A New English/Arabic Parallel Corpus for Phishing Emails

Said Salloum *

School of Science, Engineering, and Environment, University of Salford, Salford, S.A.S.Salloum@edu.salford.ac.uk

Tarek Gaber

School of Science, Engineering, and Environment, University of Salford, Salford, t.m.a.gaber@salford.ac.uk

Faculty of Computers and Informatics, Suez Canal University, Ismailia 41522, Egypt

Sunil Vadera

School of Science, Engineering, and Environment, University of Salford, Salford, s.vadera@salford.ac.uk

Khaled Shaalan

Faculty of Engineering &IT, The British University in Dubai, khaled.shaalan@buid.ac.ae

Phishing involves malicious activity whereby phishers, in the disguise of legitimate entities, obtain illegitimate access to the victims' personal and private information, usually through emails. Currently, phishing attacks and threats are being handled effectively through the use of the latest phishing email detection solutions. Most current phishing detection systems assume phishing attacks to be in English, though attacks in other languages are growing. In particular, Arabic is a widely used language and therefore represents a vulnerable target. However, there is a significant shortage of corpora that can be used to develop Arabic phishing detection systems. This paper presents the development of a new English-Arabic parallel phishing email corpus that has been developed from the anti-phishing share task text (IWSPA-AP 2018). The email content was to be translated, and the task had been allotted to 10 volunteers who had a university background and were English and Arabic language experts. To evaluate the effectiveness of the new corpus, we develop phishing email detection models using Term Frequency–Inverse Document Frequency (TF-IDF) and Multilayer Perceptron using 1258 emails in Arabic and English that have equal ratios of legitimate and phishing emails. The experimental findings show that the accuracy reaches 96.82% for the Arabic dataset and 94.63% for the emails in English, providing some assurance of the potential value of the parallel corpus developed.

CCS CONCEPTS • Computing methodologies → Language resources; Phishing Emails Detection.

Additional Keywords and Phrases: English–Arabic Parallel Corpus, Phishing Emails, Multilayer Perceptron, Term Frequency–Inverse Document Frequency.

1 INTRODUCTION

Phishing is a deceitful social engineering method. The act of phishing involves the use of electronic platforms for fraudulent access to sensitive data and information (including usernames, passwords and credit card information) with the intention of abusing this sensitive data for personal benefits. The culprit usually pretends to be a reliable party. Phishing activities mainly involve email spoofing and text messaging, which are meant to deceive the users. The phisher-generated email or message directs users to a fake website disguised as a legitimate one wherein users are coerced to provide their personal information [1]. Currently, phishing attacks and threats are handled through the use of the latest phishing email detection solutions that are developed using a corpus.

* Place the footnote text for the author (if applicable) here.

2375-4699/2023/1-ART1 $15.00
https://doi.org/10.1145/3606031

Data in a corpus can be in one or multiple languages (e.g. parallel corpora). Arabic and English are both popular languages used extensively across the globe. Arabic enjoys the status of the official language in over 20 countries. Arabic encompasses many different dialects of the Central Semitic languages. English is popularly recognised as the principal language used in business and education and establishes a common communication link between people speaking different languages. Two of the six official languages of the United Nations are Arabic and English [2].

The significance of language corpora in Natural Language Processing (NLP), specifically for Machine Translation (MT), has increased in recent years. Corpus enables the researchers to achieve better results during the training of statistical-based language modelling. Statistical-based MT systems can be developed with the help of large-scale parallel corpora. There are various applications of parallel corpora. For instance, they can be applied to train sequence-to-sequence paraphrasing models. These corpora are also helpful during summarisation, machine translation, text simplification and various paraphrastic tasks. Besides, model training, parallel corpora are applied in information retrieval, question answering and other analysis tasks. Keeping in view the current shortage of parallel English–Arabic phishing emails, this paper develops a new English–Arabic parallel corpus.

The rest of this paper is organised as follows. Section 2 highlights the features of the Arabic Language that make it different from English, Section 3 presents the related work on Arabic-English Corpora, and Section 4 presents how the new corpus was created. Sections 5 to 7 present the experimental settings, features used, and result and evaluation, Section 8 concludes the paper.

1.1 Arabic: A Global Language

For over 400 million people in the world, the primary language spoken is [3], [4]. Over the years, much advancement has been made for Arabic-related computing research and applications. Specifically, a large number of individuals use the Arabic language for Internet access [5]. There are several users who only speak the Arabic language and nothing else, which is why they are unable to comprehend vast amounts of the English data present [6]. Furthermore, there has been a worldwide expansion of interest within Arabic nations in terms of economics, politics, culture and other aspects. Since the Arabic speakers are in large numbers and English holds a significant importance, it is necessary that the language translation be carried out using high-quality parallel corpora. Yet, MT is faced with challenges due to the structural difference amongst the languages. As compared to the European languages, Arabic needs separate treatment since it has an exclusive morphology, and as indicated in Table 1, in terms of graphology features, English and Arabic are quite different.

Table 1: The difference between English and Arabic.

	English	**Arabic**
Connection	Usually, diagonal strokes link each character to the next	In Arabic letters, the baseline is connected with horizontal strokes
Character versions	Characters have limited shape variations in English	According to their relative position in the word, Arabic letters might have up to four distinct shapes
Capitalisation	Yes	No
Direction	Follows the left-to-right direction in reading/writing	Follows the right-to-left direction in reading/writing
Features	English-writing has a specific geometrical feature	The letters or segmented sub-letters are different from the segments in English
Gender differentiation	No differentiation	Verb and sentence structure
Language codes	en | eng	ar | arb
Plural forms	Singular and plural	Singular, dual and plural
Position of Adjective	Before the noun	After the noun
Place with most speakers	The United States of America	Egypt
Segmentation	Handwriting can be segmented into different letters or sub-letters using any analytical segmentation method	The letters or segmented sub-letters vary from those in English
Size of alphabet	26 letters	28 letters
Types of sentences	Verbal	Nominal and verbal
Total speakers	1.348 billion	274 million

1.2 Arabic–English parallel corpora

Recently, Arabic researchers have been focusing extensively on parallel corpora. The awareness about the importance of parallel corpora among masses has also elevated. However, there are not many Arabic–English parallel corpora available in the literature. Research indicates ([7], p. 327) that the shortage of these corpora may be attributed to the limited availability of financial and material resources and the prevailing uncertainty of the concerned authorities about the effectiveness and significance of corpora. Among the most prominent English–Arabic corpora projects is "the English–Arabic Parallel Corpus of the United Nations Texts (EAPCOUNT)", based on 341 paragraph-aligned texts [8]. A compilation of a couple of sub-corpora,

it includes 5,392,491 words. One subset includes English content in original form, while the other includes the corresponding Arabic translations. The corpus was developed by compiling textual content from UN resolutions and annual reports and texts extracted from the literature issued by international institutions [8]. Likewise, another such project was sponsored by the European Commission whereby the experts at the Language Technology Lab in Germany developed a multilingual parallel corpus, MultiUN [9]. This 300-million word long parallel corpus was actually a compilation of chunks of data obtained from the UN documents issued at the UN's official website during the period from 2000 to 2009 (see [9]). Another parallel corpus, namely the Open Parallel Corpus (OPUS), was developed by Tiedemann (2012) [10], who offered it to be used as a free multilingual parallel corpus; he obtained online translated texts and compiled them into the corpus. OPUS allowed parallel as well as monolingual data to be processed through its open-source tools; it also facilitated the research process by offering a number of search interfaces. The OPUS was developed on an automatic basis without involving any manual processing, as mentioned on the website. The European Union sponsored the development of a multilingual parallel corpus namely the EuroMatrix based on texts taken from the European Parliament proceedings; the extracted texts were originally in English and translated into Arabic and more languages. This EU-developed corpus contained 1.5 million Arabic words out of a total of 51 million words. This corpus intended to facilitate and support machine translation systems. Considering the corpus development in Arab countries, experts at the Kuwait University obtained extracts of Arabic translations from the book series "World of Knowledge" and compiled it to formulate a parallel corpus; this book series was issued in Kuwait by the National Council for Culture, Arts and Letters (NCCAL). There were a total of 3 million words in this corpus; the corpus could only be accessed and used by staff and students associated with the Kuwait University, specifically those enrolled for the lexicography and translation programs [7]. A number of projects of parallel corpora (including the Arabic language projects) had been initiated by the Linguistic Data Consortium (LDC). GALE Phase 2 Arabic Broadcast News Parallel Text is among their prominent projects containing data obtained and recorded under the LDC's supervision. The data contained extracts of news aired from 2005 to 2007 in the form of Arabic source texts with its English translations. There were 60 source-translation document pairs containing 42,089 Arabic source text words with corresponding English translations in the corpus (See [11]). Moreover, the LDC automatically incorporated texts from a couple of monolingual corpora, including the Arabic Gigaword Second Edition (LDC2006T02) and English Gigaword Second Edition (LDC2005T12) to come up with Arabic-English Automatically Extracted Parallel Text. This corpus was based on Chinese and French new articles issued by the Xinhua News Agency (Chinese) and Agence France-Presse (French). There were 1,124,609 sentence pairs in the corpus with about 31 million English words (See [11]). Another multilingual corpus was developed at UMIST by [12]. The corpus contained texts pertaining to IT in the English language with Arabic and Swedish translational corpora. There were 1 million tokens of Arabic text and 2.7 million tokens of Swedish text. The IT content was extracted from multilingual IT websites and included guides and manuals meant for instructing the users of computer systems, hardware, and software. The corpus can only be used by researchers after obtaining prior copyright approval, as it cannot be freely accessed by the public. The Qatar Computing Research Institute also executed the corpus of AMARA [13], [14] that extracted data from educational platforms like Technology, Entertainment, and Design (TED) and the Khan Academy in the form of video captions developed by the community. The corpora contained both Arabic (2.6 million) words and English (3.9 million) words. This corpus was designed for the machine translation of data. The

corpus was equipped with an editor which allowed the generation of subtitles (see [15]). Data including 27.8 million Arabic words and 30.8 million English words for a parallel corpus was collected by [16]. He extracted data from the Al-Hayat newspaper and the OPUS corpus. The corpus was anticipated to facilitate researchers exploring machine translation. Similarly, Hassan and Atwell (2016) [17] compiled about 2 million holy words of the Prophet Mohammad (Peace Be Upon Him) to develop a Hadith corpus in Arabic with translations in multiple languages of English, French, and Russian.

The main issue is the limited number of publicly available text corpora [18]–[20] for phishing email. Recently, it has become common to devise linguistic tools on the basis of parallel corpora (i.e. Tokenizer, Part-of-Speech (PoS) Taggers, Stemmers, Lemmatizers, Named Entity Recognition (ENR)). Similarly, the parallel corpora and parallel text processing are characterised with limited representation of the Arabic language. This research mainly intends to develop an English–Arabic parallel phishing email corpus created from the English and Arabic text provided by the leading security and privacy analytics anti-phishing shared task (IWSPA-AP 2018) to address the issue of a lack of superior quality English/ Arabic parallel texts; this was quite helpful for the researchers, since the absence of parallel corpus left the researchers with no choice but to resort to manual translations that proved to be inaccurate and involved lengthy procedures [21]. Additionally, most available corpora were not feasible to be used by students and researchers due to their high cost and lower quality. Hence, the research mentioned in this study was particularly aimed at coming up with a parallel corpus to fulfil the needs of the researchers intending to design statistical translation software who required parallel corpora for the training of statistical models. Hence, the success of corpora being proposed in this study is expected to bridge this gap and provide a cheap and effective alternative to those who fail to afford the expensive data. Table 2 presents the statistics about English–Arabic parallel corpus, including the amount of Arabic and English words for both phishing and legitimate e-mails.

Table 2: English–Arabic parallel corpus for email phishing.

Language	Type		Total
	Legitimate emails	Phishing emails	
Arabic Words	24138	18262	42400
English Words	23554	18079	41633
Total	47692	36341	84033

2 ARABIC-ENGLISH PHISHING EMAIL CORPUS

In this section, we will describe how the Arabic-English Phishing Email Corpus has been created.

2.1 Dataset description

This paper presents the development of a new English-Arabic parallel phishing email corpus that has been developed from the anti-phishing share task text (IWSPA-AP 2018) corresponded with the 8th ACM Conference on Data and Application Security and Privacy in detecting phishing email through an anti-phishing shared task [22], which is a common practice associated with machine learning and text analysis in the area of cybersecurity. The organisers of (IWSPA-AP 2018) [EDMB+18] provided the email corpus. Researchers in [23]–[32] evaluated their models by using IWSPA-AP 2018 dataset. The point of the anti-phishing shared undertaking is to assemble a classifier to differentiate phishing emails from spam and authentic email. The two sub-tasks can be accommodated within unconstrained categories, which implies

that members undertaking training may use any other external corpus. The anti-phishing shared tasks involve two sub-tasks: the first one is associated with the testing of emails with a header, while the second is associated with the testing of emails without a header. The descriptive statistics of training and testing email corpus related to these tasks are summed up in Table 3 and Table 4

Table 3: Training email corpus details.

Training Dataset	Legitimate	Spam	Total
With header	4082	501	4583
Without header	5088	612	5700

Table 4: Testing email corpus details.

Training Dataset	Data Samples
With header	4195
Without header	4300

2.2 Creating the Corpus

The IWSPA-AP version 2.0 training dataset will be used by us in order to build the Arabic–English parallel corpus. There exists two approaches: (i) translation by making use of the free APIs provided by service providers as Google or Microsoft (ii) working from scratch for the translation of the English text. A good translation is not provided by the first technique because of the translation quality. Therefore, the second one is used in this work, i.e., translating from scratch with the aid of 10 volunteers who are English and Arabic language experts. The 10 translators were divided into two groups, with one group translating and the other checking the translation for grammatical and spelling errors. The roles were then exchanged for the next batch and the process repeated until all the emails were translated. The translators were asked to make sure that:

1) The Modern Standard Arabic language is followed.
2) A valuable sentence is written, which must end using a period (.).
3) Multiple phrases or words should not be typed.
4) The speech style used should be polite, and punctuation marks must be correct.
5) Factual data should only be provided when commenting on the email.
6) One must not mention aspects which may occur in the future.
7) Imagination and speculation shouldn't be present.
8) Feelings related to the email scene should not be stated.
9) Poetic style shouldn't be used excessively.
10) Must not write the nationality, places or person names like American Flag or Washington City.
11) All essential details to be mentioned and non-essential ones to be ignored.

The translation and proofreading process took six months from 15 September 2021 through 15 March 2022, 12 weeks, and 12 weeks for proofreading and quality control. Every month, 200 emails were translated and audited (100 legitimate emails and 100 phishing emails). In the first week of every month, 100 e-mails presented in an MS Word file were distributed equally to the ten volunteers, with ten new e-mails for each

volunteer (5 legitimate e-mails and 5 phishing e-mails). In the first week of every month, the first group checks the emails translated by the second group, and the second group also checks the emails translated by the first group and ensures the level of the translated text. The sixth volunteer checks the emails translated by the first volunteer, while the seventh volunteer checks the emails translated by the second volunteer, and so on. The translator must complete the process of transferring the text and its content with all credibility and integrity. He should not delete or add anything on his own and according to his whims, and he can add some and a few margins for clarification if the translator wants this matter, and the translator should not highlight or show what his personal view is in the content of the text he is translating, as well as working to take into account the nature of the repeated words and their meaning. The process is repeated in the third and fourth weeks of every month.

2.3 Judging the quality of the translation

Figure 1 illustrates the process of translating texts from English to Arabic, including the quality control procedures taken. The following criteria are used to judge the level of quality of the translation process:

1. Coherence of meaning and work to achieve consistency.
2. Integration and Comprehensiveness.
3. Matching the method.
4. Grammar and spelling.

Table 5 shows a sample of the translation correction process that includes checking the punctuation, rewording some sentences, and solving the ambiguity of some cultural expressions. Figure 2 also depicts a sample sentence taken from the corpus with sentence ID number as well as English and Arabic translations of the sample sentence.

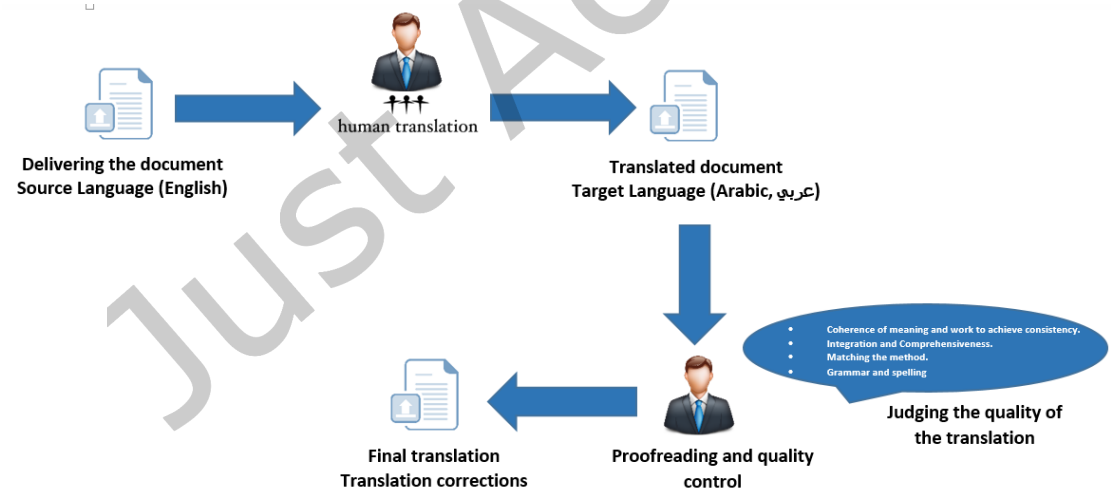

Figure 1: Human translation from English to Arabic.

Table 5: Examples of translation corrections.

No.	Source (English Text)	Translation (Arabic Text)	Corrected (Arabic Text)
1	This is an urgent notice from the board of governors federal reserve bank washington DC. Open attached letter and read carefully and respond accordingly.	هذا إشعار عاجل من مجلس محافظي بنك الاحتياطي الفيدرالي في واشنطن العاصمة. افتح الرسالة المرفقة واقرأها بعناية واستجب وفقًا لذلك.	هذا إشعار عاجل من مجلس إدارة محافظي البنك الاحتياطي الفيدرالي في واشنطن العاصمة. افتح الرسالة المرفقة واقرأها بعناية واستجب وفقًا لذلك.
2	As part of our duty to strengthening our security and improving your overall mail experience, we have detected your mail settings is out of date. We want to upgrade all email account scheduled for today. To Complete this procedure, CLICK HERE to upgrade your account. If your settings is not updated today, your account will be inactive and cannot send or receive message any longer.	كجزء من واجبنا في تعزيز أمننا وتحسين تجربة بريدك الإجمالية، اكتشفنا أن إعدادات بريدك قديمة. نريد ترقية جميع حسابات البريد الإلكتروني المقررة اليوم. لإكمال هذا الإجراء، انقر هنا لترقية حسابك. إذا لم يتم تحديث إعداداتك اليوم، فسيكون حسابك غير نشط ولا يمكنه إرسال أو استقبال الرسالة بعد الآن.	كجزء من واجبنا المتمثل في تعزيز الأمان وتحسين تجربة البريد بشكل عام ، اكتشفنا أن إعدادات البريد لديك قديمة. نريد ترقية جميع حسابات البريد الإلكتروني المجدولة لهذا اليوم. لإكمال هذا الإجراء ، انقر هنا لترقية حسابك. إذا لم يتم تحديث إعداداتك اليوم ، فسيكون حسابك غير نشط ولا يمكنه إرسال أو استقبال الرسائل بعد الآن.
3	This is to notify all Students, Staffs of organization that we are validating active accounts. Kindly confirm that your account is still in use by clicking the validation link below:	هذا لإخطار جميع الطلاب وموظفي التنظيم بأننا نتحقق من صحة الحسابات النشطة. يرجى التأكيد على أن حسابك لا يزال قيد الاستخدام من خلال النقر فوق رابط التحقق أدناه:	هذا لإخطار جميع الطلاب وموظفي المؤسسة بأننا نتحقق من صحة الحسابات النشطة. يرجى التأكد من أن حسابك لا يزال قيد الاستخدام من خلال النقر على رابط التحقق أدناه:

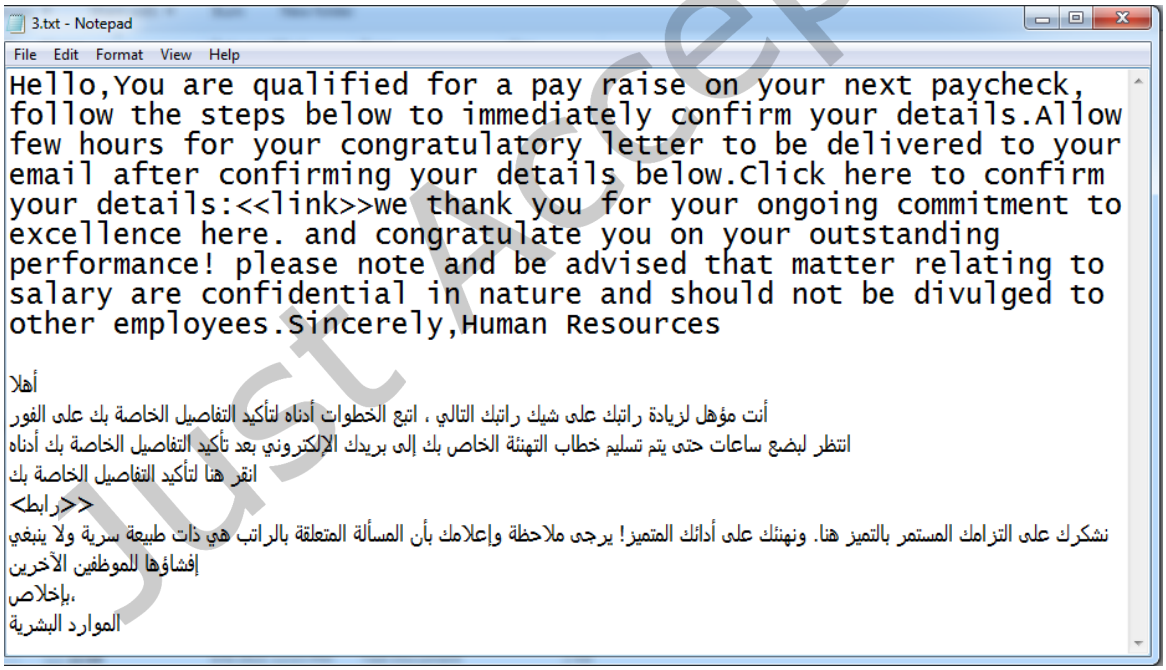

Figure 2: Example from the new Arabic/English parallel corpus.

3 EXPERIMENTAL SETTINGS

3.1 Data Pre-processing

Data pre-processing is an important step in detecting phishing emails. Cleaning of data is necessary so that all unnecessary words and characters are removed. After data cleaning, it is possible to use formal feature extraction. The following are common steps in data pre-processing for phishing emails:

- **Text Cleaning**: Remove punctuation, numbers, and special characters, lowercase all text.
- **Removing Stopwords**: Remove common words such as "and", "the", "is", etc. which do not add meaning to the text.
- **Tokenization**: Divide text into individual words or phrases.
- **Stemming/Lemmatization**: Reduce words to their root form to reduce dimensionality.
- **Vectorization**: Convert text into numerical representations, such as bag-of-words or word embeddings.
- **Balancing Classes**: Handle imbalanced datasets by oversampling or undersampling to ensure that both classes are represented equally.
- **Split Data**: Divide data into training and testing sets to evaluate the performance of the model.

By implementing these pre-processing steps, the data is prepared for further analysis and modeling in detecting phishing emails.

3.2 Email cleaning

Through this procedure, emails are cleaned, and unnecessary information and non-English/Arabic characters are removed. Using the Python Regex, the non-special characters, such as “?”, “!”, and “”, and alphanumeric characters are not removed. White spaces are removed. The pseudocode applied is mentioned in Figure 3 along with the example.

Input	*** To be automatically unsubscribed from this list, please email <Link> *** *** ليتم إلغاء الاشتراك تلقائيًا من هذه القائمة ، يرجى إرسال بريد إلكتروني إلى <رابط> ***
Output	To be automatically unsubscribed from this list please email Link ليتم إلغاء الاشتراك تلقائيًا من هذه القائمة يرجى إرسال بريد إلكتروني إلى رابط

Figure 3: Email cleansing.

3.3 Email cleaning

Through this procedure, emails are cleaned, and unnecessary information and non-English/Arabic characters are removed. Using the Python Regex, the non-special characters, such as “?”, “!”, and “”, and alphanumeric characters are not removed. White spaces are removed. The pseudocode applied is mentioned in Figure 3 along with the example.

3.3.1Tokenisation

Through tokenisation [33], each email is broken down into words keeping in mind the white spaces. The words are then considered to be tokens. The split () function is applied in the current study. The tokenisation procedure is illustrated in Figure 4.

Input	To be automatically unsubscribed from this list please email Link ليتم إلغاء الاشتراك تلقائيًا من هذه القائمة يرجى إرسال بريد إلكتروني إلى رابط
Output	[‘To’, ‘be’, ‘automatically’, ‘unsubscribed’ ,‘from’ ,‘this’, ‘list’ ‘please’, ‘email’ ,‘Link’] [‘ليتم’,’إلغاء’, ’الاشتراك’, ’تلقائيًا’, ’من’, ’هذه’, ’القائمة’, ’يرجى’, ’إرسال’, ’إلى’, ’بريد’,’إلكتروني’]

Figure 4: Email tokenization.

3.3.2Stop-word and rare-word removal

The commonly used words are referred to as stop words, and they are the ones which help create ideas. However, they do not have a significance of their own, unlike conjunctions, articles, prepositions, and the like. The stopwords list has been extracted from the Natural Language Toolkit (NLTK) [34]. Words in English such as “off”, “no”, “aren’t”, “too”, “an”, “being”, “only”, “ll”, “o”, “its”, “them”, and “might” and in Arabic such as "في“, “كل", “حيث”, “انه”, “قال”, “ من""", are part of the stop words list. A word which is present seven times or fewer is removed. Figure 5 indicates the stop-word and rare-word removal procedure.

Input	['To', 'be', 'automatically', 'unsubscribed' ,'from' ,'this', 'list' 'please', 'email' ,'Link'] ['ليتم', 'إلغاء', ' الاشتراك', 'تلقائيًا', ' من', 'هذه', 'القائمة', ' يرجى', ' إرسال', ' إلى', ' بريد', 'إلكتروني']
Output	['automatically', 'unsubscribed' , 'list' 'please', 'email' ,'Link'] ['ليتم', 'إلغاء', ' الاشتراك', 'تلقائيًا', 'القائمة', ' يرجى', 'إرسال', 'إلكتروني', 'بريد' , 'رابط']

Figure 5: Stopword removal.

4 FEATURES USED

Feature extraction is the process of transforming raw data into acceptable inputs (i.e., features) that may be processed by a machine learning algorithm. To put it another way, the extracted features must reflect the primary textual material in a manner that most suits the requirements of the classifier algorithm applied. Minimal feature extraction is usually required, except deep learning neural networks, which can conduct feature extraction on their own. Furthermore, a weak classifier fed with relevant features is thought to outperform a robust classifier fed with low-quality features. Bag-of-Words (BoW) [35]–[38], Document-Term Matrix (DTM) [39], Term Frequency–Inverse Document Frequency (TF-IDF) [40]–[45], Word Embeddings [46]–[48], and Character-level Convolutional Networks [46], [48], [49] are prominent features extraction methods for Arabic text classification. The TF-IDF is used in this research. BoW just creates a set of vectors representing the number of times a word appears in a document. The TF-IDF, on the other side, assigns a score to every word in a document that represents how important that word is in that document. The data on the more significant words, as well as the less significant ones, will be included in every document. As a result, documents containing comparable words will have identical vectors.

4.1 TF-IDF

The TF-IDF (Term Frequency-Inverse Document Frequency) weight is employed in information retrieval to measure the value of a word to a document in a set of documents. The relevance of a word rises in direct relation to the quantity of occurrence in the document (term frequency) and falls in inverse relation to the word's document frequency in the set. The IDF is a measure of the term's discriminating power. It calculates the frequency of a term over many documents. As a result, a word with a high term frequency in one document and a lower document frequency in the entire set of documents has a high TF-IDF weight.

We employed the term frequency-inverse document frequency word weights, or TF-IDF weights, as clustering features, as described in [50]. The following parameters and terminology are used to construct

these weights. Let us assume we are extracting features from a data set **E** made up of |**E**| emails. Let N (*w; m*) be the quantity w appears in *m* for a word *w* and an email m. Assume you are looking at a set T = {*t1…tk*} of terms *t1*...., *tk*. TF (*w*, *m*) denotes the repetends of a word $\boldsymbol{w \in T}$ in an email *m* and is described as the quantity ***w*** appears in *m*, normalized over the number of repetends of all words in ***m***:

$$\text{TF (w, m)} = \frac{N\,(w,m)}{\sum_{i=1}^{k} N(ti,m)} \qquad (1)$$

DF(*w*) stands for the document frequency of the word *w*, which is described as the proportion of emails in a data set in which the word *w* appears at minimum once. To determine the importance of every term, the inverse document frequency is employed. IDF (*w*) is the symbol for it, and it is determined by the formula below.

$$\text{IDF } (w) = \log\left(\frac{|E|}{DF\,(w)}\right) \qquad (2)$$

The TF-IDF weight of *w* in *m*, or term frequency-inverse document frequency of a word *w* in email *m*, is specified as

$$\text{TF-IDF } (w, m) = \text{TF } (w, m) \times \text{IDF } (w, m) \qquad (3)$$

We compiled a list of words with the maximum TF-IDF values across the entire data set of emails. The TF-IDF values of these words in the email were calculated for every email. These weights and other features were compiled into a vector. We employed Genism, a Python and NumPy package for vector space modeling of text documents, to calculate the TF-IDF values.

5 EXPERIMENTAL RESULT AND EVALUATION

5.1 Experimental Setup

All of the experiments in this paper were conducted using a PC Lenovo "LEGION 5" with (15IMH05H GAMING Core™ i7-10750H 2.6 GHz 1TB +512 GB SSD 16GB 15.6" (1920x1080) 144 Hz BT WIN10). The scikit learn, TensorFlow, and Keras libraries are used to develop the models. All the methods were evaluated using an 80/20 split of training and testing. Multilayer Perceptron (MLP), k-Nearest Neighbor (KNN), Decision Trees (DT), Logistics Regression (LR), Support Vector Machines (SVM), Random Forest (RF), Naive Bayes (NB), and XGBoost classifiers were utilized in this research.

5.1.1Multilayer Perceptron (MLP)

Multilayer Perceptron (MLP) is referred to as the feed-forward artificial neural network which includes several layers, mostly 3, of the neurons. Each of these neurons is referred to as a processing unit that can be activated by applying the activation function. This MLP is a supervised machine learning procedure where the network is trained with the help of a labelled training data set. Using a trained MLP, it would be possible to map the input data set (email features in this case) into the output set (email class). The MLP classifier for the present system contains these parameters:

*"class sklearn.neural_network.MLPClassifier(hidden_layer_sizes=(100,), activation='relu', *, sover='adam', alpha=0.0001, batch_size='auto', learning_rate='constant', learning_rate_init=0.001, power_t=0.5, max_iter =200, shuffle=True, random_state=None, tol=0.0001, verbose=False, warm_start=False, momentum=0.9, ne sterovs_momentum=True, early_stopping=False, validation_fraction=0.1, beta_1=0.9, beta_2=0.999, epsilon =1e08, n_iter_no_change=10, max_fun=15000".*

5.1.2The k-nearest neighbors algorithm (KNN)

Presently, academics prefer the KNN classifier because it is easy, polished, and straightforward. If new sample data *x* occurs, KNN will use some distance measure to find the k neighbors closest to the unlabeled data starting from the training space. These are parameters that were used in the study:

*"class sklearn.neighbors.KNeighborsClassifier(n_neighbors=5, *, weights='uniform', algorithm='auto', leaf_s ize=30, p=2, metric='minkowski', metric_params=None, n_jobs=None)".*

5.1.3Decision Trees (DT)

Regarding classification and regression, Decision Trees (DTs) are a non-parametric supervised learning approach. The objective is to learn simple decision rules from data features to develop a model that predicts the significance of a target variable. A tree is an equivalent to a piecewise constant. A Decision Tree Classifier is a class that can classify a dataset into multi-class. DecisionTreeClassifier, like other classifiers, requires two arrays as input: a sparse or dense array X of shape (n samples, n features) containing the training samples, and an array Y of integer values of shape (n samples) containing the class labels for the training samples. These are the parameters that were used in the ongoing study:

"class sklearn.tree.DecisionTreeClassifier(, criterion='gini', splitter='best', max_depth=None, min_sample s_split=2, min_samples_leaf=1, min_weight_fraction_leaf=0.0, max_features=None, random_state=None, max_leaf_nodes=None, min_impurity_decrease=0.0, class_weight=None, ccp_alpha=0.0)".*

5.1.4Logistics Regression (LR)

Logistic regression (LR) is called logit regression, maximum-entropy classification (MaxEnt), or the log-linear classifier in academia. A logistic function is used to model the probability of the probable results of a particular attempt in this model. In Logistic Regression, logistic regression is used with configurable ℓ1, ℓ2, or Elastic-Net regularization, this solution can accommodate binary, One-vs-Rest, or multinomial logistic regression. The LR topology layout for the present model contains these parameters:

*"class sklearn.linear_model.LogisticRegression(penalty='l2', *, dual=False, tol=0.0001, C=1.0, fit_intercept =True, intercept_scaling=1, class_weight=None, random_state=None, solver='lbfgs', max_iter=100, multi_cl ass='auto', verbose=0, warm_start=False, n_jobs=None, l1_ratio=None)".*

5.1.5Support Vector Machines (SVM)

The SVM has been used for pattern recognition and classification problems since it is deemed straightforward and adequate for the computation of machine learning algorithms. As opposed to other classifiers, the classification performance is relatively efficient due to the minimal training data. As a result, the textual data was categorized by employing the support vector machine in the ongoing study, with these parameters.

"class sklearn.svm.SVC(, C=1.0, kernel='rbf', degree=3, gamma='scale', coef0=0.0, shrinking=True, probab ility=False, tol=0.001, cache_size=200, class_weight=None, verbose=False, max_iter=- 1, decision_function_shape='ovr', break_ties=False, random_state=None)".*

5.1.6Random Forest (RF)

A random forest (RF) is a meta estimator that employs averaging to increase predictive accuracy and minimize overfitting by adopting a set of decision tree classifiers on different sub-samples of the dataset. If bootstrap=True (default), the sub-sample size is specified by the max sample's parameter; alternatively, the entire dataset is utilized to create every tree. These are parameters that were used in the study:

*"class sklearn.ensemble.RandomForestClassifier(n_estimators=100, *, criterion='gini', max_depth=None, mi n_samples_split=2, min_samples_leaf=1, min_weight_fraction_leaf=0.0, max_features='auto', max_leaf_node s=None, min_impurity_decrease=0.0, bootstrap=True, oob_score=False, n_jobs=None, random_state=None, verbose=0, warm_start=False, class_weight=None, ccp_alpha=0.0, max_samples=None)".*

5.1.7Naive Bayes (NB)

The "naive" supposition of conditional independence in between each pair of features provided the value of the class variable is used in Naive Bayes approaches, which are a collection of supervised learning algorithms depending on employing Bayes' theorem. Within the present model, the following parameters are present for the NB topology.

"class sklearn.naive_bayes.GaussianNB(, priors=None, var_smoothing=1e-09)".*

5.1.8Extreme Gradient Boosting (XGBoost)

Using gradient boosted decision trees, the module sklearn.ensemble offers approaches for both classification and regression. The Gradient Boosting Classifier and Gradient Boosting Regressor are given underneath, along with their parameters and application. These estimators' most essential parameters are n_estimators and learning_rate. Within the current research, the following parameters have been applied:

"*class sklearn.ensemble.GradientBoostingClassifier(*, loss='deviance', learning_rate=0.1, n_estimators=100, subsample=1.0, criterion='friedman_mse', min_samples_split=2, min_samples_leaf=1, min_weight_fraction_*

leaf=0.0, max_depth=3, min_impurity_decrease=0.0, init=None, random_state=None, max_features=None, v erbose=0, max_leaf_nodes=None, warm_start=False, validation_fraction=0.1, n_iter_no_change=None, tol= 0.0001, ccp_alpha=0.0)".

5.2 Model Evaluation

A confusion matrix is a type of table that is commonly used in machine learning to assess the performance of a classification model. The table compares the predicted and actual classes of the model to determine how well it is performing. It is a square matrix that has rows and columns representing the predicted and actual classes, respectively. The diagonal elements of the matrix indicate the number of instances that are correctly classified, while the off-diagonal elements indicate the number of instances that are misclassified. To gain a better understanding of these measures, one can examine the confusion matrix presented in Table 6.

Multilayer Perceptron (MLP), k-Nearest Neighbor (KNN), Decision Trees (DT), Logistics Regression (LR), Support Vector Machines (SVM), Random Forest (RF), Naive Bayes (NB), and XGBoost classifiers were utilized in this research.

Table 6: Confusion matrix.

	Predicted Negative	**Predicted Positive**
Actual Negative	True Negative (TN)	False Positive (FP)
Actual Positive	False Negative (FN)	True Positive (TP)

The actual classes are represented by the rows and the predicted classes are represented by the columns. The four elements of the confusion matrix are defined as follows:

- True Positive (TP): The number of instances that are actually positive and are correctly classified as positive by the model.
- False Positive (FP): The number of instances that are actually negative but are incorrectly classified as positive by the model.
- True Negative (TN): The number of instances that are actually negative and are correctly classified as negative by the model.
- False Negative (FN): The number of instances that are actually positive but are incorrectly classified as negative by the model.

Using the values in the confusion matrix, various performance metrics can be calculated, such as accuracy, precision, recall, and F1 score. These metrics can provide insight into how well the classification model is performing and can be used to make adjustments to improve the model's performance.

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \qquad (4)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}}$$

$$\text{F1-measure} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (6)$$

$$\text{Accuracy} = \frac{(\text{TP}+\text{TN})}{(\text{TP}+\text{FP}+\text{TN}+\text{FN})} \quad (7)$$

5.3 Overall Result

Table 7 presents the outcome of applying the classifiers for filtering the phishing emails written in English. The highest accuracy level is indicated by MLP 94.63%, then precision, 94.91%, recall, 94.76%, and F1-Score, 94.78%. Table 8 presents the results on the Arabic text. Outcomes indicate that the highest accuracy level, 96.82%, is for MLP followed by precision, 97.10%, recall, 96.56%, and F1-Score, 96.77%. The results indicate that the complete performance for the classifiers is much more efficient when using the Arabic Corpus as compared to the English Corpus. We found that the difference between the training accuracy between the English and Arabic text is very small and this indicates the quality of the translation. The results also showed that the Logistic Regression (LR) has a short training time to Arabic & English corpus using TF-IDF.

Table 7: Results of classical ML classifiers to English corpus using TF-IDF.

Classifier	Accuracy	Precision	Recall	F1-Score	TP	FN	FP	TN	Time (sec.)
MLP	**94.63%**	**94.91%**	**94.76%**	**94.78%**	42.06%	3.17%	1.98%	52.78%	1s
KNN	91.66%	91.63%	92.00%	91.64%	43.25%	1.98%	6.35%	48.41%	0.0974s
DT	89.68%	89.58%	89.58%	89.58%	40.08%	5.16%	5.16%	49.60%	0.163s
LR	93.74%	93.14%	93.52%	93.76%	41.27%	3.97%	1.19%	53.57%	**0.0949s**
SVM	93.62%	93.01%	93.60%	93.77%	41.67%	3.57%	1.59%	53.17%	0.238s
RF	93.54%	93.72%	93.90%	93.80%	43.25%	1.98%	3.17%	51.59%	0.133s
NB	90.47%	90.89%	90.00%	90.29%	38.49%	6.75%	2.78%	51.98%	0.219s
XGBoost	91.66%	92.39%	91.09%	91.48%	38.49%	6.75%	1.59%	53.17%	18s

Table 8: Results of classical ML classifiers to Arabic corpus using TF-IDF.

Classifier	Accuracy	Precision	Recall	F1-Score	TP	FN	FP	TN	Time (sec.)
MLP	**96.82%**	**97.10%**	**96.56%**	**96.77%**	42.46%	2.78%	0.40%	54.37%	5s
KNN	89.68%	90.00%	90.27%	89.67%	43.65%	1.59%	8.73%	46.03%	0.113s
DT	89.28%	89.27%	89.07%	89.16%	39.29%	5.95%	4.76%	50.00%	0.205s
LR	95.23%	95.63%	94.88%	95.16%	41.27%	3.97%	0.79%	53.97%	**0.0975s**
SVM	96.42%	96.63%	96.20%	96.38%	42.46%	2.78%	0.79%	53.97%	0.369s
RF	94.84%	94.76%	94.83%	94.79%	42.86%	2.38%	2.78%	51.98%	0.158s
NB	88.49%	88.91%	87.96%	88.26%	37.30%	7.94%	3.57%	51.19%	0.384s
XGBoost	91.66%	92.20%	91.17%	91.50%	38.89%	6.35%	1.98%	52.78%	51.5s

6 CONCLUSION

This paper presented the development of a new corpus (English-Arabic Phishing Email) which created based on the translation of English email bodies extracted from IWSPA-AP v2.0 dataset. To test the suitability of this new corpus, the study used 8 machine learning algorithms to develop different phishing detection models. The experimental findings demonstrated that the overall accuracy of phishing email detection models reached 96.82% for the Arabic emails and 94.63% for the English emails, with the best results obtained when using MLP classifier and TF-IDF feature extraction technique. These outcomes have facilitated researchers to align direct key issues with multilingual corpora specifically with Arabic language processing. The researchers have been able to solve various technical and linguistic issues because of the language diversity and nature of corpus. However, there is still room for further research on the various presumptions and observations entailed in this paper. The researchers can work on the horizontal and vertical extension of the corpus in future research. The horizontal extension involves enhancing the corpus in size or introducing more language diversity in the corpus.

ACKNOWLEDGMENTS

This work is a part of a thesis submitted in fulfilment of a PhD in the School of Science, Engineering, and Environment at Salford University.

REFERENCES

[1] S. Salloum, T. Gaber, S. Vadera, and K. Shaalan, "Phishing Website Detection from URLs Using Classical Machine Learning ANN Model," in *International Conference on Security and Privacy in Communication Systems*, 2021, pp. 509–523.

[2] U. Nations, "UN Official Languages," 2013. [Online]. Available: https://www.un.org/%0Aen/aboutun/languages.shtml.

[3] S. A. Salloum, M. Al-Emran, and K. Shaalan, "A Survey of Lexical Functional Grammar in the Arabic Context," *Int. J. Com. Net. Tech*, vol. 4, no. 3, 2016.

[4] A. Farghaly and K. Shaalan, "Arabic natural language processing: Challenges and solutions," *ACM Trans. Asian Lang. Inf. Process.*, vol. 8, no. 4, p. 14, 2009.

[5] A. A. Rafea and K. F. Shaalan, "Lexical analysis of inflected Arabic words using exhaustive search of an augmented transition network," *Softw. Pract. Exp.*, vol. 23, no. 6, pp. 567–588, 1993.

[6] K. Shaalan, M. Attia, P. Pecina, Y. Samih, and J. van Genabith, "Arabic word generation and modelling for spell checking.," in *LREC*, 2012, pp. 719–725.

[7] H. Al-Ajmi, "A new English–Arabic parallel text corpus for lexicographic applications," *Lexikos*, vol. 14, 2004.

[8] H. Salhi, "Investigating the complementary polysemy and the Arabic translations of the noun destruction in EAPCOUNT," *Meta J. des traducteurs/Meta Transl. J.*, vol. 58, no. 1, pp. 227–246, 2013.

[9] A. Eisele and Y. Chen, "MultiUN: A Multilingual Corpus from United Nation Documents.," in *LREC*, 2010.

[10] J. Tiedemann, "Parallel data, tools and interfaces in OPUS.," in *Lrec*, 2012, vol. 2012, pp. 2214–2218.

[11] "Linguistic Data Consortium (LDC)," *LDC catalog.*, 2013. .

[12] S. Izwaini, "A corpus-based study of metaphor in information technology," in *Corpus Linguistics*, 2003.

[13] A. Abdelali, F. Guzman, H. Sajjad, and S. Vogel, "The AMARA Corpus: Building Parallel Language Resources for the Educational Domain.," in *LREC*, 2014, vol. 14, pp. 1044–1054.

[14] F. Guzmán, H. Sajjad, S. Vogel, and A. Abdelali, "The AMARA corpus: Building resources for translating the web's educational content," in *Proceedings of the 10th International Workshop on Spoken Language Technology (IWSLT-13)*, 2013.

[15] "AMARA." [Online]. Available: www.amara.org.

[16] S. Alkahtani and W. J. Teahan, "A new parallel corpus of Arabic/English," in *Proceedings of the Eighth Saudi Students Conference in the UK*, 2016, pp. 279–284.

[17] S. M. O. Hassan and E. S. Atwell, "Design and implementing of multilingual Hadith corpus," *Int. J. Recent Res. Soc. Sci. Humanit.*, vol. 3, no. 2, pp. 100–104, 2016.

[18] J. Nazario, "'Nazario's phishing corpora,' accessed: Jan 16, 2020." [Online]. Available: https://monkey.org/~jose/phishing/.

[19] C. Project, "'Enron email dataset,' accessed: Jan 16, 2020." [Online]. Available: http://www.cs.cmu.edu/~enron/.

[20] "The Apache Spamassassin Public Corpus." [Online]. Available: https://spamassassin.apache.org/old/publiccorpus.

[21] K. Taghipour, S. Khadivi, and J. Xu, "Parallel corpus refinement as an outlier detection algorithm," *Proc. 13th Mach. Transl. Summit (MT Summit XIII)*, pp. 414–421, 2011.

[22] R. M. Verma, V. Zeng, and H. Faridi, "Data Quality for Security Challenges: Case Studies of Phishing, Malware and Intrusion Detection Datasets," in *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, 2019, pp. 2605–2607.

[23] A. E. Aassal, L. Moraes, S. Baki, A. Das, and R. Verma, "Anti-phishing pilot at ACM IWSPA 2018: Evaluating performance with new metrics for unbalanced datasets," in *Proc. IWSPA-AP Anti Phishing Shared Task Pilot 4th ACM IWSPA*, 2018, pp. 2–10.

[24] N. B. Harikrishnan, R. Vinayakumar, and K. P. Soman, "A machine learning approach towards phishing email detection," in *Proceedings of the Anti-Phishing Pilot at ACM International Workshop on Security and Privacy Analytics (IWSPA AP)*, 2018, vol. 2013, pp. 455–468.

[25] V. Ra, B. G. HBa, A. K. Ma, S. KPa, P. Poornachandran, and A. Verma, "DeepAnti-PhishNet: Applying deep neural networks for phishing email detection," in *Proc. 1st AntiPhishing Shared Pilot 4th ACM Int. Workshop Secur. Privacy Anal.(IWSPA)*, 2018, pp. 1–11.

[26] B. G. HBa, V. Ra, A. K. Ma, and S. KPa, "Distributed Representation using Target Classes: Bag of Tricks for Security and Privacy Analytics."

[27] A. Vazhayil, N. B. Harikrishnan, R. Vinayakumar, K. P. Soman, and A. D. R. Verma, "PED-ML: Phishing email detection using classical machine learning techniques," in *Proc. 1st AntiPhishing Shared Pilot 4th ACM Int. Workshop Secur. Privacy Anal.(IWSPA)*, 2018, pp. 1–8.

[28] N. A. Unnithan, N. B. Harikrishnan, S. Akarsh, R. Vinayakumar, and K. P. Soman, "Machine learning based phishing e-mail detection," *Secur. Amrita*, pp. 65–69, 2018.

[29] C. Coyotes, V. S. Mohan, J. Naveen, R. Vinayakumar, K. P. Soman, and A. D. R. Verma, "ARES: Automatic rogue email spotter," in *Proc. 1st AntiPhishing Shared Pilot 4th ACM Int. Workshop Secur. Privacy Anal.(IWSPA)*, 2018.

[30] M. Nguyen, T. Nguyen, and T. H. Nguyen, "A deep learning model with hierarchical lstms and supervised attention for anti-phishing," *arXiv Prepr. arXiv1805.01554*, 2018.

[31] M. Hiransha, N. A. Unnithan, R. Vinayakumar, K. Soman, and A. D. R. Verma, "Deep learning based phishing e-mail detection," in *Proc. 1st AntiPhishing Shared Pilot 4th ACM Int. Workshop Secur. Privacy Anal.(IWSPA)*, 2018.

[32] N. A. Unnithan, N. B. Harikrishnan, R. Vinayakumar, K. P. Soman, and S. Sundarakrishna, "Detecting phishing E-mail using machine learning techniques," in *Proc. 1st Anti-Phishing Shared Task Pilot 4th ACM IWSPA Co-Located 8th ACM Conf. Data Appl. Secur. Privacy (CODASPY)*, 2018, pp. 51–54.

[33] D. D. Palmer, "Tokenisation and sentence segmentation," *Handb. Nat. Lang. Process.*, pp. 11–35, 2000.

[34] S. G. Bird and E. Loper, "NLTK: the natural language toolkit," 2004.

[35] S. Seifollahi, A. Bagirov, R. Layton, and I. Gondal, "Optimization based clustering algorithms for authorship analysis of phishing emails," *Neural Process. Lett.*, vol. 46, no. 2, pp. 411–425, 2017.

[36] E. Castillo, S. Dhaduvai, P. Liu, K.-S. Thakur, A. Dalton, and T. Strzalkowski, "Email Threat Detection Using Distinct Neural Network Approaches," in *Proceedings for the First International Workshop on Social Threats in Online Conversations: Understanding and Management*, 2020, pp. 48–55.

[37] R. Vinayakumar, K. P. Soman, P. Poornachandran, V. S. Mohan, and A. D. Kumar, "ScaleNet: scalable and hybrid framework for cyber threat situational awareness based on DNS, URL, and email data analysis," *J. Cyber Secur. Mobil.*, pp. 189–240, 2018.

[38] R. Vinayakumar, K. P. Soman, P. Poornachandran, S. Akarsh, and M. Elhoseny, "Deep learning framework for cyber threat situational awareness based on email and url data analysis," in *Cybersecurity and Secure Information Systems*, Springer, 2019, pp. 87–124.

[39] E. S. Gualberto, R. T. De Sousa, P. D. B. Thiago, J. P. C. L. Da Costa, and C. G. Duque, "The Answer is in the Text: Multi-Stage Methods for Phishing Detection based on Feature Engineering," *IEEE Access*, 2020.

[40] R. Amin, M. M. Rahman, and N. Hossain, "A Bangla Spam Email Detection and Datasets Creation Approach based on Machine Learning Algorithms," in *2019 3rd International Conference on Electrical, Computer & Telecommunication Engineering (ICECTE)*, 2019, pp. 169–172.

[41] S. Kaddoura, O. Alfandi, and N. Dahmani, "A Spam Email Detection Mechanism for English Language Text Emails Using Deep Learning Approach," in *2020 IEEE 29th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE)*, 2020, pp. 193–198.

[42] J. Rastenis, S. Ramanauskaitė, I. Suzdalev, K. Tunaitytė, J. Janulevičius, and A. Čenys, "Multi-Language Spam/Phishing Classification by Email Body Text: Toward Automated Security Incident Investigation," *Electronics*, vol. 10, no. 6, p. 668, 2021.

[43] V. Ramanathan and H. Wechsler, "phishGILLNET—phishing detection methodology using probabilistic latent semantic analysis, AdaBoost, and co-training," *EURASIP J. Inf. Secur.*, vol. 2012, no. 1, p. 1, 2012.

[44] F. Janjua, A. Masood, H. Abbas, and I. Rashid, "Handling Insider Threat Through Supervised Machine Learning Techniques," *Procedia Comput. Sci.*, vol. 177, pp. 64–71, 2020.

[45] G. Sonowal, "Phishing Email Detection Based on Binary Search Feature Selection," *SN Comput. Sci.*, vol. 1, no. 4, 2020.

[46] M. MANASWINI and D. R. N. SRINIVASU, "Phishing Email Detection Model using Improved Recurrent Convolutional Neural Networks and Multilevel Vectors," *Ann. Rom. Soc. Cell Biol.*, vol. 25, no. 6, pp. 16674–16681, 2021.

[47] A. Baccouche, S. Ahmed, D. Sierra-Sosa, and A. Elmaghraby, "Malicious Text Identification: Deep Learning from Public Comments and Emails," *Information*, vol. 11, no. 6, p. 312, 2020.

[48] Y. Fang, C. Zhang, C. Huang, L. Liu, and Y. Yang, "Phishing email detection using improved RCNN model with multilevel vectors and attention mechanism," *IEEE Access*, vol. 7, pp. 56329–56340, 2019.

[49] C. Thapa *et al.*, "Evaluation of Federated Learning in Phishing Email Detection."

[50] T. Joachims, “A probabilistic analysis of the Rocchio algorithm with TFIDF for text,” in *Int l Conf on Machine learning (ICML). Ten-7301*, 1997, no. 9.

APPENDICES

Snapshot of English–Arabic Parallel Text Corpus for Phishing Email.

No.	English Text	Arabic Text
1	You have 1 new Important security notification regarding 2017 payroll schedule. View Message Now.	لديك إشعار أمان واحد جديد مهم بخصوص جدول الرواتب لعام 2017. عرض الرسالة الآن
2	You have new messages for your organization account. Continue here now to receive your new messages. If no action is taken in less than 24 hours, all new messages will be permanently deleted on our database Have a great day!	لديك رسائل جديدة لحساب مؤسستك. تواصل هنا الآن لتلقي رسائلك الجديدة. إذا لم يتم اتخاذ أي إجراء في أقل من 24 ساعة ، فسيتم حذف جميع الرسائل الجديدة نهائيًا من قاعدة البيانات الخاصة بنا أتمنى لك يوما عظيما
3	Our record shows that your Mailbox is Out-dated which has caused some incoming mails to be placed on pending. Kindly Click Here to update your Mailbox in order to be able to receive new mails. We apologies for any inconvenience this might cause	يُظهر سجلنا أن صندوق البريد الخاص بك قديم مما تسبب في وضع بعض الرسائل الواردة في الانتظار. يرجى النقر هنا لتحديث صندوق البريد الخاص بك لتتمكن من استقبال رسائل بريد إلكتروني جديدة نحن نعتذر عن أي إزعاج قد يسببه هذا الأمر
4	You have (2) important unread messages, Click on review read it.	لديك (2) رسائل مهمة غير مقروءة ، انقر فوق مراجعة لقراءتها
5	Your mailbox has exceeded the storage limit 1 GB, which is defined by the administrator, you are running at 99.8 gigabytes, you cannot send or receive new messages until you re-validate your mailbox. To renew the mailbox, Click Here	صندوق البريد الخاص بك قد تجاوز حد التخزين 1 جيجا بايت ، والذي تم تحديده من قبل المسؤول ، أنت تعمل على 99.8 جيجا بايت ، لا يمكنك إرسال أو استقبال رسائل جديدة حتى تقوم بإعادة التحقق من صندوق البريد الخاص بك. لتجديد صندوق البريد ، اضغط هنا
6	This organization Account is Subject to mandatory upgrade, Failure to comply would lead to Permanent closure of your account. Upgrade Account Now	يخضع حساب المؤسسة هذا للترقية الإلزامية ، وسيؤدي عدم الامتثال إلى الإغلاق الدائم لحسابك قم بترقية الحساب الآن.
7	To whom it may concern: Please contact your financial institution to get the necessary updates of the Direct Deposit software.	إلي من يهمه الامر يرجى الاتصال بمؤسستك المالية للحصول على التحديثات اللازمة لبرنامج الإيداع المباشر.
8	You have used 98.9% of the total data allocated to your mailbox. To avoid placing your incoming messages on hold or loose them permanently, we require you to re-validate your mailbox to expand your data allocation size.	لقد استخدمت 98.9% من إجمالي البيانات المخصصة لصندوق البريد الخاص بك. لتجنب وضع رسائلك الواردة قيد الانتظار أو فقدانها بشكل دائم ، نطلب منك إعادة التحقق من صحة صندوق البريد الخاص بك لتوسيع حجم تخصيص البيانات الخاص بك.
9	Dear Student, A recent security upgrade has been implement on our servers. All organization users are hereby required to update their account information by following the link below. This update is necessary in order to activate a safety feature on your account. Thank you.	عزيزي الطالب، تم تنفيذ ترقية أمنية حديثة على خوادمنا. يُطلب من جميع مستخدمي المؤسسة بموجب هذا التحديث معلومات الحساب من خلال اتباع الرابط أدناه هذا التحديث ضروري لتفعيل ميزة الأمان في حسابك شكرا لك
10	We are contacting you to remind you that our Account Review Team identified some unusual activity in your organization account. We advise to verify your account to keep it activated, <<link>>	نتصل بك لتذكيرك بأن فريق مراجعة الحساب لدينا قد حدد نشاطًا غير عادي في حساب مؤسستك. ننصحك بالتحقق من حسابك ليظل نشطًا رابط