process for sentimentally analyzing another dataset in Bahraini dialects.

**Dataset:** https://doi.org/10.17632/5rhw2srzjj.1

**Dataset License:** CC-BY-NC

**Keywords:** Bahraini dialects resources; Bahraini resources scarcity; deep learning; products reviews

check for updates

**Citation:** Omran, T.; Sharef, B.; Grosan, C.; Li, Y. Sentiment Analysis of Multilingual Dataset of Bahraini Dialects, Arabic, and English. *Data* **2023**, *8*, 68. https://doi.org/10.3390/data8040068

Academic Editor: Hiram Calvo

Received: 18 February 2023
Revised: 20 March 2023
Accepted: 24 March 2023
Published: 30 March 2023

## 1. Summary

Sentiment analysis (SA) is the computational study of people's opinions and impressions toward entities such as events, individuals, organizations, services, and products [1], whether these opinions are acceptable or not.

Sentiment analysis is an application of natural language processing (NLP) that requires a machine learning algorithm and a data source, either a lexicon or a dataset. The dataset is a key part of the sentiment analysis process, and the size of the dataset plays a significant role in obtaining the best analysis results.

Despite the growing body of literature that recognizes the importance of providing datasets for both standard and dialect Arabic NLP, the resource scarcity of Bahraini dialects has attracted very little attention from the scholarly community, leading to a notable lack of studies in such a field. The primary aim of this paper is to present a Bahraini dialect dataset and give a detailed description of creating it.

Reviewing the literature shows that most Arabic NLP datasets were created using social media platforms such as Twitter, Facebook (Meta), and Instagram. It was reported by [2] that Twitter is the most used platform in Gulf Council countries (GCC) compared to Arab world countries. For example, refs. [3–7] used Twitter to create their research datasets. It was also mentioned by [8] that researchers used Facebook comments to create their dataset. For example, refs. [9,10] used Facebook comments, while [11] generated their dataset from Instagram.

Here in this paper, a different approach was utilized to create our dataset of Bahraini dialects. This approach is translation, which can be employed in the case of resource scarcity [12]. The dataset presented here is a dataset that covers Amazon product reviews and is a balanced raw one, composed of 2500 positive and 2500 negative reviews. It was created by two cascading translation stages: machine and manual translation. Machine translation was applied to translate 5000 English Amazon product reviews to modern standard Arabic (MSA). In contrast, manual translation was employed to convert the resulting MSA reviews to Bahraini dialects by qualified native speakers utilizing constructed customized forms distributed electronically.

Creating a dataset is not an easy task. It needs time, collaborative efforts, money, and other resources. The dataset presented here will save all factors mentioned: time, money, and effort, not only for NLP community researchers but for all stakeholders. It can be used as a benchmark dataset for future NLP studies.

The presented dataset was generated to enrich the Arabic NLP community with a multilingual dataset by filling a gap in Arabic resources in general and Bahraini dialects in particular, which promotes research studies in such a field. Furthermore, it can be used in the transfer learning process as source data to sentimentally analyze another dataset that covers different domains in Bahraini dialects, as detailed in our research paper [12]. The BDs dataset is reliable because it was obtained from qualified respondents.

The remaining parts of this paper have been divided into two parts. The first part deals with data description, while the second covers the methods of creating the dataset and its pre-processing.

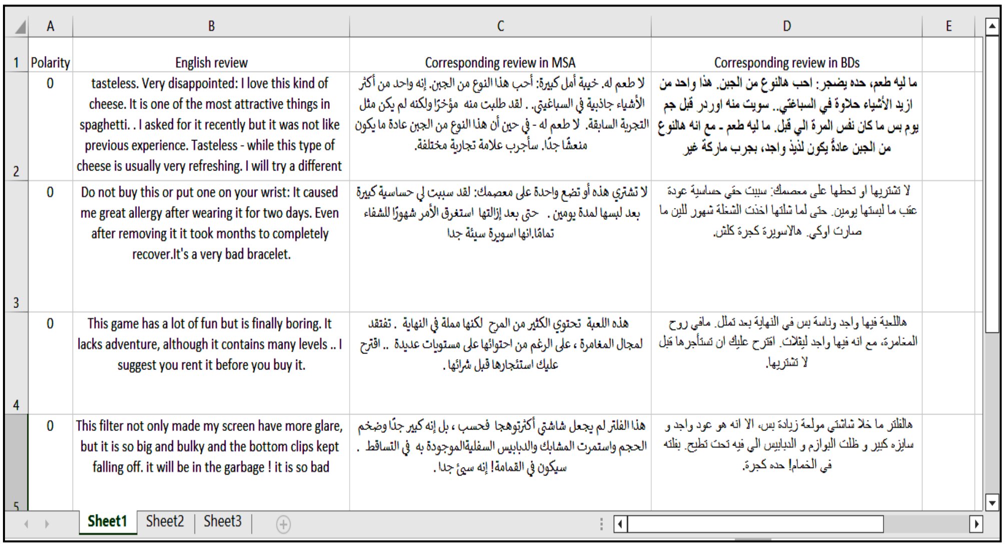
## 2. Data Description

This section gives a detailed description of the dataset repository, where the datasets and coding for the dataset preprocessing are available.

### 2.1. Dataset Repository

The created dataset was published in a Mendeley dataset repository. The repository contains two folders, “Datasets” and “Data Preprocessing”. The Dataset folder contains one to four datasets files in CSV format, one of which contains a parallel dataset of English, MSA, and Bahraini. In contrast, each of the remaining ones represents the dataset of each language individually. Two stopwords files are in txt format. The second folder will be described in Section 2.2. The CSV files of the Dataset folder are as follows:

1. E_MSA_BDs-PR-SA.csv: is composed of four columns. The first one is polarity, which is represented by 0 and 1. The second column is the “English review”, the third column is the “corresponding review in MSA”, and the fourth one is the “corresponding review in BDs”, as shown in Figure 1;
2. Bahraini Dialects Dataset.csv: is composed of two columns: the polarity and the review text in Bahraini dialects, and 5000 rows of reviews;
3. English Dataset.csv: is composed of two columns: the polarity and the review text in English, and 5000 rows of reviews;
4. MSA Dataset.csv: is composed of two columns: the polarity and the review text in modern standard Arabic (MSA), and 5000 rows of reviews.



|  | A | B | C | D | E |
|---|---|---|---|---|---|
| 1 | Polarity | English review | Corresponding review in MSA | Corresponding review in BDs |  |
| 2 | 0 | tasteless. Very disappointed: I love this kind of cheese. It is one of the most attractive things in spaghetti. . I asked for it recently but it was not like previous experience. Tasteless - while this type of cheese is usually very refreshing. I will try a different | لا طعم له. خيبة أمل كبيرة: أحب هذا النوع من الجبن. إنه واحد من أكثر الأشياء جاذبية في السباغيتي. . لقد طلبت منه مؤخرًا ولكنه لم يكن مثل التجربة السابقة. لا طعم له - في حين أن هذا النوع من الجبن عادة ما يكون منعشًا جدًا. سأجرب علامة تجارية مختلفة. | ما ليه طعم، حده يضجر: احب هالنوع من الجبن. هذا واحد من ازيد الأشياء حلاوة في السباغتي.. سويت منه اوردر قبل جم يوم بس ما كان نفس المرة الي قبل. ما ليه طعم ـ مع انه هالنوع من الجبن عادةً يكون لذيذ واجد، بجرب ماركة غير |  |
| 3 | 0 | Do not buy this or put one on your wrist: It caused me great allergy after wearing it for two days. Even after removing it it took months to completely recover.It's a very bad bracelet. | لا تشتري هذه أو تضع واحدة على معصمك: لقد سببت لي حساسية كبيرة بعد لبسها لمدة يومين . حتى بعد إزالتها استغرق الأمر شهورًا للشفاء تمامًا.انها اسويرة سيئة جدا | لا تشتريها او تحطها على معصمك: سببت حقي حساسية عودة عقب ما لبستها يومين. حتى لما شلتها اخذت الشغلة شهور للين ما صارت اوكي. هالاسويرة كجرة كلش. |  |
| 4 | 0 | This game has a lot of fun but is finally boring. It lacks adventure, although it contains many levels .. I suggest you rent it before you buy it. | هذه اللعبة تحتوي الكثير من المرح لكنها مملة في النهاية . تفتقد لمجال المغامرة ، على الرغم من احتوائها على مستويات عديدة .. اقترح عليك استئجارها قبل شرائها . | هاللعبة فيها واجد وناسة بس في النهاية بعد تملل. مافي روح المغامرة، مع انه فيها واجد ليفلات. اقترح عليك ان تستأجرها قبل لا تشتريها. |  |
| 5 | 0 | This filter not only made my screen have more glare, but it is so big and bulky and the bottom clips kept falling off. it will be in the garbage ! it is so bad | هذا الفلتر لم يجعل شاشتي أكثرتوهجا فحسب ، بل إنه كبير جدًا وضخم الحجم واستمرت المشابك والدبابيس السفليةالموجودة به في التساقط . سيكون في القمامة! إنه سيئ جدا . | هالفلتر ما خلا شاشتي مولعة زيادة بس، الا انه هو عود واجد و سايزه كبير و ظلت البوازم و الدبابيس الي فيه تحت تطيح. بفلته في الخمام! حده كجرة. |  |

**Figure 1.** The CSV file contains four columns of the parallel dataset [12].

The following are the two txt files:

5 BDs-StopWords.txt: contains a created list of Bahraini dialects stopwords. The stopwords are parts of the text that are not useful in sentiment analysis, such as punctuation, pronouns, and prepositions;

6 MSA-Stopwords.txt: MSA-Stopwords.txt: contains a list of MSA stopwords. The included words in this list represent an extension to the built one in python software.

### 2.2. Dataset Pre-Processing

The associated code for the dataset preprocessing is in the "Data Preprocessing" folder, which contains three python files; each file is dedicated to preprocessing the reviews of a specific dataset. One of the preprocessing steps is the augmentation to obtain artificial reviews from the original ones. Every file contains coding for augmenting the original 5000 reviews of each dataset to 10,000 by applying the swap augmentation technique twice. The first swap augmentation was applied using 1 and 10 as the Min and Max parameter values for the number of words to be augmented, while the second swap was applied using 1 and 5 as the Min and Max, respectively. The program's output informs the user of the number of reviews before the augmentation process and the number of reviews after it.

Additionally, it gives the percentage of the number of original and artificial reviews of the total ones. After applying the augmentation steps, a new list of the original and artificial reviews will be generated. The other preprocessing steps follow the augmentation step. The three Python files are as follows:

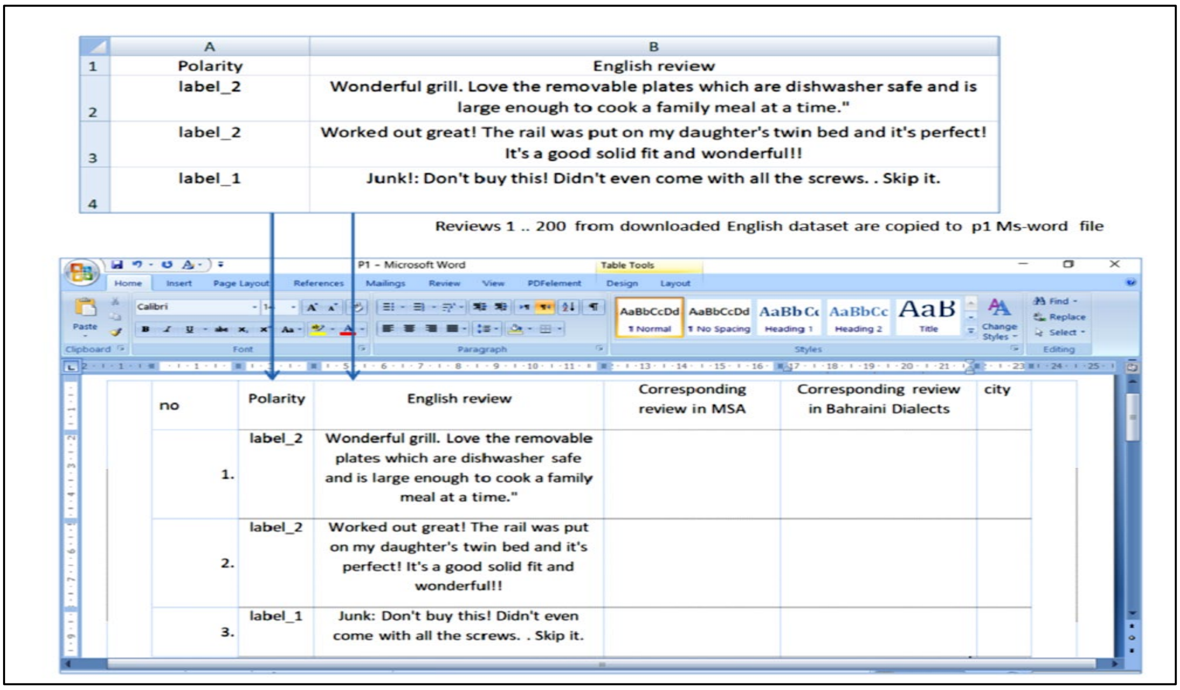
7. BDs_ PreProcessing.py: contains the coding of preprocessing of Bahraini dialects product reviews, including the augmenting of the 5000 reviews of the Bahraini dialect product dataset to 10,000;
8. MSA_ PreProcessing.py: contains the coding for preprocessing MSA product reviews, including the augmenting of the 5000 reviews of the standard Arabic product dataset to 10,000;
9. English_ PreProcessing.py: contains the coding for preprocessing English product reviews, including augmenting of the 5000 reviews of the English product dataset to 10,000.

## 3. Methods

This section described the detailed steps of our dataset design and preparation, and then it moved on to shed light on the general preprocessing steps of the dataset in Section 3.2, where the detailed preprocessing steps for modern standard Arabic and Bahraini dialects were covered in Section 3.2.1 and those of the English dataset in Section 3.2.2.

## *3.1. Dataset Design and Preparation*

1. A dataset in English of Amazon product reviews for sentiment analysis was downloaded from [13]. The downloaded English Dataset is composed of two comma-separated values (CSV) training and testing files. They contain negative and positive reviews labeled as label_1 and label_2, respectively;
2. A total of 5000 reviews, 2500 negative and 2500 positive, have been manually selected from the training and testing files of the English dataset of Amazon reviews to form a balanced dataset;
3. The selected 5000 reviews were copied into 25 separate Ms-word files. The copied reviews have been validated manually by correcting the spelling and grammar errors, using a Ms-word grammar and spelling checker supported by a human checker specializing in English–Arabic–English translation, and by deleting the sentences that are not affecting the review sentiment. This deletion was aimed to provide reviews with a reasonable length, not exceeding 200 words, which aid the capturing of meaning by participants for a manual translation of MSA reviews to BDs. Each of the 25 files includes a template for a table composed of six columns: number, label, review in English, corresponding review in MSA, corresponding review in BDs, and a city, as shown in Figure 2. The Ms-word files have been saved as P1, P2, and P3 . . . P25. Each *Pi* contains 200 balanced reviews. This numbering method helps in identifying each review in the stage of distributing the MSA reviews and collecting their corresponding ones in BDs, as will be explained later;



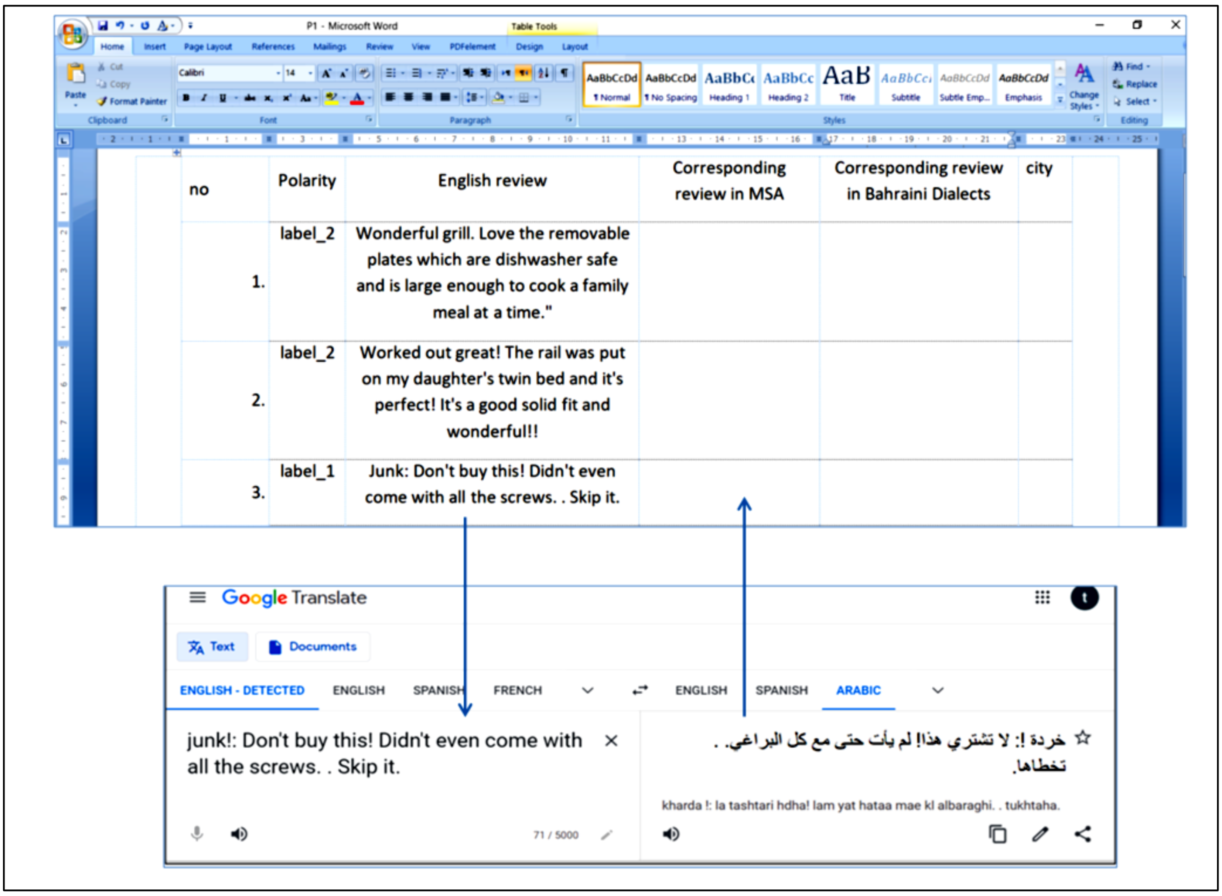
| | A | B |
|---|---|---|
| 1 | Polarity | English review |
| 2 | label_2 | Wonderful grill. Love the removable plates which are dishwasher safe and is large enough to cook a family meal at a time." |
| 3 | label_2 | Worked out great! The rail was put on my daughter's twin bed and it's perfect! It's a good solid fit and wonderful!! |
| 4 | label_1 | Junk!: Don't buy this! Didn't even come with all the screws. . Skip it. |

Reviews 1 .. 200 from downloaded English dataset are copied to p1 Ms-word file

| no | Polarity | English review | Corresponding review in MSA | Corresponding review in Bahraini Dialects | city |
|---|---|---|---|---|---|
| 1. | label_2 | Wonderful grill. Love the removable plates which are dishwasher safe and is large enough to cook a family meal at a time." | | | |
| 2. | label_2 | Worked out great! The rail was put on my daughter's twin bed and it's perfect! It's a good solid fit and wonderful!! | | | |
| 3. | label_1 | Junk: Don't buy this! Didn't even come with all the screws. . Skip it. | | | |

**Figure 2.** Example of a table template in an Ms-word file that contains the copied reviews.

4. The 200 reviews of *Pi* were translated one by one to MSA using https://translate.google.com/ (accessed on 1 December 2019). The MSA-translated review is copied to its dedicated cell in the table in the MS-word file, as shown in Figure 3. The resulting MSA reviews were manually validated by someone who specializes in English–Arabic–English translation to ensure meaning matching between the English reviews and their corresponding ones in MSA;



| no | Polarity | English review | Corresponding review in MSA | Corresponding review in Bahraini Dialects | city |
|---|---|---|---|---|---|
| 1. | label_2 | Wonderful grill. Love the removable plates which are dishwasher safe and is large enough to cook a family meal at a time." | | | |
| 2. | label_2 | Worked out great! The rail was put on my daughter's twin bed and it's perfect! It's a good solid fit and wonderful!! | | | |
| 3. | label_1 | Junk: Don't buy this! Didn't even come with all the screws. . Skip it. | | | |

junk!: Don't buy this! Didn't even come with all the screws. . Skip it.

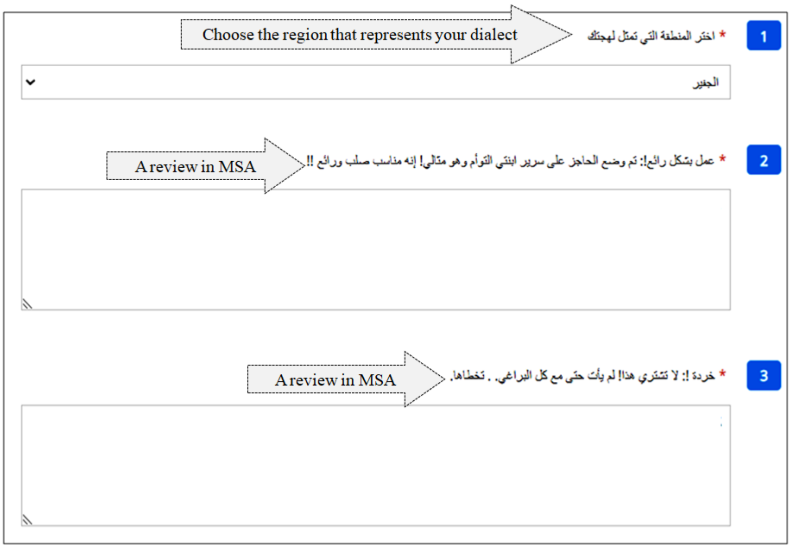
71 / 5000

خردة !: لا تشتري هذا! لم يأت حتى مع كل البراغي. .
تخطاها.

kharda !: la tashtari hdha! lam yat hataa mae kl albaraghi. . tukhtaha.

**Figure 3.** Example of translating an English review to MSA.

5. The 5000 MSA reviews that resulted from the translation process were distributed to Bahraini dialect native speakers using 500 customized forms that were constructed through https://getfoureyes.com/ (accessed on 5 December 2019). Each form includes 10 different and unique MSA reviews. Below each review, a text box was provided where the respondent can rewrite the review in their spoken dialect, as depicted in Figure 4.



1 * اختر المنطقة التي تمثل لهجتك

Choose the region that represents your dialect

الجفير

2 * عمل بشكل رائع!: تم وضع الحاجز على سرير ابنتي التوأم وهو مثالي! إنه مناسب صلب ورائع !!

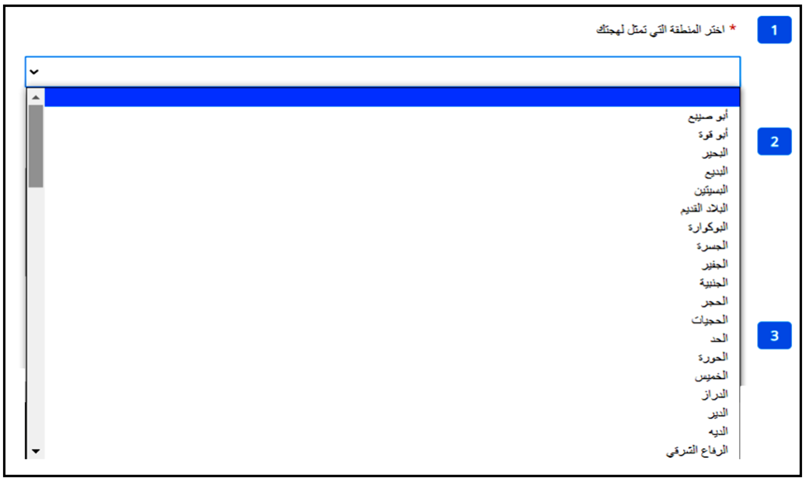
A review in MSA

3 * خردة !: لا تشتري هذا! لم يأت حتى مع كل البراغي. . تخطاها.

A review in MSA

**Figure 4.** Example of MSA reviews and the provided text box.

In addition, there is a combo box containing a list of Bahraini towns and villages, from which the respondent selects one of them that represents their dialect, as shown in Figure 5.



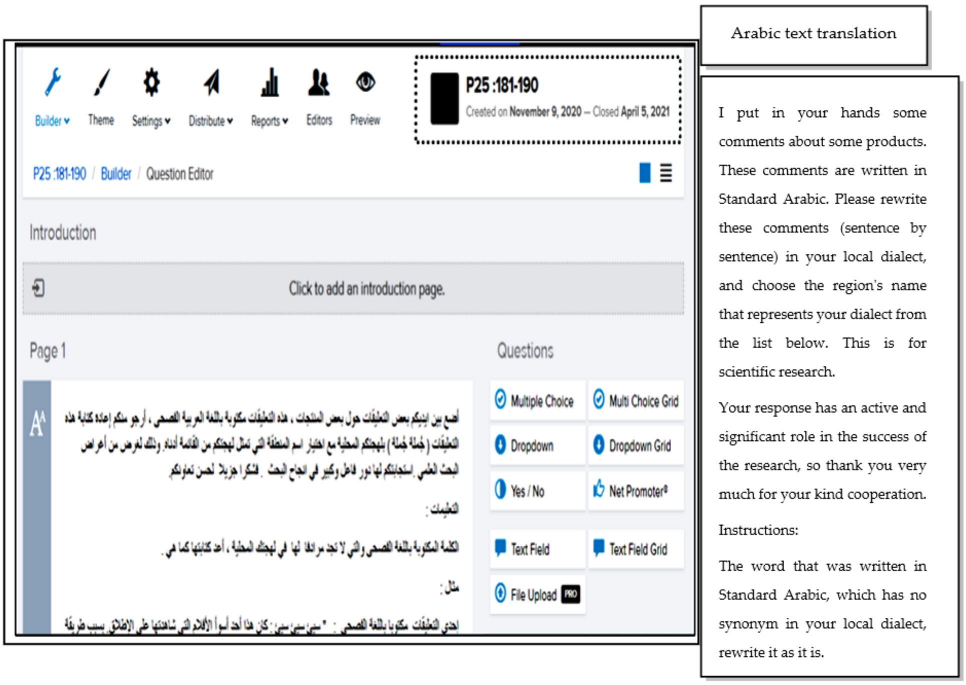
* اختر المنطقة التي تمثل لهجتك

أبو صيبع
أبو قوة
البحير
البديع
البسيتين
البلاد القديم
البوكوارة
الجسرة
الجفير
الجنبية
الحجر
الحجيات
الحد
الحورة
الخميس
الدراز
الدير
الديه
الرفاع الشرقي

**Figure 5.** A combo box shows some Bahraini towns and villages.

At the beginning of the form, an instruction was provided for the respondents requesting them to rewrite the provided MSA reviews in their dialects, sentence by sentence, and rewrite the word as it is, if there is no synonym for it in BDs.

Each form was identified according to the Ms-word file it affiliates with, followed by the range of reviews it contains, as shown with a dotted rectangle in Figure 6. For example, the form identified by (P25:181-190) represents a form that belongs to the Ms-word file named P25 and contains reviews from 181-190, keeping in mind that each file contains 200 reviews, which means creating 20 forms per Ms-word file. So, when the respondent submits the form, it is easy for the data collector to locate it and collect the responses.



P25 :181-190
Created on November 9, 2020 – Closed April 5, 2021

P25 :181-190 / Builder / Question Editor

Introduction

Click to add an introduction page.

Page 1

أضع بين إيديكم بعض التعليقات حول بعض المنتجات ، هذه التعليقات مكتوبة باللغة العربية الفصحى ، أرجو منكم إعادة كتابة هذه
التعليقات (جملة جملة) بلهجتكم المحلية مع اختيار اسم المنطقة التي تمثل لهجتكم من القائمة أدناه. وذلك لغرض من أغراض
البحث العلمي. إستجابتكم لها دور فاعل وكبير في انجاح البحث . فشكرا جزيلا لحسن تعاونكم

التعليمات :

الكلمة المكتوبة باللغة الفصحى والتي لا تجد مرادفا لها في لهجتك المحلية ، أعد كتابتها كما هي .

مثال :

إحدى التعليقات مكتوبا باللغة الفصحى : "سي سي سي: كان هذا أحد أسوأ الأفلام التي شاهدتها على الإطلاق بسبب طريقة

Arabic text translation

I put in your hands some comments about some products. These comments are written in Standard Arabic. Please rewrite these comments (sentence by sentence) in your local dialect, and choose the region's name that represents your dialect from the list below. This is for scientific research.

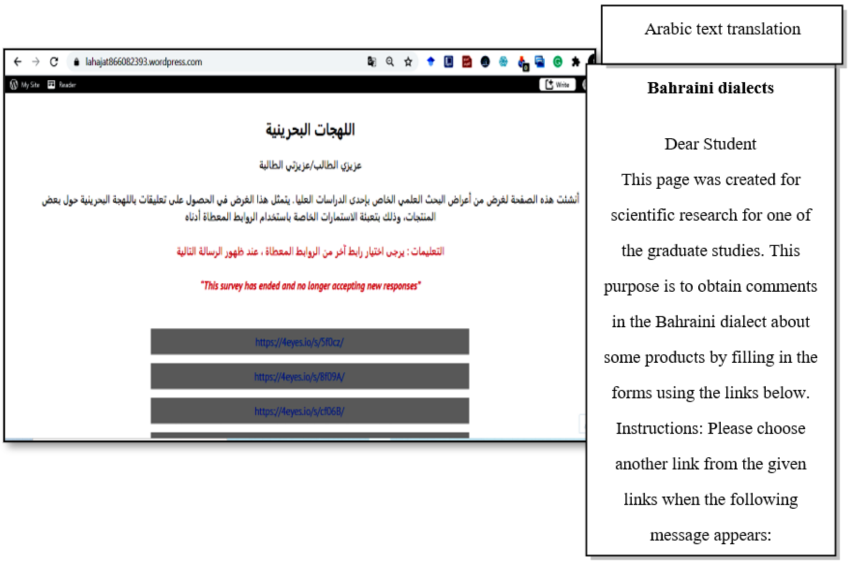
Your response has an active and significant role in the success of the research, so thank you very much for your kind cooperation.

Instructions:

The word that was written in Standard Arabic, which has no synonym in your local dialect, rewrite it as it is.

**Figure 6.** The identification of the distributed form.

Each form was automatically assigned a link during its setup by https://getfoureyes.com/ (accessed on 5 December 2019). Some forms' links were listed on a created web page: https://lahajat866082393.wordpress.com/ (accessed on 20 April 2021). The URL of the created web page was distributed to the respondents via WhatsApp application and text message. When the respondent visits the URL page, they will find the list of links as shown in Figure 7, where they were asked to choose a link. If a message such as "This survey has ended and no longer accepting new responses" is popped out, they should choose another form link. Such a procedure was followed to guarantee that all forms had been responded to and there were no duplicate responses.



lahajat866082393.wordpress.com

اللهجات البحرينية

عزيزي الطالب/عزيزتي الطالبة

أنشئت هذه الصفحة لغرض من أعراض البحث العلمي الخاص بإحدى الدراسات العليا. يتمثل هذا الغرض في الحصول على تعليقات باللهجة البحرينية حول بعض المنتجات، وذلك بتعبئة الاستمارات الخاصة باستخدام الروابط المعطاة أدناه

التعليمات : يرجى اختيار رابط آخر من الروابط المعطاة ، عند ظهور الرسالة التالية

"This survey has ended and no longer accepting new responses"

https://4eyes.io/s/5f0cz/

https://4eyes.io/s/8f09A/

https://4eyes.io/s/cf06B/

Arabic text translation

**Bahraini dialects**

Dear Student

This page was created for scientific research for one of the graduate studies. This purpose is to obtain comments in the Bahraini dialect about some products by filling in the forms using the links below.

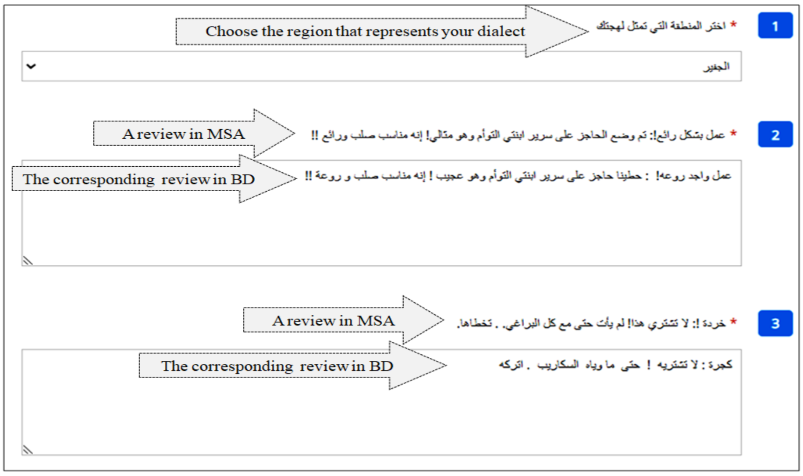
Instructions: Please choose another link from the given links when the following message appears:

**Figure 7.** A snippet for the web page.

All respondents are holders of academic qualifications, such as diplomas, bachelors, masters, and doctorates, in various disciplines.

Several criteria were considered when selecting the participants, such as filling out the forms on time, being aware of the research's importance, and adequately responding as instructions were received. The overall response was very positive. The respondents were thanked for their time and efforts;

6 The submitted forms were opened one by one by the data collector, a native speaker specializing in Arabic, to ensure semantic and spelling corrections. Each submitted form contains 10 MSA reviews and their corresponding ones in BDs. Figure 8 shows two out of ten reviews;



1 * اختر المنطقة التي تمثل لهجتك

Choose the region that represents your dialect

الجفير

2 * عمل بشكل رائع!: تم وضع الحاجز على سرير ابنتي التوأم وهو مثالي! إنه مناسب صلب ورائع !!

A review in MSA

عمل واجد روعه! : حطينا حاجز على سرير ابنتي التوأم وهو عجيب ! إنه مناسب صلب و روعة !!

The corresponding review in BD

3 * خردة !: لا تشتري هذا! لم يأت حتى مع كل البراغي. . تخطاها.

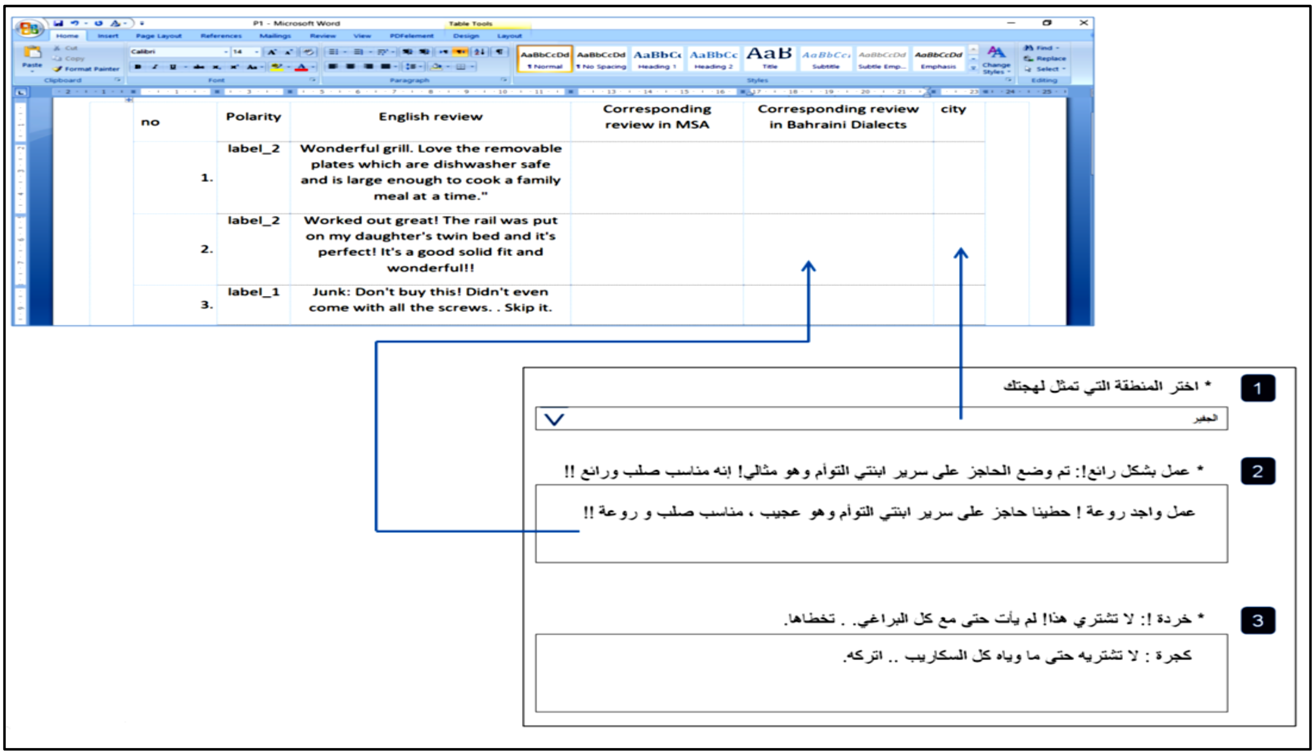
A review in MSA

كجرة : لا تشتريه ! حتى ما وياه السكاريب . اتركه

The corresponding review in BD

**Figure 8.** An example of MSA reviews and the corresponding ones in BDs [12].

7 After the checking, revising, and validation processes, the responses were collected by copying them to the dedicated cell in the table of the corresponding file, as shown in Figure 9.



| no | Polarity | English review | Corresponding review in MSA | Corresponding review in Bahraini Dialects | city |
|---|---|---|---|---|---|
| 1. | label_2 | Wonderful grill. Love the removable plates which are dishwasher safe and is large enough to cook a family meal at a time." | | | |
| 2. | label_2 | Worked out great! The rail was put on my daughter's twin bed and it's perfect! It's a good solid fit and wonderful!! | | | |
| 3. | label_1 | Junk: Don't buy this! Didn't even come with all the screws. . Skip it. | | | |

1 * اختر المنطقة التي تمثل لهجتك

الجفير

2 * عمل بشكل رائع!: تم وضع الحاجز على سرير ابنتي التوأم وهو مثالي! إنه مناسب صلب ورائع !!

عمل واجد روعة ! حطينا حاجز على سرير ابنتي التوأم وهو عجيب ، مناسب صلب و روعة !!

3 * خردة !: لا تشتري هذا! لم يأت حتى مع كل البراغي. . تخطاها.

كجرة : لا تشتريه حتى ما وياه كل السكاريب .. اتركه.

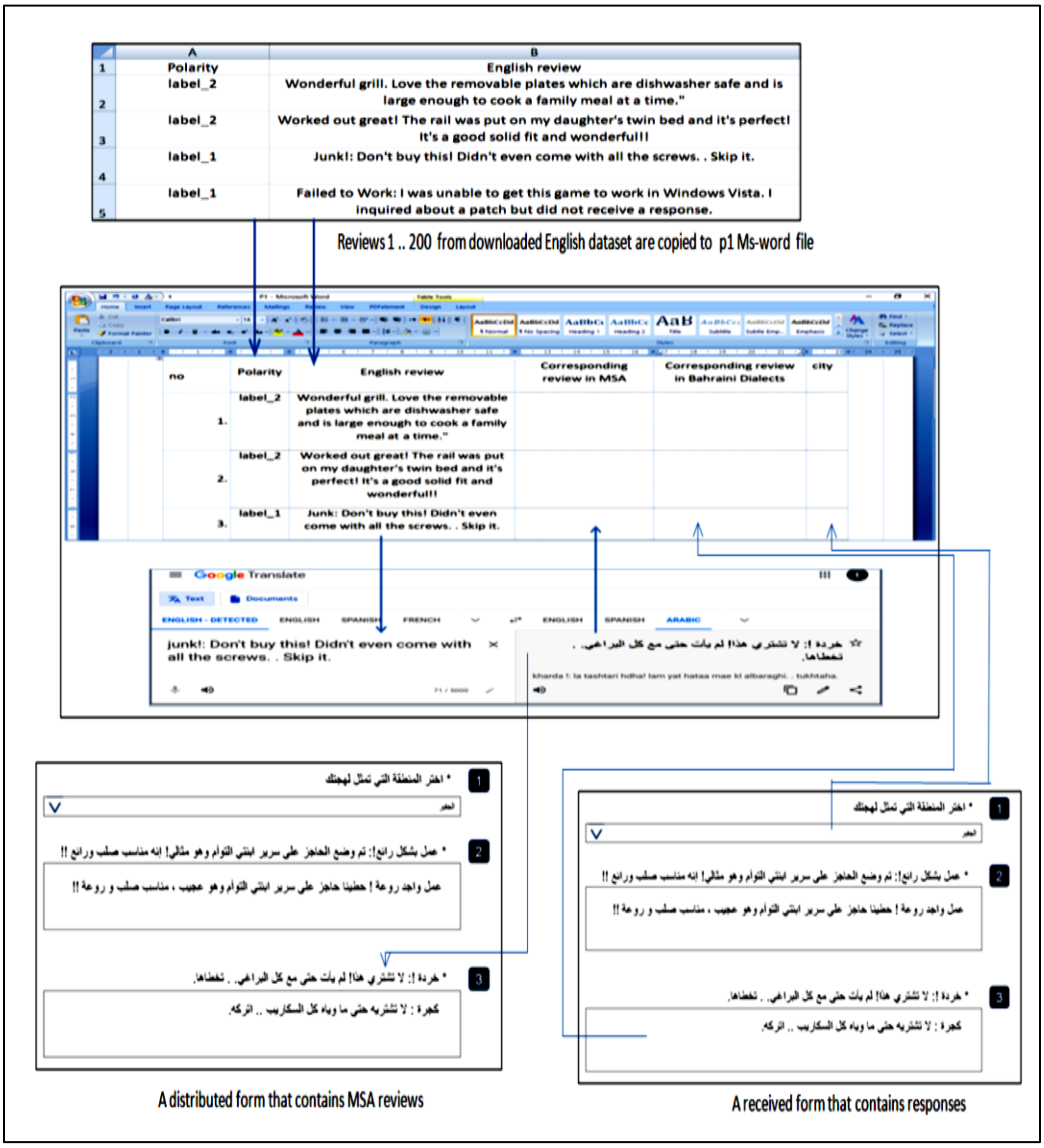
**Figure 9.** Copying the responses to the corresponding Ms-word file.

The collected responses cover most Bahraini cities and villages, such as Manama, Isa Town, Al Nuwaidart, Sitra, East Riffa, and many others, with different numbers of responses, some of which are low and others high. It is worth mentioning that the main objective is to obtain the reviews in Bahraini dialects for their corresponding ones in MSA, regardless of which city they represent, especially when keeping in mind that the sentiment classification process included in the reviews was based on the polarity as positive or negative, not on the city or village. Therefore, there is no big issue with the low number of collected dialects because the high number of responses from native speakers of some villages will compensate for the low ones of others due to the similarity in sentiment vocabulary between the villages and cities. In other words, the obligatory equal number of responses from each village or city is not a condition. A list of covered cities and villages is provided in Appendix A. The complete steps of preparing the datasets are depicted in Figure 10.

8 The resulting parallel dataset of English, MSA, and BDs of 5000 reviews was copied to the Ms-Excel file, where the polarity labels (label_1, label_2) were replaced by (0, 1), respectively. After the step of copying reviews and replacing the labels, the datasets have been put through one of the accuracy checks by checking the missing and null values of the polarity and the review text in English and its corresponding ones in MSA and BDs, using the count function that guaranteed the required numbers of rows are not empty;

9 The Ms-Excel file was converted to a text file with UTF-8 encoding, a suitable format for processing Arabic text in Python;

10 Each dataset was separated individually in a step to prepare them for the SA process.



| | A | B |
|---|---|---|
| 1 | Polarity | English review |
| 2 | label_2 | Wonderful grill. Love the removable plates which are dishwasher safe and is large enough to cook a family meal at a time." |
| 3 | label_2 | Worked out great! The rail was put on my daughter's twin bed and it's perfect! It's a good solid fit and wonderful!! |
| 4 | label_1 | Junk!: Don't buy this! Didn't even come with all the screws. . Skip it. |
| 5 | label_1 | Failed to Work: I was unable to get this game to work in Windows Vista. I inquired about a patch but did not receive a response. |

Reviews 1 .. 200 from downloaded English dataset are copied to p1 Ms-word file

| no | Polarity | English review | Corresponding review in MSA | Corresponding review in Bahraini Dialects | city |
|---|---|---|---|---|---|
| 1. | label_2 | Wonderful grill. Love the removable plates which are dishwasher safe and is large enough to cook a family meal at a time." | | | |
| 2. | label_2 | Worked out great! The rail was put on my daughter's twin bed and it's perfect! It's a good solid fit and wonderful!! | | | |
| 3. | label_1 | Junk: Don't buy this! Didn't even come with all the screws. . Skip it. | | | |

junk!: Don't buy this! Didn't even come with all the screws. . Skip it.

خردة !: لا تشتري هذا! لم يأت حتى مع كل البراغي. . تخطاها.

kharda !: la tashtari hdha! lam yat hataa mae kl albaraghi. . tukhtaha.

1 * اختر المنطقة التي تمثل لهجتك

2 * عمل بشكل رائع!: تم وضع الحاجز على سرير ابنتي التوأم وهو مثالي! إنه مناسب صلب ورائع !!

عمل واجد روعة ! حطينا حاجز على سرير ابنتي التوأم وهو عجيب ، مناسب صلب و روعة !!

3 * خردة !: لا تشتري هذا! لم يأت حتى مع كل البراغي. . تخطاها.

كجرة : لا تشتريه حتى ما وياه كل السكاريب .. اتركه.

A distributed form that contains MSA reviews

1 * اختر المنطقة التي تمثل لهجتك

2 * عمل بشكل رائع!: تم وضع الحاجز على سرير ابنتي التوأم وهو مثالي! إنه مناسب صلب ورائع !!

عمل واجد روعة ! حطينا حاجز على سرير ابنتي التوأم وهو عجيب ، مناسب صلب و روعة !!

3 * خردة !: لا تشتري هذا! لم يأت حتى مع كل البراغي. . تخطاها.

كجرة : لا تشتريه حتى ما وياه كل السكاريب .. اتركه.

A received form that contains responses

**Figure 10.** Steps of preparing the dataset.

## 3.2. Dataset Preprocessing

To ensure obtaining an accurate dataset, two steps of preprocessing were taken to reduce the noise from the reviews of all datasets: manual and automatic preprocessing. The manual preprocessing was conducted during the translation stage, as in steps 3, 4, and 6. In contrast, the automatic preprocessing steps were carried out using Python regular expression libraries, as detailed in Sections 3.2.1 and 3.2.2.

An augmentation technique was one of the steps in the dataset preprocessing. Data augmentation (DA) is one of the techniques that deals with the lack of data. It contributes to boosting data [14] and model performance, especially with deep learning algorithms such as long short-term memory (LSTM), which are greedy for data.

Four powerful methods of DA consist of easy data augmentation (EDA) techniques [15] such as synonym replacement, random swap, random deletion, and random insertion. The Python NLP augmentation library provides various textual augmenters at the sentence, word, and character levels [16]. The random swap method was explicitly selected due to the restrictions of the BDs dataset.

Figure 11 shows the steps for obtaining 10,000 reviews from the original 5000 reviews by applying the random swap method twice.

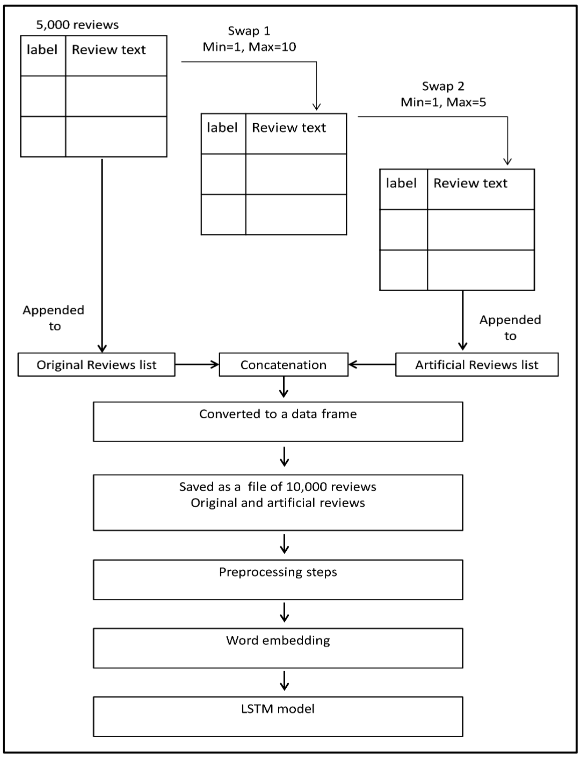


**Figure 11.** Augmentation steps for obtaining 10,000 reviews of 5000 ones.

The augmentation experiments were run on Intel(R) Core (TM) i7-6500U CPU @2.50GH, with a 64-bit operating system ×64-based processor.

### 3.2.1. Modern Standard Arabic and Bahraini Dialect Datasets Preprocessing

The preprocessing steps of the MSA and Bahraini dialect datasets were as follows [12]:

1. Applying the "swap" augmentation technique twice on each review of the dataset. The first one with a minimum number of words to be swapped equals 1 and a maximum equals 10; in contrast, the second time applying the augmentation technique, a minimum number of words to be swapped equals 1 and a maximum equals 5;
2. Normalizing some characters by replacing them with other ones; for example, the character "گ, whose spelling is "Ga" is replaced by "ك", "ة" is replaced by "ه", "ؤ" is replaced by "ء", and "ئ" is replaced by "ء";
3. Removing the digits, repeated characters, punctuation, diacritics, and English words;
4. Tokenizing the text by breaking it into chunks of words;
5. Removing the stop words.

### 3.2.2. English Dataset Preprocessing

The preprocessing steps for the English dataset are as follows [12]:

1. Applying the "swap" augmentation technique twice on each review of the dataset. The first one with a minimum number of words to be swapped equals 1 and a maximum equals 10; in contrast, the second time applying the augmentation technique, a minimum number of words to be swapped equals 1 and a maximum equals 5;
2. Normalizing text by converting uppercase characters to lower ones;
3. Removing special characters such as "@", hash, and digits;
4. Removing the stop words.

## 4. Remarks for the User

1. The dataset has been created to be analyzed at the document level;

2. The dataset has been augmented using a random swap technique through the Python program;
3. The dataset has been analyzed sentimentally using the LSTM model using different evaluation metrics such as accuracy, F1-score, and ROC AUC at different k-fold cross-validation values, in addition to the train-validate-test split. More details are in [12];
4. The results showed slight differences in the LSTM model's performance on all datasets. For example, the AUC value was 98.79% on the English dataset, 98.67% on the MSA, and 98.46% on the BDs;
5. To obtain a more enhanced performance of the LSTM model, the model has been integrated into the ensemble learning technique as detailed in [17];
6. The BDs dataset was utilized as a source in the transfer learning process to analyze a target dataset of movie comments in Bahraini dialects, as detailed in [12].

**Author Contributions:** T.O.: Conceptualization, Investigation, Methodology, Software, Writing—original draft, Writing—Review and Editing, Project administration, B.S.: Writing—Review and Editing, Supervision, Project administration, C.G.: Writing—Review and Editing, Supervision, Project administration, Y.L.: Writing—Review and Editing, Supervision, Project administration. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The dataset is openly available at: https://data.mendeley.com/datasets/5rhw2srzjj (accessed on 15 February 2023).

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

## Appendix A

**Table A1.** List of the towns and villages [12].

| Number | Name of Town/Villagein in Arabic | Name of Town/Villagein in English | Number of Responses (Filled Forms) |
|---|---|---|---|
| 1 | أبو قوة | AbuQuwah | 3 |
| 2 | البديع | Al Budaiya | 2 |
| 3 | البلاد القديم | Al Bilad Al Qadeem | 3 |
| 4 | الجفير | Al Juffair | 4 |
| 5 | الحد | Al Hidd | 4 |
| 6 | الدراز | Al Diraz | 13 |
| 7 | الدير | Al Dair | 9 |
| 8 | الديه | Al Daih | 4 |
| 9 | الرفاع الشرقي | East Rifaa | 18 |
| 10 | السنابس | Al Sanabis | 8 |
| 11 | السهلة الشمالية | North Sehla | 2 |
| 12 | العكر | Al Eker | 1 |
| 13 | القرية | Al Qurrayah | 1 |
| 14 | الكورة | Al Kawarah | 6 |
| 15 | المالكية | Al Malikiyah | 6 |
| 16 | المحرق | Al Muharraq | 8 |

**Table A1.** *Cont.*

| Number | Name of Town/Villagein in Arabic | Name of Town/Villagein in English | Number of Responses (Filled Forms) |
|---|---|---|---|
| 17 | المصلى | Al Musalla | 2 |
| 18 | المعامير | Al Ma'ameer | 19 |
| 19 | المنامة | Al Manama | 27 |
| 20 | النعيم | Al Naim | 3 |
| 21 | النويدرات | Al Nuwaidrat | 264 |
| 22 | أم الحصم | Umm AlHassam | 4 |
| 23 | بوري | Buri | 5 |
| 24 | توبلي | Tubli | 4 |
| 25 | جبلة حبشي | Jeblat Hebshi | 1 |
| 26 | جدحفص | Jidhafs | 3 |
| 27 | جدعلي | JidAli | 5 |
| 28 | دمستان | Damistan | 1 |
| 29 | رأس رمان | Ras Romman | 10 |
| 30 | سار | Saar | 4 |
| 31 | سترة | Sitra | 21 |
| 32 | سماهيج | Samaheej | 4 |
| 33 | سند | Sanad | 4 |
| 34 | عراد | Arad | 3 |
| 35 | كرانة | Karranah | 10 |
| 36 | كرزكان | Karzakkan | 5 |
| 37 | مدينة عيسى | Isa Town | 7 |
| 38 | مقابة | Maqabah | 2 |

## References

1. Zhang, L.; Wang, S.; Liu, B. Deep learning for sentiment analysis: A survey. In *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*; John Wiley & Sons Inc.: Hoboken, NJ, USA, 2018; Volume 8, p. 1253.
2. El-Masri, M.; Altrabsheh, N.; Mansour, H.; Ramsay, A. A web-based tool for Arabic sentiment analysis. *Procedia Comput. Sci.* **2017**, *117*, 38–45. [CrossRef]
3. Abdul-Mageed, M.; Alhuzali, H.; Elaraby, M. You tweet what you speak: A city-level dataset of Arabic dialects. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan, 7–12 May 2018.
4. Abo, M.E.M.; Shah, N.A.K.; Balakrishnan, V.; Kamal, M.; Abdelaziz, A.; Haruna, K. SSA-SDA: Subjectivity and Sentiment Analysis of Sudanese Dialect Arabic. In Proceedings of the 2019 International Conference on Computer and Information Sciences (ICCIS), Aljouf, Saudi Arabia, 3–4 April 2019; pp. 1–5.
5. Mohammed, A.; Kora, R. Deep learning approaches for Arabic sentiment analysis. *Soc. Netw. Anal. Min.* **2019**, *9*, 1–12. [CrossRef]
6. Alahmary, R.M.; Al-Dossari, H.Z.; Emam, A.Z. Sentiment analysis of Saudi dialect using deep learning techniques. In Proceedings of the 2019 International Conference on Electronics, Information, and Communication (ICEIC), Auckland, New Zealand, 22–25 January 2019; pp. 1–6.
7. Alsarsour, I.; Mohamed, E.; Suwaileh, R.; Elsayed, T. Dart: A large dataset of dialectal arabic tweets. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan, 7–12 May 2018.
8. Al Shamsi, A.A.; Abdallah, S. A Systematic Review for Sentiment Analysis of Arabic Dialect Texts Researches. In *Proceedings of International Conference on Emerging Technologies and Intelligent Systems: ICETIS 2021*; Springer International Publishing: Berlin/Heidelberg, Germany, 2022; Volume 2, pp. 291–309.
9. Mdhaffar, S.; Bougares, F.; Esteve, Y.; Hadrich-Belguith, L. Sentiment analysis of tunisian dialects: Linguistic ressources and experiments. In Proceedings of the Third Arabic Natural Language Processing Workshop (WANLP), Valencia, Spain, 3 April 2017; pp. 55–61.
10. Itani, M.; Roast, C.; Al-Khayatt, S. Developing resources for sentiment analysis of informal Arabic text in social media. *Procedia Comput. Sci.* **2017**, *117*, 129–136. [CrossRef]
11. Al Shamsi, A.; Abdallah, S. Sentiment Analysis of Emirati Dialect. *Big Data Cogn. Comput.* **2022**, *6*, 57. [CrossRef]

12. Omran, T.M.; Sharef, B.T.; Grosan, C.; Li, Y. Transfer learning and sentiment analysis of Bahraini dialects sequential text data using multilingual deep learning approach. *Data Knowl. Eng.* **2023**, *143*, 102106. [CrossRef]
13. Amazon Reviews for Sentiment Analysis. 2022. Available online: https://www.kaggle.com/datasets/bittlingmayer/amazonreviews (accessed on 1 December 2019).
14. Luque, F.M. Atalaya at tass 2019: Data augmentation and robust embeddings for sentiment analysis. *arXiv* **2019**, arXiv:1909.11241.
15. Wei, J.; Zou, K. Eda: Easy data augmentation techniques for boosting on text classification tasks. *arXiv* **2019**, arXiv:1901.11196.
16. Makcedward/Nlpaug. 2021. Available online: https://github.com/makcedward/nlpaug/blob/master/example/quick_example.ipynb (accessed on 6 April 2021).
17. Omran, T.; Sharef, B.; Grosan, C.; Li, Y. Ensemble Learning for Sentiment Analysis of Translation-Based Textual Data. In Proceedings of the 2022 International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME), online, 16–18 November 2022; pp. 1–9.

**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.