Contents

Nomenclature

AraBERT - Arabic Bidirectional Encoder Representations from Transformers

AUC - Area Under the Curve

CNNs - Convolutional Neural Networks

DT - Decision Tree

FN - False Negative

FP - False Positive

LSTM - Long Short-Term Memory

mBERT - multilingual Bidirectional Encoder Representations from

MLM - Masked Language Modelling

NUS - National University of Singapore

RF - Random Forest

RNNs - Recurrent Neural Networks.

ROC - Receiver Operating Characteristic

TN - True Negative

TP - True Positive

SMS - Short Message Service

SVC - Support Vector Classifier

SVM - Support Vector Machine

TF-IDF - Term Frequency-Inverse Document Frequency

XLM-RoBERTa - Cross-lingual Language Model pretraining from Facebook AI Research

Abstract

The proliferation of spam and phishing attempts via messaging platforms threatens users across languages. While research exists on SMS spam detection, most focus on monolingual datasets, usually English. This report explores cross-lingual transfer learning to improve multilingual SMS spam classification across English, French, German, and Hindi. Transfer learning leverages pretrained multilingual word embeddings to transfer knowledge from high to low-resource languages. On an SMS dataset, a simple LSTM baseline achieved 87% accuracy. The cross-lingual transfer model improved accuracy to 97%, demonstrating the effectiveness of transfer learning. However, hyperparameter tuning led to some overfitting. Still, tuned transfer learning outperformed the LSTM baseline, validating its ability to boost multilingual performance when data is limited. This research provides a strong baseline for cross-lingual learning in multilingual SMS spam detection.

1 Introduction

Messaging platforms have always been used for the proliferation of unwanted and unsolicited content to users of these platforms; the Short Message Service (SMS) is not an exception. This has led to users receiving SMS spam. Liu et al., (2021) stated SMS spam is unwanted and immaterial messages sent via mobile communication networks. There have been several research on SMS spam detection due to the phishing, scamming, and unsolicited marketing attempts sent to the over 6 billion active users of the service which continues to increase (Abayomi-Alli et al., 2019). Reasons for the advancement of spam messages range from its relatively cheap fee to sending a message to the population of users of mobile telecommunication gadgets (Liu et al., 2021). Several machine learning algorithms have been deployed to create models that are capable of detecting spam texting messages, these attempts have been largely on traditional algorithms like SVC, Naïve Bayes, RF, and DT (Gaurav et al., 2020, Almeida et al., 2018). These algorithms have largely depended on handcrafted feature engineering (Salloum et al., 2021), which limits the adaptability of the model to evolve patterns in the text data.

In recent times, the deep learning framework has shown great promise in SMS spam detection systems (Sheneamer, A., 2021). Models like the CNNs, RNNs, LSTM and even more recently Transformer-based models have proven to do a decent job in accurate spam detection (Salman et al., 2022). However, the majority of the research has focused on using the monolingual language datasets. While English remains the language of business, expanding SMS spam detection capabilities to multiple languages is an important challenge (Nicholas., 2023). Cross-lingual transfer learning techniques show promise for transferring knowledge from high-resource languages like English to lower-resource languages to improve multilingual SMS spam detection (Mozafari et al., 2022).

1.1 Background and Review of Literature

Ample time has been placed into researching the possibility of accurately detecting SMS spam for over a decade. The fundamental challenge of using traditional machine learning approaches is the reliance on handcrafted features like bag-of-words, TF-IDF vectors, and stylometric features from SMS text from classifiers like Naive Bayes, SVC, and Random Forests (Gaikwad et al., 2021). This feature involves counting word frequencies, assigning higher weights, and acquiring sentence length and the use of punctuation. They are prone to overfitting on insignificant keywords rather than semantic meaning.

Application of deep learning has transformed the approach towards spam detection, there is more adoption of neural network architectures like CNNs which can analyse patterns in short text sequences and capture features like n-grams (Parwez and Abulaish., 2019), LSTMs recognize patterns in long-range messages which was used to achieve a 98.3% accuracy in email spam classification (Nooraee, M. and Ghaffari, H., 2022.), and attention-based Transformer networks excels by assigning different weights to words thereby capturing complex relationships in a message (Salman et al., 2022).

While great strides have been recorded in the usage of deep learning in SMS spam detection, the monolithic use of a single language structure in model building which is not representative of current realities has been an issue. Capturing the complexity of diverse languages is important as it would cover unique challenges like varied lexical semantics, syntax, morphology, and sentence structuring (Chauhan and Daniel., 2022; Zampieri et al., 2020). Cross-lingual transfer learning has been researched to show promise, the goal is to transfer knowledge to a different language which is usually lower from the primary language which is mostly English (Pantraki et al., 2022; Kostić et al., 2023; Ruder et al., 2019). It uses MLM as the main pretraining objective which is a process of masking some random tokens and predicting them using a model (Jiang et al., 2022). Similar research exists like Catelli et al., (2022) using a mBERT model on an English and Italian dataset on a travel blog with an 86% accuracy. The model was built by fine-tuning different architectures like attention heads, learning rates, sequence length, and hidden layers. El-Alami et al., (2022) combined transfer-learning and AraBERT to classify offensive and non-offensive statements on social media with a 91% accuracy. Gohl., (2022) used zero-shot cross-lingual transfer learning which combined a feedforward neural network, an LSTM, and an XLM-RoBERTa transformer to transfer the knowledge of an English language to a Swedish language. The average accuracy, recall, and precision score for the different models were 81%, 80%, and 83% respectively. However, their application to multilingual SMS spam detection remains relatively underexplored, most research focuses on bi-lingual cross learning which limits the model in a more complex language repository. The key challenges this report aims to achieve in the Cross-lingual transfer learning model are vocabulary mismatch, syntactic variance, semantic diversity, and data imbalance.

The proposed framework for this report is to develop a multilingual SMS spam detector using Cross-lingual transfer learning through word embeddings. The model is trained using four languages in other to build a very robust model with the capacity to detect SMS spam across various languages regardless of their complexities.

1.3 Research Question

How does the cross-lingual transfer learning technique perform in terms of SMS spam classification compared to traditional LSTM recurrent neural network classifiers when deployed across multiple languages?

1.2 Objectives of Study

1. To develop a multilingual SMS spam detection model using cross-lingual word embeddings for transfer learning across 3 languages - English, French, and German.

3. To evaluate the model on SMS spam datasets in each of the 4 languages and analyse performance in terms of metrics like accuracy, precision, recall and F1-score.

2 Methodology

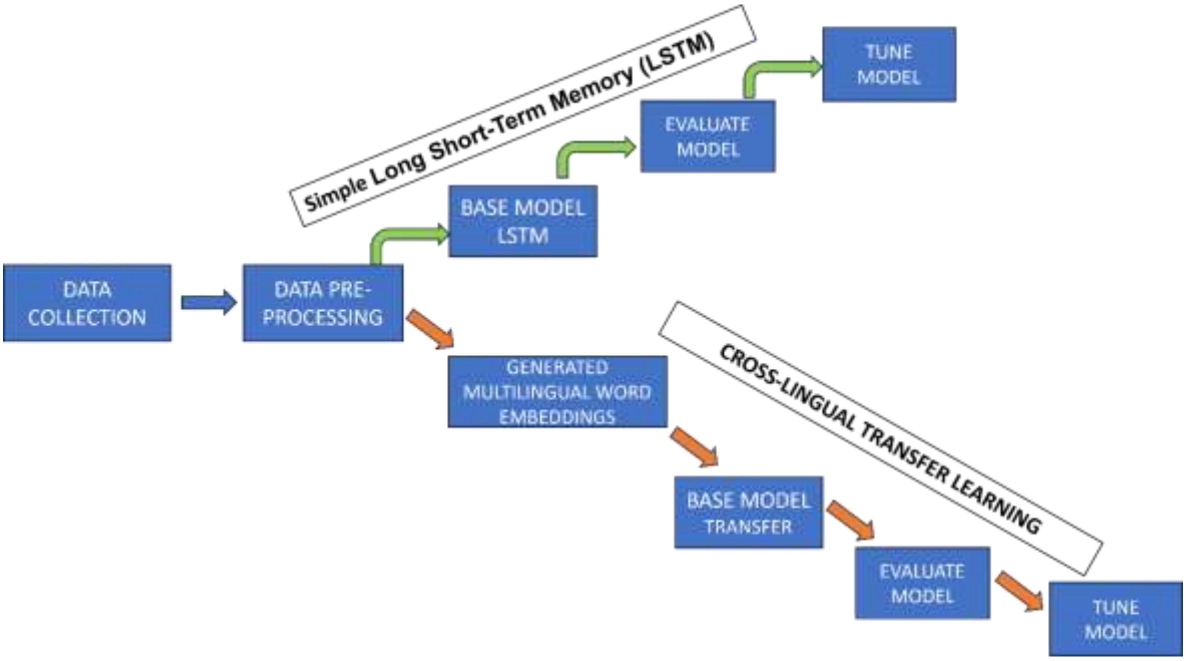

Figure 1: Workflow of the study

2.1 Dataset Collection

The SMS spam dataset was acquired from Kaggle which has its primary source from Almeida et al., (2012). The original text was in English and Machine Translated to German, and French. It contains 5,574 SMS messages which are labelled as ham (i.e. considered to be desirable), or spam (having the capacity to be harmful).

The messages were collected from the following sources:

- 425 spam SMS messages were extracted from a website called Grumbletext.
- The National University of Singapore researched several issues of which 3,375 of SMS messages were randomly sampled.
- 450 of the SMS data was from a Ph.D. thesis of Caroline Tag's.
- The dataset from Corpus v.0.1 which contained 1,002 ham and 322 spam SMS messages.

  These messages largely originate from English speakers, primarily in the UK and Singapore. The datasets were collected from public sources and with consent where applicable.

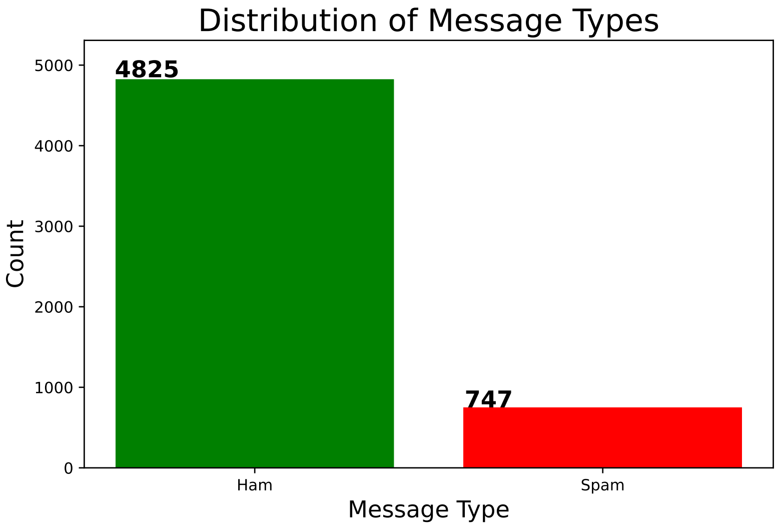

Figure 2: Distribution of ham (non-spam) and spam messages in the dataset represented as a histogram with ham in green and spam in red. The class imbalance skews heavily towards ham.

A wordcloud to display the most common words in each of the language in the dataset.

2.2 Data Preprocessing

The data preprocessing step is fundamental since raw data are always unclean with noise and inconsistencies, the format of the data may also not be compatible with the machine learning models so there is a need for certain changes to be executed before deploying the data to the algorithm. In this study, the following preprocessing steps were applied to the text data:

- **Lowercasing**: Text in upper or lowercase does not carry a great deal of semantic information so it is appropriate to convert to lowercase letters. The conversion was done using the string.lower() method which aids in the normalization of the data.
- **Punctuation removal**: Machine learning models do not require punctuation marks in their model, instead they create a great deal of noise if they are left in the data. Regular expression substitution was used to remove the punctuations.
- Lemmatization: Lemmatization is done to return words to their root form. It helps to create consistency in their semantic meanings. Lemmatization was executed using the WordNetLemmatizer class from NLTK.
- Stopword removal: Stopwords generally do not contribute significantly to the semantic structure of a sentence. Hence, they are removed before an algorithm is deployed with the data. This is done to avoid redundancy of the model. The NLTK stopwords corpus was used in removing stopwords.

After the data has been completely pre-processed, the train_test_split from the sklearn module will be used to randomly split the data into a training set of 80% and a validation of 20%.

2.3 Multilingual Word Representations

Multilingual Word representation is a similar concept to word embedding which is a process where diverse words from different languages with comparable meanings are mapped to a space of continuous high dimensionality with a comparable represented vector. The essence of embedded words is to convert the strings in the dataset into numerical values which can be understood by the computer before supplying the data to the model (Nooraee et al., 2022). Multilingual word representations can be executed using Multilingual BERT or FastText.

- Multilingual BERT: This is a transformer-based network that is designed to support multiple languages through its pre-trained models in an unsupervised process.

- FastText: This technique uses neural networks to vectorize words by representing each word as a bag of character n-grams.

In this study, multilingual word presentation will be executed with fastText which has a repository of pre-trained word vectors.

2.4 Model Architecture

Deep Learning LSTM Model

An LSTM network is one of the recurrent neural network algorithms with an architecture that can gather information and use a feedback mechanism through the learning of order dependence to solve sequence prediction problems (Nowak et al., 2017). LSTM is generally made up of gates and cell state which is critical in its infrastructure. The cell state is deployed for memory retention while the gates are divided into three which are the input, forget, and output gates. The gates bring new information into the model, decide on the information to throw away, and activate the final output respectively. The mathematical representations of the different gates in an LSTM are;

$$\text{Input gate } (i_t) = \sigma(w_i \,.\, [h_{t-1}, x_t\,] + \, b_i)$$
$$\text{forget gate } (f_t) = \sigma(w_f \,.\, [h_{t-1}, x_t\,] + \, b_f)$$
$$\text{Output gate } (o_t) = \sigma(w_o \,.\, [h_{t-1}, x_t\,] + \, b_o)$$

Where:

σ is the sigmoid function,
h_{t-1} previous LSTM output at time (t-1)
x_t is the input at the current timestamp.
i_t is the input gate.
f_t is the forget gate.
o_t is the output gate.
b represents the biases
w represents the weights

The diagram below shows the architecture of the LSTM.

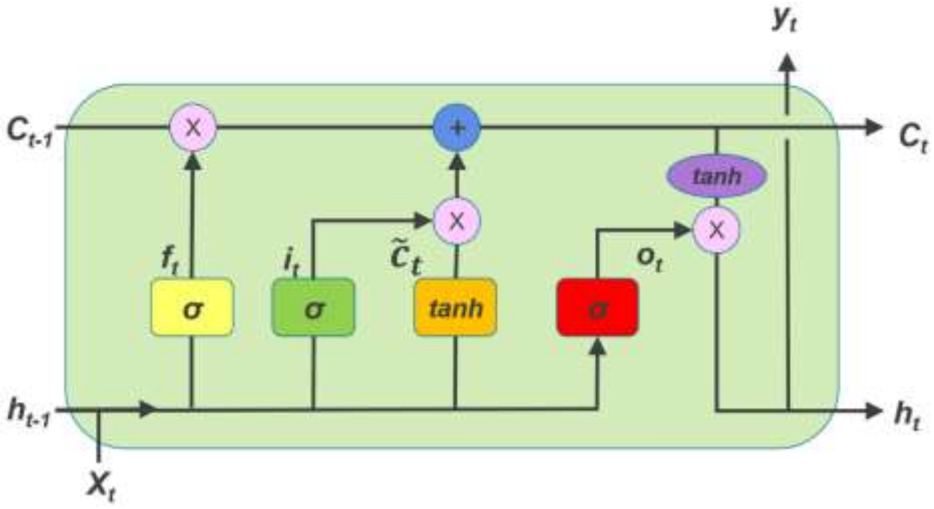

Figure 3: A LSTM architecture (Reddy, 2021)

Where:

Ct is the cell state at a time (t)
Ft is the forget gate
Ot is the output gate
σ is the sigmoid function

2.5 Cross-Lingual Transfer Learning

Transfer learning is the process of transmitting learning information from a particular domain which is most likely to a language with rich pedigree and lots of labels available to a lesser domain with very few (Wang et al., 2019). Through the use of cross-lingual transfer learning, models trained in a different language can be applied to another language. Cross-lingual transfer learning becomes very useful when there is very little label data in the target language. In this study, cross-lingual transfer learning is applied using fastText embeddings and LSTM networks. The use of fastText is to provide pre-trained word embeddings that obtain semantic relationships across languages. The model is built to first train in the high source language (English). The trained embedding layer is then transferred to a target LSTM model for a low-resource language (Hindu, German, French).

2.6 Evaluation Metrics

Model performance will be evaluated using accuracy, precision, recall, F1-score, and AUC-ROC. This is dependent on the model's success and failure in its classification using the TP, FP, TN and FN. The computation process of the evaluation metrics is shown below.

Accuracy = $\dfrac{TP+TN}{Total\ Population}$ Precision = $\dfrac{TP}{TP+FP}$ Recall = $\dfrac{TP}{TP+FN}$

F1-Score = $\dfrac{2(Precision * Recall)}{(Precision+ Recall)}$

AUC-ROC:

AUC-ROC is the area under the Receiver Operating Characteristic (ROC) curve which plots the True Positive Rate against the False Positive Rate at different classification thresholds. It summarizes the trade-off between true positives and false positives.

2.7 Results

- Base Model
  The result of the simple LSTM is used as a base model. The model architecture consists of input, embedding, and concatenated LSTM. The classification report of the model shows an accuracy of 87%, precision of 87% recall of 100% and F1-score of 93%.
  Some parameters were tuned like an increase in the learning rate from 10-5 to 10-10, adding dropout for regularization, and freezing the embedded layers. However, it made no input to the accuracy but applied RandomOverSampler from mblearn.over_sampling adversely reduces the accuracy from 0.87 to 0.45.

- Cross-Lingual Transfer Learning

  The model architecture of cross-lingual transfer learning is given in the table below.

  Table 1: The model is composed of LSTM layers that transform the input data sequentially.

Layer (type	Output Shape
Lstm_6 (LSTM)	(None, 4, 124)
Dropout_3 (Dropout)	(None, 4, 124)
Lstm_7 (LSTM)	(None, 64)
Dense_3 (Dense)	(None, 1)

The results show the model could predict the accuracy of 97%, multi-lingual precision, recall and F1-score of 94%, 77%, and 85% respectively.

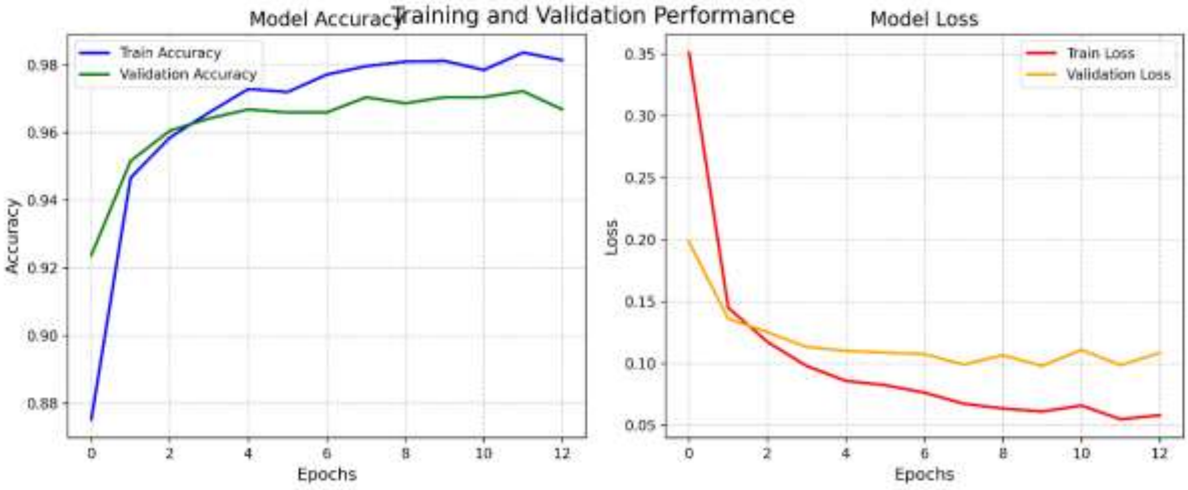

Figure 4: Visualizng the training and validation performance

The model evaluation on the high resource domain language (English) shows an accuracy of 97%, precision of 93%, recall of 85%, f1 of 89%, and auc_roc of 99%. However, the multi-lingual evaluation indicates a reduction with Multilingual Dataset Evaluation, accuracy of 93%, precision of 72%, recall of 80%, f1 of 76%, auc_roc of 96%
Hyper-tuning the cross-lingual transfer learning for the multilingual evaluation gives an accuracy of 93%, precision of 82%, recall of 62%, f1 of 71%, and an auc_roc of 96%.

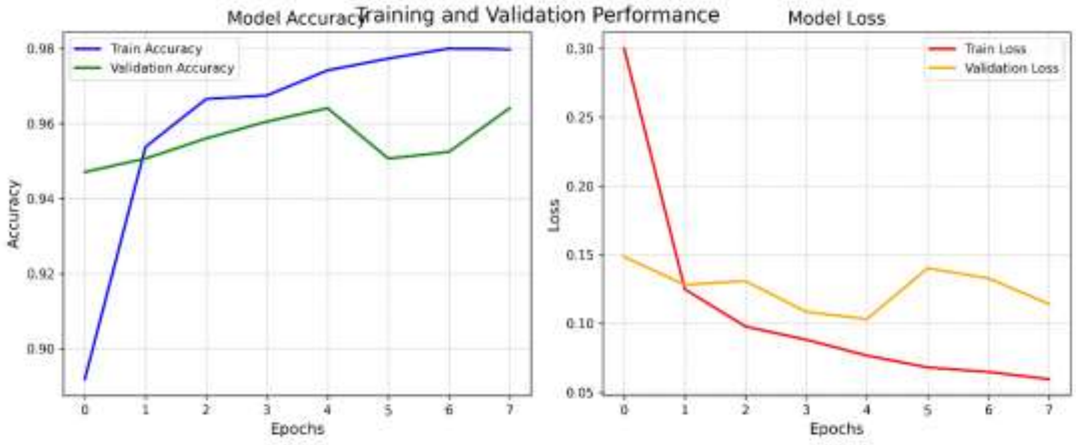

Figure 5: traning and validation performance of the hyper-tuned cross-lingual transfer learning.

2.8 Discussion

The results demonstrate the effectiveness of transfer learning for improving model performance on this multilingual text classification task.

Table 2: Model evaluation results on the multilingual SMS classification dataset. It compares the performance of the baseline LSTM model, transfer learning model, and hyperparameter-tuned transfer learning model across accuracy, precision, recall, F1 & auc_roc

Evaluation Metric	LSTM Base Model	Transfer Learning	Hyper-Tuned Transfer Learning
Accuracy	0.87	0.97	0.93
Precision	0.87	0.93	0.82
Recall	1.00	0.85	0.62
f1	0.93	0.89	0.71
auc_roc		0.99	0.96

From the table, there is a clear 10% increase in the model's accuracy from the baseline LSTM model of an accuracy of 0.87 to the cross-lingual transfer learning model with an improved accuracy of 0.97. This validates the ability of transfer learning which can be attributed to its pretrained rich resources where diverse languages can be leveraged upon. However, the addition of the hyperparameter tuning and regularization to the transfer model resulted in a partial decrease in all the model's evaluation metrics. The accuracy was reduced by 4%, the precision was reduced by 11%, the precision was reduced by 23%, the f1 was reduced by 18, and the auc_roc was reduced by 3%. The decrease in the evaluation metric indicates a potential overfitting to the majority class which implies the need to strike a balance between the usage of transfer learning and how it should be deployed to a specific domain. The combination of the accuracy and the auc_roc signifies the possibility of a range of classification thresholds.

2.9 Limitations

There are several limitations to this study which are listed below;

- Size of the dataset: The dataset was less than 10,000 rows which is relatively low for a neural network classifier model.
- Limited language diversity: Only four languages were used for the study at which three of the languages are European languages. This lack of more languages and its dispersion across the globe will limit the model to continental capacity.

- Limited model architectures: The study was predominantly anchored on the use of LSTM. Failures to expand into other architectures like Transformers, attention mechanisms, and class imbalance reduced the robustness of the study.

2.10 Future of Work

The study can be advanced within the following areas;

- Attention Mechanisms: The addition of attention mechanisms can enhance the model's ability to be more inclined towards the most important words in a particular language. This would prove useful for languages with different sentence structures and word order.

- Data Augmentation: The use of data augmentation will increase the volume of the data used in the research. This will improve model robustness and address the class imbalance observed.

- Architecture Optimization: Additional experiments can be executed with CNN and RNN architectures to uncover the optimal model design.

- Multilingual Pretraining: Other Pretraining transformer models like mBERT can be used to train the model and compare the performance.

- Low-Resource Language Tuning: Further techniques like self-training and semi-supervised learning can be explored to adapt the models to lower-resource languages with minimal labelled data.

3 Conclusion

This study demonstrated the potential of using cross-lingual transfer learning for multilingual SMS spam detection across English, French, German, and Hindi. The transfer learning approach leveraged pretrained multilingual word embeddings to improve model performance over a baseline LSTM network. On the SMS dataset, transfer learning boosted accuracy from 87% to 97%.

However, hyperparameter tuning led to some overfitting, indicating the need to balance transfer knowledge with task-specific adaptation. Still, even after tuning, the transfer model outperformed the base LSTM. The results validate cross-lingual transfer learning as an effective technique for multilingual NLP when labelled data is limited.

While promising, several limitations exist, including small dataset sizes, limited languages and architectures, and class imbalance. Future work should focus on larger-scale multilingual data, more diverse languages, optimized neural architectures, data augmentation, and model introspection. Overall, this research establishes a strong baseline for cross-lingual learning in SMS spam detection.

References

Abayomi-Alli, O., Misra, S., Abayomi-Alli, A. and Odusami, M., 2019. A review of soft techniques for SMS spam classification: Methods, approaches and applications. *Engineering Applications of Artificial Intelligence*, *86*, pp.197-212.

Almeida,Tiago and Hidalgo,Jos. (2012). SMS Spam Collection. UCI Machine Learning Repository. https://doi.org/10.24432/C5CC84.

Almeida, T.A., Yamakami, A. and Almeida, J., 2018. Filtering spam mobile text messages using support vector machines. Journal of Information Security and Applications, 40, pp.76-82.

Brownlee, J., 2019. A Gentle Introduction to Long Short-Term Memory Networks by the Experts. [online] Machine Learning Mastery. Available at: https://machinelearningmastery.com/gentle-introduction-long-short-term-memory-networks-experts/ [Accessed 24 February 2023].

Catelli, R., Bevilacqua, L., Mariniello, N., di Carlo, V.S., Magaldi, M., Fujita, H., De Pietro, G. and Esposito, M., 2022. Cross lingual transfer learning for sentiment analysis of Italian TripAdvisor reviews. *Expert Systems with Applications*, p.118246.

Chauhan, S. and Daniel, P., 2022. A comprehensive survey on various fully automatic machine translation evaluation metrics. *Neural Processing Letters*, pp.1-55.

El-Alami, F.Z., El Alaoui, S.O. and Nahnahi, N.E., 2022. A multilingual offensive language detection method based on transfer learning from transformer fine-tuning model. *Journal of King Saud University-Computer and Information Sciences*, *34*(8), pp.6048-6056.

Gaikwad, M., Ahirrao, S., Phansalkar, S. and Kotecha, K., 2021. Online extremism detection: A systematic literature review with emphasis on datasets, classification techniques, validation methods, and tools. *Ieee Access*, *9*, pp.48364-48404.

Gaurav, D., Tiwari, S.M., Goyal, A., Gandhi, N. and Abraham, A., 2020. Machine intelligence-based algorithms for spam filtering on document labeling. *Soft Computing*, *24*, pp.9625-9638.

Göhl, S.A., 2022. Zero-shot cross-lingual transfer learning for sentiment analysis on Swedish chat conversations.

Jiang, X., Liang, Y., Chen, W. and Duan, N., 2022, June. XLM-K: Improving cross-lingual language model pre-training with multilingual knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 36, No. 10, pp. 10840-10848).

Kostić, M., Batanović, V. and Nikolić, B., 2023. Monolingual, multilingual and cross-lingual code comment classification. *Engineering Applications of Artificial Intelligence*, *124*, p.106485.

Liu, X., Lu, H. and Nayak, A., 2021. A spam transformer model for SMS spam detection. *IEEE Access*, *9*, pp.80253-80263.

Mozafari, M., Farahbakhsh, R. and Crespi, N., 2022. Cross-lingual few-shot hate speech and offensive language detection using meta learning. *IEEE Access*, *10*, pp.14880-14896.

Nicholas, G. and Bhatia, A., 2023. Lost in Translation: Large Language Models in Non-English Content Analysis. *arXiv preprint arXiv:2306.07377*.

Nooraee, Mohsen, and Hamidreza Ghaffari. "Optimization and Improvement of Spam Email Detection Using Deep Learning Approaches." *Journal of Computer & Robotics* 15, no. 2 (2022): 61-70.

Nowak, J., Taspinar, A. and Scherer, R., 2017. LSTM recurrent neural networks for short text and sentiment classification. In *Artificial Intelligence and Soft Computing: 16th International Conference, ICAISC 2017, Zakopane, Poland, June 11-15, 2017, Proceedings, Part II 16* (pp. 553-562). Springer International Publishing.

Pantraki, E., Tsingalis, I. and Kotropoulos, C., 2022. Cross-lingual transfer learning: A PARAFAC2 approach. *Pattern Recognition Letters*, *159*, pp.167-173.

Parwez, M.A. and Abulaish, M., 2019. Multi-label classification of microblogging texts using convolution neural network. *IEEE Access*, *7*, pp.68678-68691.

Ruder, S., Vulić, I. and Søgaard, A., 2019. A survey of cross-lingual word embedding models. Journal of Artificial Intelligence Research, 65, pp.569-631.

Salloum, S., Gaber, T., Vadera, S. and Shaalan, K., 2021. Phishing email detection using natural language processing techniques: a literature survey. *Procedia Computer Science*, *189*, pp.19-28.

Salman, M., Ikram, M. and Kaafar, M.A., 2022. An Empirical Analysis of SMS Scam Detection Systems. *arXiv preprint arXiv:2210.10451*.

Sheneamer, A., 2021. Comparison of Deep and Traditional Learning Methods for Email Spam Filtering. *International Journal of Advanced Computer Science and Applications*, *12*(1).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I., 2017. Attention is all you need. Advances in neural information processing systems, 30.

Wang, J., Chen, Y., Yu, H., Huang, M. and Yang, Q., 2019, July. Easy transfer learning by exploiting intra-domain structures. In *2019 IEEE international conference on multimedia and expo (ICME)* (pp. 1210-1215). IEEE.

Zampieri, M., Nakov, P., Rosenthal, S., Atanasova, P., Karadzhov, G., Mubarak, H., Derczynski, L., Pitenis, Z. and Çöltekin, Ç., 2020. SemEval-2020 task 12: Multilingual offensive language identification in social media (OffensEval 2020). *arXiv preprint arXiv:2006.07235*.