This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3110285, IEEE Access

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

The Impact of Translating Resource-Rich Datasets to Low-Resource Languages Through Multi-Lingual Text Processing

ABDUL GHAFOOR[1], ALI SHARIQ IMRAN[2] (Member, IEEE), SHER MUHAMMAD DAUDPOTA[1], ZENUN KASTRATI[3], ABDULLAH[1], RAKHI BATRA[1], MUDASIR AHMAD WANI[2]

[1]Dept. of Computer Science, Sukkur IBA University, Pakistan (e-mail: Sher Muhammad Doudpota sher@iba-suk.edu.pk, rakhi@iba-suk.edu.pk)
[2]Dept. of Computer Science, Norwegian University of Science & Techology (NTNU), Norway, (e-mail: ali.imran@ntnu.no, mudasir.a.wani@ntnu.no)
[3]Dept. of Informatics, Linnaeus University, Sweden, (zenun.kastrati@lnu.se)

Corresponding author: Ali Shariq Imran (e-mail:ali.imran@ntnu.no).

ABSTRACT Urdu is still considered a low-resource language despite being ranked as the world's 10^{th} most spoken language with nearly 230 million speakers. The scarcity of benchmark datasets in low-resource languages has led researchers to utilize more ingenious techniques to curb the issue. One such option widely adopted is to use language translation services to replicate existing datasets from resource-rich languages such as English to low-resource languages, such as Urdu. For most natural language processing tasks, including polarity assessment, words translated via Google translator from one language to another often change the meaning. It results in a polarity shift causing the system's performance degradation, particularly for sentiment classification and emotion detection tasks. This study evaluates the effect of translation on the sentiment classification task from a resource-rich language to a low-resource language. It identifies and enlists words causing polarity shift into five distinct categories. It further finds the correlation between the language with similar roots. Our study shows 2-3 percentage points performance degradation due to polarity shift as a result of translation from resource-rich languages to low-resource languages.

INDEX TERMS Multilingual text processing, sentiment classification, polarity assessment, low resource language, language translation, BiLSTM, Conv1D, English, Urdu, German, Hindi

I. INTRODUCTION

British physicist Tim Berners-Lee published the first website in 1991 at CERN lab Switzerland [1] and as of 2019, there were 1.72 billion websites online[1]. Technology adoption, economic opportunities, and domestic pressure are some of the important factors in the spread of the Internet around the globe [2]. People use the internet for social networking [3], entertainment [4], education [5], online shopping [6], and so-on. This rapid increase in the use of the internet is producing loads of data. Leaders, celebrities, athletes, and other individuals use micro-blogging sites to share their stories, events, and opinions (negative, positive, and neutral) about entities. These opinions can be about the quality of product or service whether it is good or not, social events, natural disasters, and so on [7], [8].

[1]https://www.statista.com/chart/19058/how-many-websites-are-there/

Sentiment analysis is almost two decades old research area that primarily focuses to extract the polarity and emotions from the text data. Polarity can be measured as positive, negative, or neutral while emotions are divided into six categories namely: joy, surprise, sad, disgust, fear, and anger. Joy and Surprise emotions are assigned Positive polarity whereas Sad, Disgust, Fear, and Anger are classified as Negative [9]. Sentiment analysis approaches can broadly be categorized into two distinct categorizes: (1) Lexicon-based approach [10] [11] [12] [13] [14] [15] [16] assign the sentiment score to each word of sentence from lexicon corpus. (2) Machine learning approach [17] [18] [19] [9] consist of supervised and unsupervised learning algorithm. Mainly machine learning and deep learning polarity detection algorithms require labeled data to train the model and evaluate the performance of the model. While lexicon-based models do not require the labeled data. They use

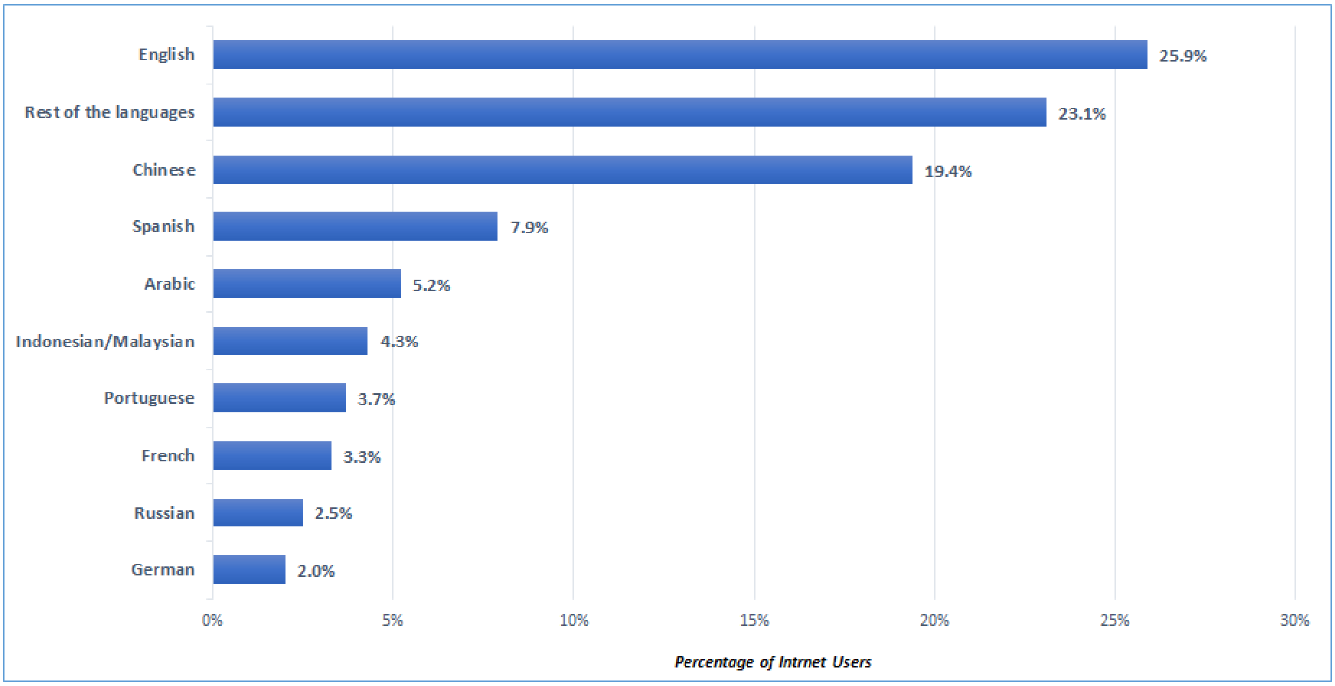

FIGURE 1. Share of Internet users by languages as of January 2020

corpus of opinion words and sentiment score associated with them to predict the sentiment from text [20]. Sentiment analysis is widely used to classify product reviews [21], the sentiment of social media posts [22], peoples' reactions towards different situations like COVID-19 pandemic [23] [24], and so on.

Users on social media have most of their friends and followers from the same community or same country, so, they prefer to communicate in their own language to share their opinions. According to Statista[2], the most common language used on the internet is English. It can be observed from the Figure 1 that English is used by 25.9% of internet users followed by Chinese (19.4%) and Spanish (7.9%), which makes the 53.2% users of the internet. Remaining 46.8% users communicate in other languages such as Arabic, Indonesian, Malaysian, Portuguese, French, Hindi, Urdu, and all the rest. The main problem is that the majority of these non- English languages are resource-poor in the context of machine learning because of the small size of the labeled dataset. This study considering Urdu as the primary low-resource language for experiments through this study. And to the best of our knowledge, currently available Urdu sentiment analysis datasets have a low number of instances, at most 11,000 [25]. The study has explored the machine translation approach to create a large dataset for low-resource languages by simply translating the English dataset into Urdu, German, and Hindi.

Machine Translation is the automated translation of text or speech from one language to another [26] [27]. The lack of labeled data available for low-resource languages have motivated researchers to use multilingual approaches to fill-up the gap between low-resource languages and resource-rich languages. Recently, many corpora are developed for multilingual text processing [26] [28]. Transformer based models have improved the quality of text-to-text machine translation [27] [29] [30], but there is a trade off between the number of languages and efficiency in language model [31]. So, the lower number of languages in a model, the higher the efficiency of the model. There is a need to create a corpus for low-resource languages that require language experts and are time-consuming. The alternate approach is to translate the data set using well-known translators (Google, Bing, Yandex, and DeepL), which support 100 languages, and use it for further processing the machine learning task [32].

This rapid increase in the online content has created the opportunities for researchers to come up with effective approaches to transform this huge data into useful information. Majority of the languages used on the internet are not resource-rich languages like English and Chinese. For Example, for sentiment classification, supervised learning is the most successful approach [33] and requires labeled data to train and evaluate the model. Dataset development process requires data and annotation method to label the data in corresponding class. Annotation can be done by the human expert or also achieved by automatic computer programs. If language-specific text or data labeling resources are not available then authors have suggested the multilingual approach to develop a dataset [26] [28] [34] [35]. This approach uses the machine translation technique

[2]https://www.statista.com/statistics/262946/share-of-the-most-common-languages-on-the-internet

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3110285, IEEE Access

to translate the dataset from resource-rich language to low-resource language. For example in English, there are many labeled datasets available, but in Urdu language, dataset availability is rare [34] [35] [36] [37].

Transformers based models have brought lots of improvement in the quality of translation and now machine translation is considered as a mature field [27]. Although transformer models have achieved significant translation quality for many languages, yet these models have not been completely explored on low-resource languages [38]. Lingual similarities between two languages is also an important factor for the translation quality. For example, the English language lexically is more similar to German, and to Spanish rather than Japanese, that is why translation from English to German and Spanish can yield better quality than Japanese [39]. Many open-source machine translation solutions such as Google translator[3], Bing Translator[4], Yandex[5], and DeepL Translator[6] are available on the Internet and support more than 100 languages spoken all around the globe.

Machine translation approaches include (i) either translate data from resource-rich language (i.e. English) to low-resource language (like Urdu) and train the model on particular low-resource language, or (ii) create a model in resource-rich language then translate the instance from low-resource language to resource-rich language and evaluate it using already created model in resource-rich language [40]. Quite often the models trained on low-resource language text report the low accuracy. For such cases, the researchers in study [28] have suggested to train the model on resource-rich language such as English because learning algorithms understand the English text better than low-resource languages.

A. OBJECTIVES & RESEARCH QUESTIONS

The main objectives of this study are to: (i) explore the translation approach to develop a sentiment analysis dataset for low-resource languages, (ii) to study the effect of translating the English reviews into German, Urdu, and Hindi and compare the classification results of all languages and, (iii) to conduct error analysis to find word categories responsible for polarity shift and performance degradation, if so. The aim is to investigate the Multi-lingual approach and explore the translated German, Urdu, and Hindi text as a case-study to answers the following research questions:

1) How does the dataset translation affect the classifiers performance?
2) What kind of language structures and constructs are important to be paid proper attention while translating dataset from one language to other?
3) Can the translation be an alternative method to developing large-scale datasets for low-resource languages?

[3]https://translate.google.it/
[4]https://www.bing.com/translator
[5]https://translate.yandex.com/
[6]https://www.deepl.com/en/translator

B. CONTRIBUTIONS

The major contributions of this study are listed below:

1) IMDB English movie review dataset translated into German, Urdu, and Hindi using Google translator.
2) English and corresponding three translated datasets trained and validated on machine and deep learning models. Further, the performance of translated datasets compared with original dataset results.
3) Wrongly classified 130 Urdu translated reviews were translated manually and compared with equivalent Google translated reviews. Comprehensive analysis on both machine and human translated reviews were done to identify language structures and constructs shifting the polarity of machine learning translated text.
4) Identified 104 language structures and constructs were labeled into five categories i.e. ambiguous, idiom phrase, negation, sarcasm, and slang.
5) Finally, to empirically establish the fact whether translating a dataset from English to other languages is a right approach or not.

The rest of the article is structured as following. Section II presents a literature review on the topic of multi-lingual analysis. Section III describes the dataset and classification algorithms. Models configuration and performance metrics are presented in Section III-C followed by results and their analysis in Section IV. Error analysis and its causes are discussed in Section V. Lastly, the conclusion is presented in Section VIII.

II. RELATED WORK

Recent developments in the field of NLP are mostly related to deep neural networks which require huge amount of data for training the model. Most of the success in the field of sentiment analysis is through supervised machine learning which requires the availability of labeled datasets. There are many sentiment analysis datasets available in rich-resource languages such as English but in low resource languages such as Urdu, the dataset availability is scarce. Many researchers have used the multi-lingual approach to solve this problem such as by translating the huge English dataset into corresponding low-resource languages. This related work section will focus on the overall multi-lingual approach used to solve the low dataset issue for sentiment analysis.

Improvement in machine translation has attracted researchers to explore the multi-lingual approach for data labeling and sentiment analysis as presented in Figure 2. Kerstin Denecke in his article [34] used the multi-lingual approach to label the German text for sentiment analysis. The author has translated German movie reviews into English and used the SentiWordNet lexicon to assign the polarity score [41] [16]. Alexandra Balahur et al. translated the original English dataset into French, German, and Spanish using three different translators for training data namely: Google, Bing, and Moses. For test data, the authors used

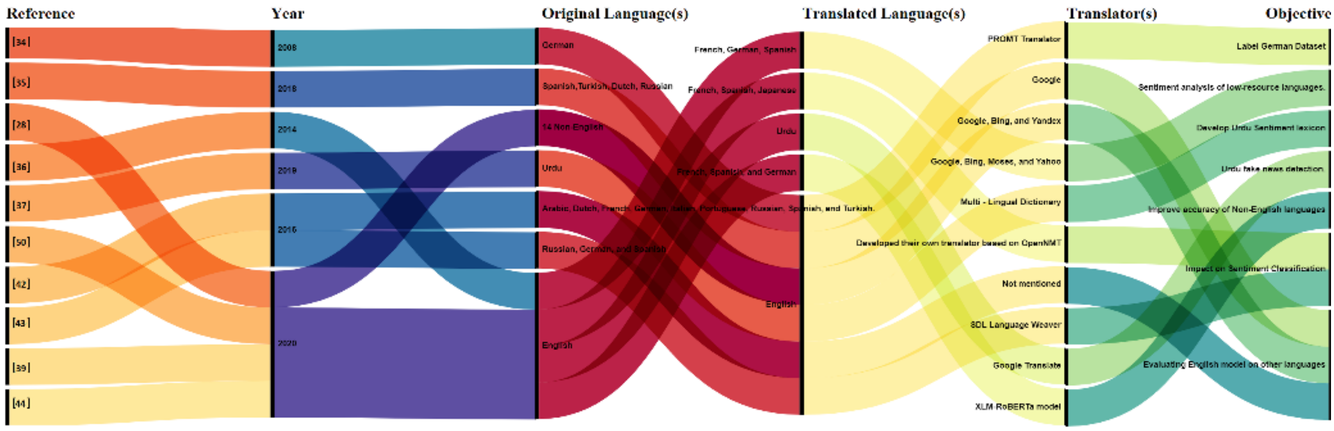

FIGURE 2. Related Work Summary

four translators, same three used for training data and additionally Yahoo translation [36]. Experimental results suggested that translation systems are producing good quality data and the classification performance gap between English and translated data is also less. The study conducted by Arujo et al. has evaluated machine translation in nine different languages [42]. They used 21 models trained on English text and two models developed for non-English text. Experimental results showed that the Sentistrength non-English method was more accurate. Gayane Shalunts et al. conducted an investigation to find the impact of translation on sentiment analysis [43]. Russian, German, and Spanish datasets were translated into English. Experimental results proved that the performance gap remains within 5% range which hints that translation method can be an alternative approach to create corpora for sentiment analysis, though with bit compromise on performance.

Ethem F. Can et al. tried to find out the answer to the question: can a model trained on English sentiment analysis dataset be reused for other languages? where data is limited such as Russian, Spanish, Turkish, and Dutch [35]. Two sets of datasets were used for experiments. (1) Training dataset: Train set consists of three English datasets, namely very large amazon reviews dataset, Yelp restaurant reviews, and competition restaurant reviews datasets. Last two datasets selected to make model learn in a specific domain, i.e., restaurant review. (2) Testing dataset: For evaluation of the multilingual approach, this study used datasets of restaurant review for Russian, Spanish, Turkish, and Dutch languages. RNN architecture with pre-trained word embedding was used to train the model on the English dataset. Experimental results proved that the multi-lingual approach outperforms the baseline. In research paper [28], the authors performed multiple experiments to find the effectiveness of language-specific methods. They evaluated sixteen methods proposed for English and three languages specific methods on fourteen human-labeled datasets. Results suggested that it is better to translate language-specific text into English and use the best model proposed for English than the language-specific method. Alberto Poncelas et al. has discussed the benefits and drawbacks of classifying translated sentences [39]. They used four languages for the experiment: English, French, Spanish, and Japanese. Paracrawl and JParaCrawl corpuses used for experiments and results proved that translation from English to French and Spanish was of better quality than English to Japanese. The reason is French and Spanish are grammatically and lexically closer to English than Japanese. Another important outcome of the study was sentiment classifier performed better on original data than translated.

Valentin Barriere et al. has proposed the multilingual transformer model and automatic translation approach to resolving the problems of sentiment analysis dataset for non-English tweets [44]. They used an English dataset to train the model and translated it into French, Spanish, German, and Italian. These translated datasets merged with a small corresponding language dataset to built a huge dataset for non-English languages. Experimental results proved that the merged dataset (English Translated and Original) produced better performance than small original corpora. For more detail on multi-lingual sentiment analysis, the readers are advised to read the paper by Siaw Ling Lo et al. [45]. This study has discussed both formal, informal, and low-resource languages for sentiment analysis.

Recently, many researchers have worked on the Urdu sentiment analysis. These studies faced the common problem of the small size of a dataset, the datasets are restricted to a few thousand instances only. The current deep learning algorithms, which have outperformed traditional machine learning algorithms required a huge amount of data. The researchers have proposed several data labeling approaches to assign polarity to Urdu text. The majority of authors have used a human-annotated approach [46] [47] [48] [49] for this task, however, few studies have also explored the multi-lingual approach [37] [50] [51].

Asghar et al. in their paper have used the multi-lingual approach to develop a lexicon based dataset for Urdu sentiment analysis [37]. They extracted the adjective from

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3110285, IEEE Access

the Urdu text using Urdu POS Tagging, then translated Urdu adjectives into English using multi-lingual Urdu to English dictionary. The SentiWordNet lexicon was used to get a sentiment score for translated English adjectives [41] [16]. Maaz Amjad et al. investigated that an English to Urdu translated dataset will be useful to train the model to classify Urdu fake news [50]. Experimental results suggested that the current state of English to Urdu translation did improve the performance of fake news classification for the Urdu language. The research study [51] has proposed the structural correspondence learning method for Urdu sentiment analysis. The study used the IIIT POS Hindi dataset which was already in Latin script format. Hindi dataset has many pure Sanskrit word which need be replaced by Urdu, this replacement is done using online dictionaries. Many authors have work on Urdu sentiment analysis.

Thakkar et at. in study [52] have proposed the cross-lingual zero-shot and few-shot learning model to classify Croatian news articles as negative, positive, and neutral using the Slovene language dataset. Further, they train the Bert-based model with 3 languages namely, English, Slovene, and Croatian. The authors proposed single task and multiple task models, experiment results concluded that the multiple task model outperformed the single-task model. The research study [53] has proposed the Dual-trained lazy CNN model for sentiment analysis in Slavic languages. Neural networks perform better on the big amount of data so they require lots of computational power and to handle this issue the study proposed a lazy NN model.

This study has done a comprehensive error analysis to find reasons why performance degradation has been reported in translated low-resource languages such as Urdu and Hindi and identify, what kind of language structures and constructs shifted the polarity of translated reviews which caused accuracy drop in low-resource languages. Further, the terms that changed the sentiment polarity in translated text are categorized as Ambiguous words, Idiom and Phrase, Negation, Sarcasm, and Slang. To the best of our knowledge, studies reported in the literature have not done this kind of detailed analysis to discuss the causes of performance reduction in translated low-resource languages and what kind of words are shifting the polarity in low-resource languages.

III. METHODOLOGY

Figure 3 illustrates the methodology of multi-lingual text processing. The IMDB English movie review dataset translated into German, Urdu and Hindi using Google machine translation API. All four datasets are split into train and test using scikit-learn [7]. Cross validation and simple test-train split is performed where train set is used to train the model and test set to validate the model. Different machine learning and deep learning models are used to train and validate the model and the experimental results are shown in Table 3 and 2. Furthermore, 130 incorrectly predicted Urdu translated reviews, are translated manually to identify the language structures and constructs shifting the polarity of machine learning translated text. In the end, both manually translated and machine-translated reviews are compared to identify language structures and constructs which are categorized as Ambiguous words, Idiom and Phrase, Negation, Sarcasm, and Slang.

A. DATASET

This study has used the IMDB English movie review dataset [54] to investigate the effectiveness of the multi-lingual dataset development approach for German, Urdu, and Hindi. This dataset is well known to the research community and contain both long and short reviews, which are very important to explore the performance of Google translator on different size of the text. Dataset is download from the official webpage and consists of 50 thousand movie reviews, 25 thousand for each, negative and positive class.

B. TRANSLATION

Google-trans-new [8], a google translation API, is used to translate the review dataset from English to German, Urdu, and Hindi. Figure 4 shows a sample review in English and the corresponding translation in German, Urdu, and Hindi. API takes two parameters as input. First input takes text from source language and second input takes code for language to be translated, also called target language as shown in Equation 1. Google translation language codes are available on web page [9]. For Urdu code is "ur", for German code is "de" and for Hindi code is "hi".

$$translate(text, lang\_tgt = code) \tag{1}$$

Google announced the launch of Google Translate in April 2006 based on the Phrase-Based Machine (PBMT) Translation algorithm. And in September 2016 Google announced that Google translate is switching to a new translation system called Google Neural Machine Translation system (NMT) [10]. PBMT method breaks the complete sentences into words and phrases and these terms are translated independently. Whereas NMT learns a mapping between input language (sentence in input language) and output language (equivalent sentences in output language) [55]. Google NMT model contains the deep LSTM network with 8 encoder and 8 decoder layers using residual connections as well as attention connections from the decoder network to the encoder. NMT systems initially showed the same performance as PBMT on publicly available benchmark datasets. Since then, researchers have worked to improve NMT, includes the study on handling rare words [56] and align input and output words using attention [57].

[7]https://scikit-learn.org/stable/index.html

[8]https://pypi.org/project/google-trans-new/

[9]https://cloud.google.com/translate/docs/languages

[10]https://ai.googleblog.com/2016/09/a-neural-network-for-machine.html

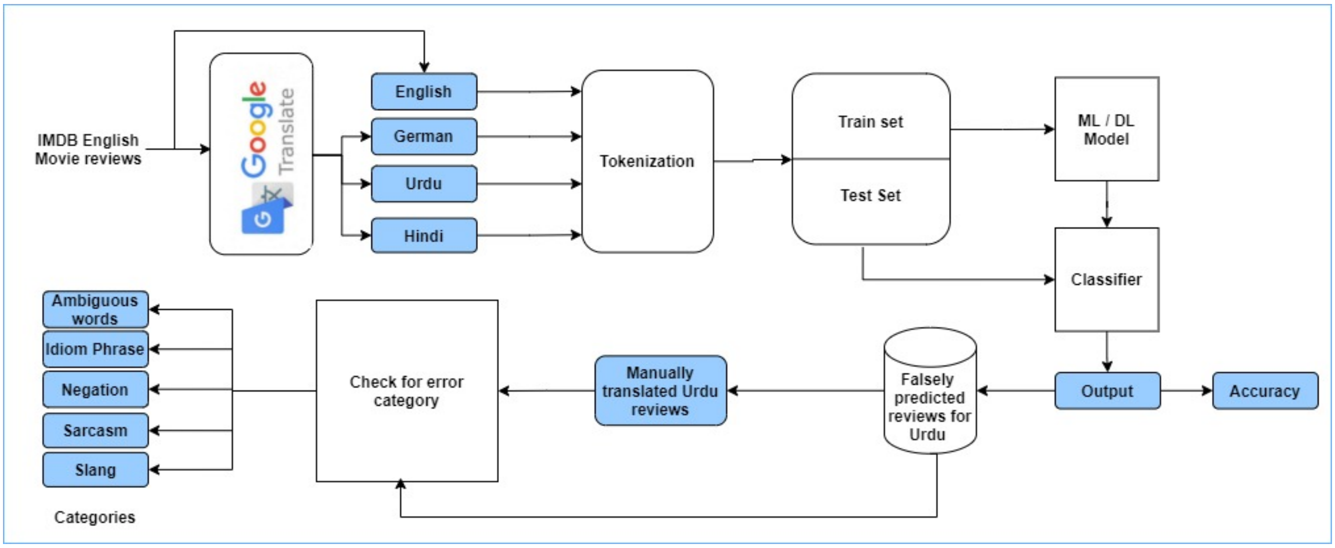

FIGURE 3. Methodology: Multi-Lingual Text Processing

The previous literature [36] [35] [37] in multi-lingual text processing suggest that Google translator is the most popular machine translator and trusted by most of the researchers as illustrated in Figure 2. Due to this reason, Google translator is used in this study to create a hypothesis whether the translation is a good solution to develop a dataset for low resource language or not?

To the best of our knowledge, we have not found any official document that specifically claims, whether Google translator API uses pivot language or not. But some online sources [11], [12] claims that that Google API uses pivot language. Web links also have mentioned that Google translator uses English as an intermediate language when translating two non-English languages for example French to Russian. In our opinion, Google uses the pivot language when translation is done between two non-English but if English is the source or target language then there is a low chance of pivot language, and this study has translated English into other languages and no translation has been done between the two non-English languages.

C. EXPERIMENTAL SETUP

This section presents the model configurations and evaluation metrics. Six models in total, including two conventional machine learning models (Naive Bayes, SVM) and four deep learning models (DNN, LSTM, Bi-LSTM, Conv1D) are employed to train and test the classification accuracy on original and translated reviews. Initially, random parameters were selected for all deep learning models, and parameters were updated to observe the change in the results, and this technique was used till we achieved the best results. The first experiment set was performed on 5 epochs and, we set the size of 16 dimensions for embedding and hidden

[11]https://www.teachyoubackwards.com/extras/pivot/
[12]https://en.wikipedia.org/wiki/Google$_Translate$

Language	Text
English	The distribution was good, the subject could have been interessant and comic. whereas, he described the wandering of an old non credible communist looking for loving sensations. Instead of this, the atmosphere is nor lively nor heavy.
German	Die Verteilung war gut, das Thema hätte interessant und komisch sein können. Er beschrieb die Wanderung eines alten, nicht glaubwürdigen Kommunisten, der nach liebevollen Empfindungen suchte. Stattdessen ist die Atmosphäre weder lebhaft noch schwer.
Urdu	ڈسٹری بیوشن اچھی تھی ، اس موضوع کو انٹر ایکٹنٹ اور مزاحیہ کیا جاسکتا تھا. جبکہ ، انہوں نے ایک پرانے غیر معتبر کمیونسٹ کے بھٹکتے ہوئے محبت انگیز جذبات کی تلاش میں بتایا. اس کے بجائے ، ماحول نہ جیونت ہے اور نہ ہی بھاری ہے.
Hindi	वितरण अच्छा था, इस विषय में निरंतर और हास्य हो सकता था। जबकि, उन्होंने एक पुराने गैर-विश्वसनीय कम्युनिस्ट को भटकने का वर्णन किया जो प्रेमपूर्ण संवेदनाओं की तलाश में था। इसके बजाय, वातावरण न तो जीवंत है और न ही भारी है।

FIGURE 4. English and Translated German, Urdu, and Hindi review

layers, and 2 for the output layer. Then performed many sets of experiments on different sets of parameters and hyper-parameters considering previous results. Experiments observations reported best results on parameter reported in Table on 10 epochs But 10 epochs only improvement reported on training data but validation accuracy remained the same and this created the problem of model over-fitting. Model over-fitting was avoided by halting the model training after 10 epochs. Another Hidden layer was also added but performances remained so the second hidden layer was removed from the final model setting because it was not improving the performance but was increasing computation time. The same approach was also used for machine learning models, many experiments were performed on different hyper-parameters: SVM (C, kernel, degree, gamma) and Naive Bayes (alpha).

Table 1 presents the parameters setting of different algorithms used in this study based upon the empirical analysis.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3110285, IEEE Access

Model Name	Model Configurations / Parameters
DNN [28]	Embedding Layer with 64 Dimension, Dense Layer with 32 Dimension + ReLU, Dense Layer with 2 Dimension + Sigmoid
LSTM [58]	Embedding Layer with 64 Dimension, LSTM Layer with 32 Dimension + Dropout + Recurrent Dropout = 0.2, Dense Layer with 2 Dimension + Sigmoid
Bi-LSTM [59]	Embedding Layer with 64 Dimension, Bi-LSTM Layer with 32 Dimension + Dropout + Recurrent Dropout = 0.2, Dense Layer with 2 Dimension + Sigmoid
Conv1D [60]	Embedding Layer with 64 Dimension, Conv1D Layer with 32 Dimension + ReLU + Global Max Pooling 1D Layer, Dense Layer with 2 Dimension + Sigmoid
SVM [61]	C=1.0, kernel='linear', degree=3, gamma='auto'
Naive Bayes [62]	MultinomialNB with alpha=0.19

TABLE 1. Deep / Machine Learning Models Configuration

D. EVALUATION METRICS

Text classification tasks mostly compute accuracy, F1-score, precision, recall to evaluate performance of the model. These metrics are derived from the confusion matrix [63] which is composed of the true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) values as shown in the Figure 5.

	Predicted No	Predicted Yes
Actual No	TN	FP
Actual Yes	FN	TP

FIGURE 5. Confusion Matrix

Accuracy - Percentage of correctly predicted instances from total instances.

$$Accuracy = \frac{(\mathrm{TP} + \mathrm{TN})}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}} \tag{2}$$

Precision - Precision is the percentage of correctly classified samples for the particular class out of all predicted labels for that class.

$$Precision = \frac{\mathrm{TP}}{(\mathrm{TP} + \mathrm{FP})} \tag{3}$$

Recall - The recall is the percentage of all predicted samples for the particular class relation with actual labels for that class.

$$Recall = \frac{\mathrm{TP}}{(\mathrm{TP} + \mathrm{FN})} \tag{4}$$

F1-score - F1 score is a combination of both precision and recall, it can be interpreted as the harmonic mean of precision and recall.

$$F1 - Score = \frac{2 \times (\mathrm{Precision} \times \mathrm{Recall})}{(\mathrm{Precision} + \mathrm{Recall})} \tag{5}$$

IV. RESULTS

In this study we have applied multiple deep learning and machine learning models on the selected dataset and on its translated versions i.e. German, Urdu, and Hindi. To evaluate the performance of models we have used accuracy, precision, recall and F1 score measures. Simple train-test split and 5-fold cross-validation used to validate all models as shown in Table 2 and 3. It can be inferred from experimental results that the SVM has performed better than other models on English, German and Urdu datasets using both validation approaches but for the Hindi language deep learning models performed better than machine learning algorithms.

Moreover, it is observed that the accuracy of each model on English and German dataset is nearly the same with both train-test and cross-validation. As illustrated in Table 3, the average accuracy of all models is nearly the same for the English and German languages. The highest difference of 1.14% reported the DNN model. SVM performed better than models for both languages with an accuracy of 90.06% for English and 89.92% for the German language. Results of simple train-test split showed same trend maximum difference of 1.43% observed in DNN model and like cross validation SVM was best performing algorithm for both languages with accuracy 90.45% and 90.01% respectively for English and German as shown in the Table 3. According to [64], English language belongs to German language family and English and German have similar lexical structure. This fact is one possible reason for the similarity in accuracy. On the other hand Urdu and Hindi are very much different from English, which is also reflected by difference in accuracy of all the models on these two datasets. We can assume that the translated Urdu and Hindi text doesn't reflect the true semantics of original text.

Overall, both validation approaches produce nearly same results. SVM was performing better for English, German and Urdu while Hindi better performed on Deep learning algorithms. There was a slight difference in terms of accuracy for each language for all models.

Further, the class-wise performance of each language is calculated as shown in Table 4. The best performing model in terms of accuracy from the Table 3 for every language was selected to calculate the class-wise performance. Experimental results concluded that both German and English have same recall scores 91.00% and 89.00% for Positive and Negative classes, respectively. Also, same precision score 91.00% for the negative class was observed but a slight difference was noticed in the precision of positive class 90.00% and 89.00% for English and German correspondingly. Similarly, major difference between the English dataset and its translated Urdu and Hindi version in terms of recall and precision is observed. Both classes for Urdu and Hindi generated less recall and precision than English dataset and this insight proved that translation affects all classes of low-resource languages.

Finally, both precision and recall were combined to

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3110285, IEEE Access

Cross Validation Test Sets	DNN				LSTM				Bi-LSTM			
	English	German	Urdu	Hindi	English	German	Urdu	Hindi	English	German	Urdu	Hindi
1	88.58%	87.07%	85.42%	85.42%	87.46%	86.71%	82.36%	82.36%	87.83%	81.81%	85.92%	85.92%
2	87.84%	87.43%	86.20%	86.20%	87.36%	86.82%	86.23%	86.23%	87.58%	85.33%	85.36%	85.36%
3	87.82%	85.41%	86.12%	86.12%	87.41%	86.85%	86.00%	86.00%	86.28%	86.56%	85.09%	85.09%
4	87.69%	87.42%	84.84%	84.84%	86.96%	86.91%	85.29%	85.29%	87.79%	87.30%	85.70%	85.70%
5	87.96%	86.85%	83.73%	83.73%	88.06%	86.13%	85.84%	85.84%	87.29%	87.06%	85.37%	85.37%
Average	**87.98%**	**86.84%**	**85.26%**	**85.26%**	**87.45%**	**86.68%**	**85.14%**	**85.14%**	**87.35%**	**85.61%**	**85.49%**	**85.49%**
	Conv1-D				**SVM**				**Naive Bayes**			
	English	German	Urdu	Hindi	English	German	Urdu	Hindi	English	German	Urdu	Hindi
1	88.77%	87.49%	86.43%	86.43%	90.22%	89.99%	87.65%	82.43%	86.23%	86.73%	83.15%	78.28%
2	87.85%	87.92%	86.18%	86.18%	89.76%	90.09%	87.71%	82.73%	86.72%	87.12%	83.19%	79.35%
3	88.36%	87.31%	86.44%	86.44%	90.09%	89.71%	87.82%	81.87%	86.49%	86.93%	83.96%	78.40%
4	87.88%	87.39%	86.03%	86.03%	89.85%	89.91%	87.83%	82.56%	85.99%	86.24%	83.13%	78.76%
5	88.20%	87.71%	86.48%	86.48%	90.38%	89.88%	87.75%	85.65%	86.63%	86.60%	83.50%	78.93%
Average	**88.21%**	**87.56%**	**86.31%**	**86.31%**	**90.06%**	**89.92%**	**87.75%**	**83.05%**	**86.41%**	**86.72%**	**83.39%**	**78.74%**

TABLE 2. Deep and Machine Learning Algorithms (Accuracy): 5-Fold Cross Validation

Model	Original language	Translated languages		
	English	German	Urdu	Hindi
DNN	88.37%	86.94%	81.62%	85.56%
LSTM	87.82%	86.85%	81.45%	85.62%
Bi-LSTM	87.76%	87.68%	80.59%	**85.99%**
Conv1-D	88.29%	87.66%	80.78%	85.83%
SVM	**90.45%**	**90.01%**	**87.26%**	82.30%
Naive Bayes	86.32%	86.75%	82.97%	78.13%

TABLE 3. Deep/Machine Learning Algorithms (Accuracy)

Model	Language	Precision		Recall	
		Positive	Negative	Positive	Negative
SVM	English	90.00%	91.00%	91.00%	89.00%
SVM	German	89.00%	91.00%	91.00%	89.00%
SVM	Urdu	87.00%	88.00%	88.00%	86.00%
Bi-LSTM	Hindi	84.00%	89.00%	90.00%	82.00%

TABLE 4. Best performing model for each language (Precision and Recall)

calculate the F1-Score as can be depicted from Figure 6. The figure clearly shows that English and German have an equal F1-score 90% the but Urdu 87% and Hindi 86% much below than performance of the English dataset.

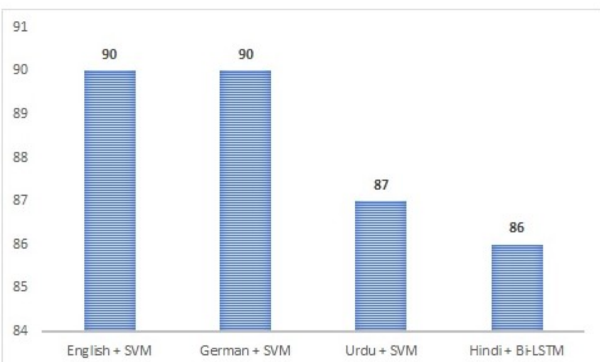

FIGURE 6. Best performing model for each language (F1 Score)

	Positive (Predicted)	Negative (Predicted)
Positive (Actual)	4600	428
Negative (Actual)	527	4445

TABLE 5. Confusion Matrix of IMDB English Dataset

	Positive (Predicted)	Negative (Predicted)
Positive (Actual)	4437	591
Negative (Actual)	683	4289

TABLE 6. Confusion Matrix of IMDB Urdu translated Dataset

V. ERROR ANALYSIS AND DISCUSSION

To find the causes why translated Urdu dataset is producing less accuracy than English, this sub-section of study reports error analysis on English and Urdu languages.

Error analysis started with confusion matrix for both English shown in Table 5 and Urdu dataset presented in Table 6. Confusion matrix shows patterns that translation is affecting both classes. In English dataset, the number of correct predicted instances were 4600 for Positive class, which were decreased in Urdu dataset to 4437. Same is the case with Negative class, in original dataset correctly predicted samples were 4445 which decreased to 4289 in Urdu dataset. .

After confusion matrix, we identified 656 sample which were correctly classified by the model on English dataset but wrongly predicted for the translated Urdu dataset. To find the cause, 130 samples out of 656 were translated manually in Urdu. SVM trained classifier is used to evaluate manually translated 130 samples and model classified 86 samples correctly. Further, these 86 correctly classified samples were compared with equivalent Google translated samples to identify the errors done by Google translator. We identified 104 words (or group of words) in Google

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3110285, IEEE Access

FIGURE 7. Word cloud of 104 words causing the change in polarity of translated Urdu text

Category	Total Terms	Examples
Ambiguous words	73	substance, Shot,hang, sick, right wing, silly movie.
Negation	11	not diminish, faultless, noes, washed out.
Slang	9	daym, holy shit, geeze, cool, ass off.
Idiom and Phrase	8	evil spoons, deadly dull, add insult to injury, ripoff of Blade, Let-down.
Sarcasm	3	unbelievable, Extremely dull entertainment at its best, wow.

TABLE 7. Categories of errors causing change in polarity in translated text

translated Urdu reviews which were changing the sentiment of movie review, words are shown in the Figure 7. We identified and categorized these 104 words into 5 categories namely: *Ambiguous words*, *Idiom and Phrase*, *Negation*, *Sarcasm*, and *Slang* as shown in Table 7.

Ambiguous words: This category contains list of words which have more than one interpretation. These words are shown in Table 7. For example, word hang can be used to kill any one and same word hang can be used for wait. It is necessary to understand the context of text to get exact meaning of ambiguous words. Google translator also face the ambiguous word problem as reported in Table 7, 73 out of 104 are ambiguous word. Nevertheless, Google Translator has been widely used for English to many languages translation including Urdu. However, Google Translator more often translates the sentences word by word and does not take context of a language or sentence in consideration. It can be depicted from Figure 8, that the sentence translated word by word, for instances, the word 'shot' translated as 'shot a fire', whereas, here the word 'shot' potentially describing the 'movie's shot of a clip or frame'. It is evident that Google Translator only translate the English word into Urdu using dictionary based approach and fails to capture context of language. Thus, the reliability of Google Translator based translations for most NLP tasks in Urdu specifically for sentiment analysis is very low where context of language has to preserved.

Negation: In sentiment analysis, it is very important to handle words affected by negation. Negation occurs in many forms such as (1) explicit negation (not, no, etc) that reverses the meaning of a word. Good is a positive word, but if we add not in front of good (not good) then words meaning will be opposite. (2) Implicit negation weakens the polarity of other words, for example words bachelor and spinster, both have the same meaning but both contain different sentiment when using them with Male and Female gender. Bachelor is positive for male and negative for the female. Use of spinster with female taken as the positive but negative with the male [65]. Morphological negations have two variants: prefix (un-, non- and etc.) and suffix (-less) [66].

Experimental results have proven that negation also affects the translation from English to Urdu. We found 11 negation problems in translated text as shown in Table 7. The sentence 'that did not diminish my enjoyment of the movie ' contains the positive polarity, but the translator wrongly translated the' not diminish' that cause the change polarity for the whole movie review. It can be depicted from Figure 8 that Google translator translated 'Faultless production' into Urdu which means 'bad production' which is incorrect. It is another proof that negation affects translation.

English	Google translated	Manual translated
Shot by shot, edit by edit, the film unfolds itself around a disturbed boxer discovering his own purpose (or lack thereof)	شاٹ کے ذریعہ گولی مار دی گئی ، ترمیم کے ذریعہ ترمیم کرکے ، فلم اپنے آپ کو (یا اس کی کمی) اپنے ایک مقصد کو دریافت کرنے والے پریشان باکسر کے گرد پھیلتی ہے	عکس بہ عکس، لفظ بہ لفظ یہ فلم ایک پریشان حال باکسر کے گرد گھومتی دکھائی دیتی ہے جو اپنے مقصد کی تلاش میں ہے
Technically somewhat of a mess and boasting a stock of amateur New Yawk types, this film never bores. I highly recommend tracking this down	تکنیکی طور پر کسی حد تک گڑبڑ اور شوقیہ نیو یاق اقسام کے ذخیرے پر فخر کرنے سے یہ فلم کبھی بور نہیں ہوتی ہے- میں اس سے باخبر رہنے کی انتہائی سفارش کرتا ہوں	تکنیکی طور پر نیو یاق اقسام کے ذخیرہ پر فخر کرتے ہوئے ، یہ فلم کبھی بور نہیں کرے گی- میں اس سے باخبر رہنے کی انتہائی سفارش کرتا ہوں- یہ ایک بہت ہی مزاحیہ ہے
Faultless production values round off a never to be forgotten movie experience	ناقص پیداواری قدروں کا مقابلہ فلمی تجربہ کو کبھی بھی فراموش نہیں کیا جائے گا	کامل پروڈکشن اس فلمی تجربہ کو کبھی بھی فراموش نہیں کیا جائے گا.
Extremely dull entertainment at its best	انتہائی سست تفریحی بہترین کام	انتہائی بدترین تفریح اس میں تھی
holy Sh*t this was god awful	یہ خدا کا خوفناک تھا	ارے یہ بہت بیکار تھی ہے

FIGURE 8. Sample of English and Urdu translated text

Slang: People use slang such as ' daym, ass off, oh em gee, etc' to express their feelings. These term do not exist in the dictionaries. Slang can be a new word or misrepresentation of an existing word. Google Translator translates sentence word by word, so if a word is not available in Google Translator dictionary then the translator is not able to translate the word properly and it affects the polarity of the whole text. We identified 9 slang that changed the polarity of translated Urdu text sees Table 7. Slang 'holy sh*t', see Figure 8, is mostly used in an unpleasant situation, so it contains the negative polarity. But Google Translator considers both words holy and sh*t separately, if we look at holy as a single word it contains the positive polarity and due to cause the machine learning model classified Urdu text

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3110285, IEEE Access

as a positive review. We manually corrected this translation mistake and evaluate the review again and this time model classified the review correctly as negative.

Idiom and Phrase: An idiom is a saying (the type of phrase) that means different from its literal meaning. For example idiom 'It's raining cats and dogs' means 'It's raining hard'. But when you translate this same idiom into Urdu language using Google Translator, equivalent Urdu translation looks, the sentence is about a rain of cats and dogs animals and does not about heavy rain. We identified 8 idiom and phrase related issues in Urdu translated text as shown in Table 7. Idiom 'hoots and a half' which means very funny but Google Translator translate the text word by word, so it does not capture the context of idiom and translated text means different in Urdu that causes the change in the polarity of translated Urdu movie review as stated in the Figure 8.

Sarcasm: Sarcasm in sentiment analysis is when people use positive words to taunt and express negative feelings towards entities such as individuals, topics, products, services, events, and issues. Sarcasm can be easily detected in voice through voice tone but in textual data, it is a serious research problem [67]. Lots of scholars have contributed to English but in the Urdu language, it still requires the researcher's attention. This study has investigated that sarcasm also occurs in the translated Urdu text as shown in Table 7. It can be depicted from Figure 8 the example of sarcasm, 'best' is a positive word but in a sentence, this word is used to criticize the movie. Machine learning classifier detected these issues in English text but for the translated Urdu text model did not identify the sarcasm and classified the movie as a positive.

VI. FINDINGS CONCERNING RQS

This study has addressed all RQs by proposing the machine and deep learning models, and detailed analysis of experiment results. For RQ1, performance degradation is reported in translated low-resource languages such as Urdu and Hindi, however performance drop in translated medium-resource language like German is very low as presented in Table 3. Concerning RQ2, 104 lexeme (comprising of one word or few words) were identified and categorized as Ambiguous words, Idiom and Phrase, Negation, Sarcasm, and Slang, as illustrated in Table 7. These terms shifted the polarity of Urdu reviews and caused the incorrect prediction of Urdu reviews which leads to performance degradation in the Urdu language, therefore it is important to pay attention to these five categories of words while employing translation as a method to produce dataset for low-resource languages. Concerning RQ3, experimental findings suggest that translation is not a good way to develop a large-scale dataset for low-resource languages because a 2-3% decrease in performance from English to low-resource is not negligible. The translation approach can be an alternative if the target language is medium-resource, i.e. German.

VII. RECOMMENDATIONS

One of the problems in multilingual text processing is unavailability of datasets of low-resource languages, therefore, we used Google translator to translate the English text into Urdu and Hindi for creating a multilingual text processing model but results suggest that the translation mechanism is not an appropriate way to process multilingual text. Based on our observations we have come up with following recommendations:

- Gather data in low-resource languages from online sources like Facebook, Twitter, news websites etc.
- Build machine learning models on low-resource languages directly rather than creating models on resource rich language to process multilingual text.
- Creating a resource like SentiWordNet for low-resource languages to improve the accuracy of multi-lingual text processing models.
- Explore suitable pre-processing toolkit for low-resource languages.
- Languages that share the same lexical structure with English can be translated for multi-lingual text processing.

VIII. CONCLUSION

Using internet has become a routine for almost everyone specially for online shopping and using social networking applications like Facebook, Twitter, etc. The wide range of features offered by applications have increased their usability; one such feature is multilingual support which means people can now post their queries or write anything on internet in their own language. Increase in use of internet has increased the volume of data and also the need to process this data for meaningful insights but there is very limited research that addresses how to process the multilingual text accurately. Considering the need and importance of multilingual text processing, the aim of this study is to develop a machine learning model for a resource-rich language like English and use that model to process the translated text from low-resource languages. Also, we studied the effect of translating the text from resource-rich language to low-resource language by comparing the classification results for all languages and conducting error analysis to find word categories responsible for polarity shift and performance degradation, if any.

To achieve this objective, we designed a case study of IMDB movie review dataset. The dataset is translated into German, Urdu and Hindi using Google translator. Original and translated dataset are used to train and test the four deep learning models, namely DNN, LSTM, Bi-LSTM and Conv1D and two traditional machine learning models i.e. SVM and Naive Bayes. The performance of these models is evaluated by calculating the accuracy, precision, recall and F1-score. Results suggest that the accuracy of SVM on English is 90.45% and German dataset is 90.01%, which is nearly same and it performs better than other five models on both datasets with the F1-score of 90%. SVM has also proved to be best model for Urdu dataset with an accuracy of 87.26% while Bi-LSTM seems to be best model for Hindi dataset with an accuracy of 85.99%. Literature suggests that the English and German languages share the same lexical structure, therefore, we can assume that the language translation can work to design machine learning models for low resource languages that are similar to resource rich languages.

Furthermore, to investigate the performance degradation on Urdu and Hindi dataset, we have performed error analysis on Urdu dataset. 86 incorrectly classified reviews were manually translated into Urdu. The comparison of Google translated review and corresponding manually translated review identified 104 words that shifted the polarity of review. Those words fall into five categories i.e. Ambiguous words, Idiom and Phrase, Negation, Sarcasm, and Slang. We, therefore, experimentally establish the fact that translation is not the way to develop datasets for low-resource languages such as Urdu and Hindi.

In the future work, researchers can explore other translators such as Microsoft Bing or can develop their own translator and compare the results with Google translator. It would be interesting to see how lexicon-based sentiment analysis and transformer based models can perform on auto-translated text.

REFERENCES

[1] S. S. McPherson, Tim Berners-Lee: Inventor of the World Wide Web. Twenty-First Century Books, 2009.
[2] H. V. Milner, "The global spread of the internet: The role of international diffusion pressures in technology adoption," in 2nd Conference on Interdependence, Diffusion, and Sovereignty, UCLA, California, 2003.
[3] A. Lenhart, K. Purcell, A. Smith, and K. Zickuhr, "Social media & mobile internet use among teens and young adults. millennials." Pew internet & American life project, 2010.
[4] L. Leung, "Stressful life events, motives for internet use, and social support among digital kids," CyberPsychology & Behavior, vol. 10, no. 2, pp. 204–214, 2006.
[5] S. Livingstone and M. Bober, "Taking up online opportunities? children's uses of the internet for education, communication and participation," E-Learning and Digital Media, vol. 1, no. 3, pp. 395–419, 2004.
[6] S. Park and D. Lee, "An empirical study on consumer online shopping channel choice behavior in omni-channel environment," Telematics and Informatics, vol. 34, no. 8, pp. 1398–1407, 2017.
[7] L. Yue, W. Chen, X. Li, W. Zuo, and M. Yin, "A survey of sentiment analysis in social media," Knowledge and Information Systems, vol. 60, no. 2, pp. 617–663, 2019.
[8] Z. Kastrati, F. Dalipi, A. S. Imran, K. Pireva Nuci, and M. A. Wani, "Sentiment analysis of students' feedback with NLP and deep learning: A systematic mapping study," Applied Sciences, vol. 11, no. 9, p. 3986, 2021.
[9] A. S. Imran, S. M. Daudpota, Z. Kastrati, and R. Batra, "Cross-cultural polarity and emotion detection using sentiment analysis and deep learning on covid-19 related tweets," IEEE Access, vol. 8, pp. 181 074–181 090, 2020.
[10] A. Esuli and F. Sebastiani, "Determining term subjectivity and term orientation for opinion mining," in 11th Conference of the European chapter of the association for computational linguistics, 2006.
[11] G. A. Miller, "Wordnet: a lexical database for english," Communications of the ACM, vol. 38, no. 11, pp. 39–41, 1995.
[12] S. Mohammad, C. Dunne, and B. Dorr, "Generating high-coverage semantic orientation lexicons from overtly marked words and a thesaurus," in Proceedings of the 2009 conference on empirical methods in natural language processing, 2009, pp. 599–608.
[13] A. Andreevskaia and S. Bergler, "Mining wordnet for a fuzzy sentiment: Sentiment tag extraction from wordnet glosses," in 11th conference of the European chapter of the Association for Computational Linguistics, 2006.
[14] A. Esuli and F. Sebastiani, "Determining the semantic orientation of terms through gloss classification," in Proceedings of the 14th ACM international conference on Information and knowledge management, 2005, pp. 617–624.
[15] X. Ding, B. Liu, and P. S. Yu, "A holistic lexicon-based approach to opinion mining," in Proceedings of the 2008 international conference on web search and data mining, 2008, pp. 231–240.
[16] A. Esuli and F. Sebastiani, "Sentiwordnet: A publicly available lexical resource for opinion mining." in LREC, vol. 6. Citeseer, 2006, pp. 417–422.
[17] K. Dave, S. Lawrence, and D. M. Pennock, "Mining the peanut gallery: Opinion extraction and semantic classification of product reviews," in Proceedings of the 12th international conference on World Wide Web, 2003, pp. 519–528.
[18] G. Paltoglou and M. Thelwall, "A study of information retrieval weighting schemes for sentiment analysis," in Proceedings of the 48th annual meeting of the association for computational linguistics, 2010, pp. 1386–1395.
[19] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? sentiment classification using machine learning techniques," in Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002). Association for Computational Linguistics, 2002, pp. 79–86.
[20] D. Alessia, F. Ferri, P. Grifoni, and T. Guzzo, "Approaches, tools and applications for sentiment analysis implementation," International Journal of Computer Applications, vol. 125, no. 3, 2015.
[21] L. Yang, Y. Li, J. Wang, and R. S. Sherratt, "Sentiment analysis for e-commerce product reviews in chinese based on sentiment lexicon and deep learning," IEEE Access, vol. 8, pp. 23 522–23 530, 2020.
[22] S. M. Yimam, H. M. Alemayehu, A. Ayele, and C. Biemann, "Exploring amharic sentiment analysis from social media texts: Building annotation tools and classification models," in Proceedings of the 28th International Conference on Computational Linguistics, 2020, pp. 1048–1060.
[23] A. Alamoodi, B. Zaidan, A. Zaidan, O. Albahri, K. Mohammed, R. Malik, E. Almahdi, M. Chyad, Z. Tareq, A. Albahri et al., "Sentiment analysis and its applications in fighting covid-19 and infectious diseases: A systematic review," Expert systems with applications, p. 114155, 2020.
[24] M. Bhat, M. Qadri, M. K. Noor-ul Asrar Beg, N. Ahanger, and B. Agarwal, "Sentiment analysis of social media response on the covid19 outbreak," Brain, Behavior, and Immunity, 2020.
[25] K. Mehmood, D. Essam, K. Shafi, and M. K. Malik, "Discriminative feature spamming technique for roman urdu sentiment analysis," IEEE Access, vol. 7, pp. 47 991–48 002, 2019.
[26] R. Cattoni, M. A. Di Gangi, L. Bentivogli, M. Negri, and M. Turchi, "Must-c: A multilingual corpus for end-to-end speech translation," Computer Speech & Language, vol. 66, p. 101155, 2021.
[27] M. Popel, M. Tomkova, J. Tomek, Ł. Kaiser, J. Uszkoreit, O. Bojar, and Z. Žabokrtskỳ, "Transforming machine translation: a deep learning system reaches news translation quality comparable to human professionals," Nature communications, vol. 11, no. 1, pp. 1–15, 2020.
[28] M. Araújo, A. Pereira, and F. Benevenuto, "A comparative study of machine translation for multilingual sentence-level sentiment analysis," Information Sciences, vol. 512, pp. 1078–1102, 2020.
[29] A. Fan, S. Bhosale, H. Schwenk, Z. Ma, A. El-Kishky, S. Goyal, M. Baines, O. Celebi, G. Wenzek, V. Chaudhary et al., "Beyond english-centric multilingual machine translation," Journal of Machine Learning Research, vol. 22, no. 107, pp. 1–48, 2021.
[30] Y. Liu, J. Gu, N. Goyal, X. Li, S. Edunov, M. Ghazvininejad, M. Lewis, and L. Zettlemoyer, "Multilingual denoising pre-training for neural machine translation," Transactions of the Association for Computational Linguistics, vol. 8, pp. 726–742, 2020.
[31] A. Siddhant, M. Johnson, H. Tsai, N. Ari, J. Riesa, A. Bapna, O. Firat, and K. Raman, "Evaluating the cross-lingual effectiveness of massively multilingual neural machine translation," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, no. 05, 2020, pp. 8854–8861.
[32] F. Gao, J. Zhu, L. Wu, Y. Xia, T. Qin, X. Cheng, W. Zhou, and T.-Y. Liu, "Soft contextual data augmentation for neural machine translation," in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 5539–5544.
[33] K. Kowsari, K. Jafari Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown, "Text classification algorithms: A survey," Information, vol. 10, no. 4, p. 150, 2019.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3110285, IEEE Access

[34] K. Denecke, "Using sentiwordnet for multilingual sentiment analysis," in 2008 IEEE 24th international conference on data engineering workshop. IEEE, 2008, pp. 507–512.
[35] E. F. Can, A. Ezen-Can, and F. Can, "Multilingual sentiment analysis: An rnn-based framework for limited data," arXiv preprint arXiv:1806.04511, 2018.
[36] A. Balahur and M. Turchi, "Comparative experiments using supervised learning and machine translation for multilingual sentiment analysis," Computer Speech & Language, vol. 28, no. 1, pp. 56–75, 2014.
[37] M. Z. Asghar, A. Sattar, A. Khan, A. Ali, F. Masud Kundi, and S. Ahmad, "Creating sentiment lexicon for sentiment analysis in urdu: The case of a resource-poor language," Expert Systems, vol. 36, no. 3, p. e12397, 2019.
[38] A. Araabi and C. Monz, "Optimizing transformer for low-resource neural machine translation," in Proceedings of the 28th International Conference on Computational Linguistics. International Committee on Computational Linguistics, 2020, pp. 3429–3435.
[39] A. Poncelas, P. Lohar, J. Hadley, and A. Way, "The impact of indirect machine translation on sentiment classification," in Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track). Virtual: Association for Machine Translation in the Americas, 2020, pp. 78–88.
[40] A. Conneau, R. Rinott, G. Lample, H. Schwenk, V. Stoyanov, A. Williams, and S. Bowman, "Xnli: Evaluating cross-lingual sentence representations," in Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018, E. Riloff, D. Chiang, J. Hockenmaier, and J. Tsujii, Eds. Association for Computational Linguistics, 2020, pp. 2475–2485.
[41] S. Baccianella, A. Esuli, and F. Sebastiani, "Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining." in Lrec, vol. 10, no. 2010, 2010, pp. 2200–2204.
[42] M. Araujo, J. Reis, A. Pereira, and F. Benevenuto, "An evaluation of machine translation for multilingual sentence-level sentiment analysis," in Proceedings of the 31st Annual ACM Symposium on Applied Computing, 2016, pp. 1140–1145.
[43] G. Shalunts, G. Backfried, and N. Commeignes, "The impact of machine translation on sentiment analysis," Data Analytics, vol. 63, pp. 51–56, 2016.
[44] V. Barriere and A. Balahur, "Improving sentiment analysis over non-English tweets using multilingual transformers and automatic translation for data-augmentation," in Proceedings of the 28th International Conference on Computational Linguistics. Barcelona, Spain (Online): International Committee on Computational Linguistics, 2020, pp. 266–271.
[45] S. L. Lo, E. Cambria, R. Chiong, and D. Cornforth, "Multilingual sentiment analysis: from formal to informal and scarce resource languages," Artificial Intelligence Review, vol. 48, no. 4, pp. 499–527, 2017.
[46] M. Bilal, H. Israr, M. Shahid, and A. Khan, "Sentiment classification of roman-urdu opinions using naïve bayesian, decision tree and knn classification techniques," Journal of King Saud University-Computer and Information Sciences, vol. 28, no. 3, pp. 330–344, 2016.
[47] N. Mukhtar, M. A. Khan, and N. Chiragh, "Lexicon-based approach outperforms supervised machine learning approach for urdu sentiment analysis in multiple domains," Telematics and Informatics, vol. 35, no. 8, pp. 2173–2183, 2018.
[48] A. Rafique, M. K. Malik, Z. Nawaz, F. Bukhari, A. H. Jalbani et al., "Sentiment analysis for roman urdu," Mehran University Research Journal of Engineering & Technology, vol. 38, no. 2, p. 463, 2019.
[49] Z. Nasim and S. Ghani, "Sentiment analysis on urdu tweets using markov chains," SN Computer Science, vol. 1, no. 5, pp. 1–13, 2020.
[50] M. Amjad, G. Sidorov, and A. Zhila, "Data augmentation using machine translation for fake news detection in the urdu language," in Proceedings of The 12th Language Resources and Evaluation Conference, 2020, pp. 2537–2542.
[51] S. Mukund and R. K. Srihari, "Analyzing urdu social media for sentiments using transfer learning with controlled translations," in Proceedings of the Second Workshop on Language in Social Media, 2012, pp. 1–8.
[52] G. Thakkar, N. M. Preradovic, and M. Tadic, "Multi-task learning for cross-lingual sentiment analysis." in CLEOPATRA@ WWW, 2021, pp. 76–84.
[53] V. Ivanyuk and E. Tsapina, "Creating emotion recognition algorithms based on a convolutional neural network for sentiment analysis," in International Workshop on Reproducible Research in Pattern Recognition. Springer, 2021, pp. 66–79.
[54] A. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, "Learning word vectors for sentiment analysis," in Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies, 2011, pp. 142–150.
[55] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, J. Klingner, A. Shah, M. Johnson, X. Liu, Łukasz Kaiser, S. Gouws, Y. Kato, T. Kudo, H. Kazawa, K. Stevens, G. Kurian, N. Patil, W. Wang, C. Young, J. Smith, J. Riesa, A. Rudnick, O. Vinyals, G. Corrado, M. Hughes, and J. Dean, "Google's neural machine translation system: Bridging the gap between human and machine translation," CoRR, vol. abs/1609.08144, 2016.
[56] T. Luong, I. Sutskever, Q. Le, O. Vinyals, and W. Zaremba, "Addressing the rare word problem in neural machine translation," in Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Beijing, China: Association for Computational Linguistics, 2015, pp. 11–19.
[57] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2015.
[58] S. Pouyanfar, S. Sadiq, Y. Yan, H. Tian, Y. Tao, M. P. Reyes, M.-L. Shyu, S.-C. Chen, and S. Iyengar, "A survey on deep learning: Algorithms, techniques, and applications," ACM Computing Surveys (CSUR), vol. 51, no. 5, pp. 1–36, 2018.
[59] B. Jang, M. Kim, G. Harerimana, S.-u. Kang, and J. W. Kim, "Bi-lstm model to increase accuracy in text classification: combining word2vec cnn and attention mechanism," Applied Sciences, vol. 10, no. 17, p. 5841, 2020.
[60] A. Severyn and A. Moschitti, "Twitter sentiment analysis with deep convolutional neural networks," in Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval, 2015, pp. 959–962.
[61] W. S. Noble, "What is a support vector machine?" Nature biotechnology, vol. 24, no. 12, pp. 1565–1567, 2006.
[62] H. Zhang and D. Li, "Naïve bayes text classifier," in 2007 IEEE International Conference on Granular Computing (GRC 2007). IEEE, 2007, pp. 708–708.
[63] A. Luque, A. Carrasco, A. Martín, and A. de las Heras, "The impact of class imbalance in classification performance metrics based on the binary confusion matrix," Pattern Recognition, vol. 91, pp. 216–231, 2019.
[64] A. C. Baugh and T. Cable, A history of the English language, 2020.
[65] M. A. F. Muslah, "Implicit negation in selected romantic poems in the english and arabic (poetry contrastive study)," Al-Bahith Journal, vol. 14, no. 7, pp. 366–382, 2015.
[66] U. Farooq, H. Mansoor, A. Nongaillard, Y. Ouzrout, and M. A. Qadir, "Negation handling in sentiment analysis at sentence level." JCP, vol. 12, no. 5, pp. 470–478, 2017.
[67] S. K. Bharti, B. Vachha, R. Pradhan, K. S. Babu, and S. K. Jena, "Sarcastic sentiment detection in tweets streamed in real time: a big data approach," Digital Communications and Networks, vol. 2, no. 3, pp. 108–121, 2016.

ABDUL GHAFOOR is currently working as Instructor Computer Science at IBA Institute of Emerging Technologies, Khairpur, Pakistan. He is doing his Master degree in Computer science from Sukkur IBA University. His major research interests are Deep Learning, Natural Language Processing, Machine Learning and Data Mining.

ALI SHARIQ IMRAN (M'14) obtained his Ph.D. from the University of Oslo (UiO), Norway in Computer Science in 2013, and a Masters in Software Engineering and Computing from the National University of Science & Technology (NUST), Pakistan in 2008. He specializes in applied research with a focus on deep learning technology and its application to signal processing, natural language processing, and the semantic web. He has over 65 peer-reviewed journals and conference publications to his name and has served as a reviewer for many reputed journals over the years, including IEEE Access as an Associate Editor. He is a member of the Norwegian Colour and Visual Computing Laboratory (Colourlab) and an IEEE member, Norway section. He is currently associated with the Department of Computer Science at the Norwegian University of Science and Technology (NTNU), Norway, as an Associate Professor.

SHER MUHAMMAD DAUDPOTA received his Masters and PhD degrees from Asian Institute of Technology, Thailand in the year 2008 and 2012, respectively. His research areas include deep learning, natural language processing, video and signal processing. He is author of more than 35 peer-reviewed journal and conference publications. Presently he is serving as a Professor of Computer Science at Sukkur IBA University, Pakistan. Alongside his Computer Science contribution, he is also a Quality Assurance expert in higher education. He has reviewed more than 50 universities in Pakistan for quality assurance on behalf of Higher Education Commission in the role of educational quality reviewer.

ZENUN KASTRATI is currently associated with the Department of Informatics at Linnaeus University, Sweden, as a Senior Lecturer. He obtained his PhD from Norwegian University of Science and Technology (NTNU), Norway, in Computer Science in 2018 and a Master degree in Computer Science through the EU TEMPUS Programme developed and implemented jointly by University of Pristina, Kosovo, Universite de La Rochelle, France, and Institute of Technology Carlow, Republic of Ireland. His research interests focus on the field of artificial intelligence with a special focus on NLP, machine/deep learning and sentiment analysis. He is the author of more than 40 peer-reviewed journals and conferences and has served as a reviewer for many reputed journals over the years.

ABDULLAH received his MS degree in Computer Science with specialization in Data Knowledge Engineering in 2019 from Sukkur IBA University, Sukkur, Pakistan. He is currently working as Lecturer in Computer Science at FGEI (C/G), Quetta region, Pakistan. His research interest includes Machine Learning, Deep Learning, Natural Language Processing and Computer Vision.

RAKHI BATRA received her BS degree in Computer Science and MS degree in Data and Knowledge Engineering from Sukkur IBA University, Sukkur, Pakistan, in 2015 and 2019, respectively. Since 2018, she has been working as Assistant Manager of the ORIC Department, Sukkur IBA University. Her research interests include knowledge discovery, data mining, artificial intelligence, and deep learning.

MUDASIR AHMAD WANI is currently working as a Lecturer and Researcher at Department of Information Security and Communication Technology (IIK) at the Norwegian University of Science and Technology (NTNU), Norway. He pursued his postdoctoral research at Norwegian Biometrics Laboratory, NTNU, Norway and obtained his Ph.D. from Jamia Millia Islamia (A Central University), New Delhi, India in 2019 in Computer Science. He holds a Masters in computer Applications (MCA) and M.Phil. (Data Mining) from University of Kashmir (UoK) in 2012 and 2014 respectively. His research focuses on extraction and analysis of social data, and application of different statistical and machine/deep learning techniques in developing prediction models. He is actively involved in organizing and reviewing international conferences, workshops and journals.

• • •