Sentiment Analysis Using XLM-R Transformer and Zero-shot Transfer Learning on Resource-poor Indian Language

AKSHI KUMAR, Department of Computer Science & Engineering, Delhi Technological University, New Delhi, India
VICTOR HUGO C. ALBUQUERQUE, Laboratory of Industrial Informatics, Electronics and Health, University of Fortaleza (UNIFOR), Ceará, Brazil

Sentiment analysis on social media relies on comprehending the natural language and using a robust machine learning technique that learns multiple layers of representations or features of the data and produces state-of-the-art prediction results. The cultural miscellanies, geographically limited trending topic hash-tags, access to aboriginal language keyboards, and conversational comfort in native language compound the linguistic challenges of sentiment analysis. This research evaluates the performance of cross-lingual contextual word embeddings and zero-shot transfer learning in projecting predictions from resource-rich English to resource-poor Hindi language. The cross-lingual XLM-RoBERTa classification model is trained and fine-tuned using the English language Benchmark SemEval 2017 dataset Task 4 A and subsequently zero-shot transfer learning is used to evaluate the classification model on two Hindi sentence-level sentiment analysis datasets, namely, IITP-Movie and IITP-Product review datasets. The proposed model compares favorably to state-of-the-art approaches and gives an effective solution to sentence-level (tweet-level) analysis of sentiments in a resource-poor scenario. The proposed model compares favorably to state-of-the-art approaches and achieves an average performance accuracy of 60.93 on both the Hindi datasets.

CCS Concepts: • **Information systems → Information retrieval**; **Retrieval tasks and goals**; **Sentiment analysis**; • **Computing methodologies → Artificial intelligence**; **Natural language processing**; **Information extraction**;

Additional Key Words and Phrases: Sentiment analysis, resource-poor language, transformer, deep learning

Authors' addresses: A. Kumar, Department of Computer Science & Engineering, Delhi Technological University, New Delhi, India; email: akshikumar@dce.ac.in; V. H. C. de Albuquerque, Laboratory of Industrial Informatics, Electronics and Health, University of Fortaleza (UNIFOR), Ceará, Brazil; email: victor.albuquerque@ieee.org.
Author's Current Address: Victor Hugo C. de Albuquerque, Graduate Program on Tele-Informatics Engineering, Federal University of Ceará, Fortaleza, Fortaleza/CE, Brazil. Graduate Program on Telecommunication Engineering, Federal Institute of Education, Science and Technology of Ceará, Fortaleza/CE, Brazil.

1 INTRODUCTION

Social media has become extremely popular in today's time as people often view it as a platform to voice their opinions about various organizations, people, companies, products, or events. There is endless possibility to use this wealth of information for getting to know the sentiments of people, their attitudes, and their emotions. The quantum of opinionated data on social media is so vast that it now finds use in tracking customer reactions, monitoring competitions, anticipating election outcomes, and predicting investment trends and box office revenues. **Sentiment analysis (SA)** [1, 2] is the field of study that automatically ascertains the polarity of opinions expressed by users toward entities in a chunk of text or review. In general, polarity is categorized into one of three possible classes, namely, positive, negative, or neutral. But determining real-time sentiments is challenging owing to volume of available reviews and variety in review expression. Biased and fake/spam reviews further complicate predictive analytics. The studies within the domain of SA are primarily divided into six types, which include studies based on granularity, language, modality, techniques, application areas, and sub-tasks. On the basis of granularity of analysis, SA are either categorized as coarse-grained (document level and sentence level) or fine-grained (phrase level or aspect level). On the basis of language analyzed, SA can either be monolingual (one language like English, Hindi, Chinese, Arabic, etc.) or Multilingual (combination of two or more languages (code-mix and code-switch). With the recent proliferated use of multimedia content in social media, the studies on SA can either be categorized into mono-modal (single modality, primarily text only) or Multimodal (combination of two or more media types, for example, text +image). The techniques used for SA broadly fall into four categories, that is, lexicon-based, machine learning-based, deep learning-based, and hybrid techniques. There are various application areas where the use of SA can assist intelligence and decision making. For example, business, marketing, governance and politics, and so on. Finally, the study on SA can further be done based on the related sub-tasks such as subjectivity analysis, sarcasm/irony/pun/humour detection, and emotion analysis, to name a few.

Automatic detection of sentiments in social media platforms is typically a natural language understanding-based classification task [3]. The data scientists have accomplished a lot at creating more accurate sentiment classifiers, but there's still a lot to do. Few key challenges of automated sentiment analysis include identifying subjectivity and tone; context and polarity; irony and sarcasm; negations; comparisons; emojis or emoticons and defining neutral. The cultural miscellanies, geographically limited trending topic hash-tags, access to aboriginal language keyboards and conversational comfort in native language exemplify the variability and size of user-generated content, thereby compounding the linguistic challenges of SA. Hindi is the official language of India and a sizeable population speaks/writes Hindi, while the rest are comfortable in their regional language. The availability of keyboards with "Devanagari" scripts on mobile phones has made it a popular language choice. But like other Indian languages, Hindi is also a resource-poor language [4]. Sentiment analysis in Indian languages is largely unexplored due to the scarcity of various resources and tools such as annotated corpora, lexicons, dependency parser, **Part-of-Speech (PoS)** tagger and benchmark datasets, and so on. Also, like most of the Indian languages, Hindi too has free-word order. For example, गाने अच्छे हैं इस फिल्म के, इस फिल्म के गाने अच्छे हैं, अच्छे गाने हैं इस फिल्म के, all three statements convey the same meaning with different word order.

There have been several recent studies published on sentiment analysis focusing on different aspects of affect recognition such as emotion [5, 6], aspect [7, 8], and mood [9], to name a few. While there are a few studies published on languages such as Arabic [10, 11] and Urdu [12], most studies and datasets created primarily contain English data. The use of monolingual word embeddings is pervasive in **natural language processing (NLP)**. Data augmentation [13, 14] and multilingual word embeddings [15, 16] have also been applied to take advantage of existing English datasets to improve the performance in systems dealing with languages other than

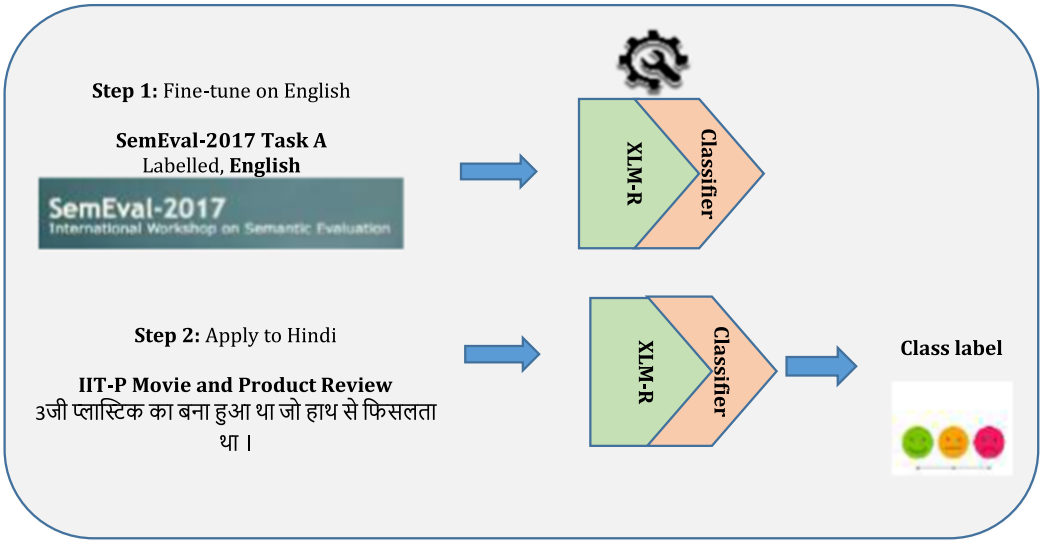

Fig. 1. Cross-lingual XLM-R transformer model.

English. Even though, multilingual model like BERT-m [13] show some cross-lingual characteristics it has not been trained on cross-lingual data [17]. To represent meaning and transfer knowledge across different languages, cross-lingual word embeddings can be used. These are embeddings that are obtained from a dataset in any available language and then used to train a model that will be used to predict polarity of texts in resource-poor language. To the best of our knowledge, the state-of-the-art cross-lingual contextual embeddings such as **XLM-RoBERTa (XLM-R)** [18] have not been applied yet to sentiment analysis in resource-poor Hindi language. To address this gap, we evaluate the performance of cross-lingual contextual word embeddings and transfer learning in projecting predictions from English to Hindi language. The classification model is trained on a resource rich language, typically English, using a cross-lingual transformer model and zero-shot transfer learning (while training, the model is not seeing even a single example of test language, but it can still make predictions in test language) is performed on a resource-poor language, i.e., Hindi (Figure 1)

The proposed model compares favorably to state-of-the-art approaches and gives an effective solution to sentence-level (tweet-level) analysis of sentiments in a resource-poor scenario. Thus, the primary contribution of this work is:

- Use of cross-lingual XLM-R transformer model for tweet-level sentiment classification for resource-poor Hindi language
- Apply zero-shot transfer learning for target resource-poor languages that lack extensive labelled and unlabelled data sets

Organization of the article is as follows: Section 2 briefs about the related work within the domain of Hindi sentiment analysis, followed by the discussion on methodology used in Section 3. Section 4 presents the details of model performance, and the conclusion is given in Section 5.

2 RELATED WORK

Sentiment analysis is the detection of attitudes, defined using a 4-tuple {Holder, Target, Type, Text}, where the holder is the source of attitude, target is the aspect of attitude, type defines the

commonly weighted polarity type such as positive, negative or neutral and text is the piece of information containing the attitude (sentence or entire document).

Having started with simple polarity detection, contemporary SA has advanced to a more nuanced analysis of context, affect and emotion sensing. But detecting fine-grained sentiment in natural language is tricky even for humans, making its automated detection more complicated. Moreover, online opinions can be put forth in the form of text reviews or ratings, for a product as a whole, or each of its individual aspects. Multiple and lengthy reviews, usage of casual dialect with micro-text (wordplay, neologism and slang), use of figurative language (sarcasm, irony), multilingual content (code-mixed and code-switched), and opinion spamming add challenges to the task of extracting opinions. Recently, memes, GIFs, typo-graphic (artistic way of text representation), info-graphic (text embedded along with an image) visual content and edited videos dominate the social feeds. Hence the task of SA can range from a simplistic classification of text attitude as positive or negative to a more advanced fine-grain emotion analysis and complex attitude type detection. In recent times the "text" component has compounded as multimedia text owing to the extensive use of various semiotic modalities (aural, visual, and textual) on social media. At the same time, multilingual sentiment analysis has also emerged as a prominent research problem as most of the existent work has been done on English only text [19]. Figure 2 depicts the various sentiment analysis tasks and the approaches used.

Various studies on diverse resource-poor languages have been reported in literature. Delbrouck et al. [20] proffered a **transformer-based joint-encoding (TBJE)** for the task of emotion recognition and SA in CMU-MOSEI dataset. González et al. [21] gave a transformer encoder model for Twitter SA in Spanish language. Sultan et al. [22] used XLM-R and transfer learning on both the monolingual English and Spanish data and Spanish-English code-mixed data. Kuratov and Arkhipov [23] used transfer learning from a multilingual BERT model to monolingual model for SA in Russian language. Sarkar et al. [24] proposed a Hierarchical Attentive Network using BERT for multilingual document-level SA. Recent studies report SA in code-mix Hindi-English text [25, 26], which was provided as the SemEval 2020 Task 9 [27]. Literature also reports emotion analysis in Hindi text. In 2019, Kumar et al. [28] introduced a Hindi emotion annotated text corpus, BHAAV. In 2020, Garg and Lobiyal [29] propose Hindi EmotionNet for SA in Hindi.

Studies related to SA on Hindi have been reported since the start of decade. Benchmark studies have been reported on three datasets, namely, the Twitter (SAIL 2015) dataset,[1] IIT-Patna Product/Service reviews dataset (sentence-level and aspect-level), and IIT-Patna Movie review dataset (sentence-level).[2] In 2012, Bakliwal et al. [30] provided Hindi subjective lexicon for SA. In the same year, Balamurali et al. [31] linked WordNets of Hindi and Marathi for cross-lingual SA. In 2015, various studies have been reported on the shared task SAIL dataset [32-34], which contained Hindi, Bengali, and Tamil tweets. Akhtar et al. [35] presented work on aspect-based SA (ABSA) in Hindi have also been reported. Cross-lingual and multi-lingual ABSA has also been reported [36, 37]. Pertinent studies also report deep learning architectures for SA in Hindi. In 2016, Akhtar et al. [38] proposed a hybrid deep learning model and evaluated its performance on four Hindi language datasets. The model achieved an average accuracy of 51.11% on the IITP-Movie and IIIT-Product review datasets. Most recently, in 2020 Kunchukuttan et al. [16] presented the IndicNLP word embeddings and achieved state-of-the-art results for the IITP-Movie and IIIT-Product review datasets.

[1] http://amitavadas.com/SAIL/data.html.
[2] https://www.iitp.ac.in/~ai-nlp-ml/resources.html.

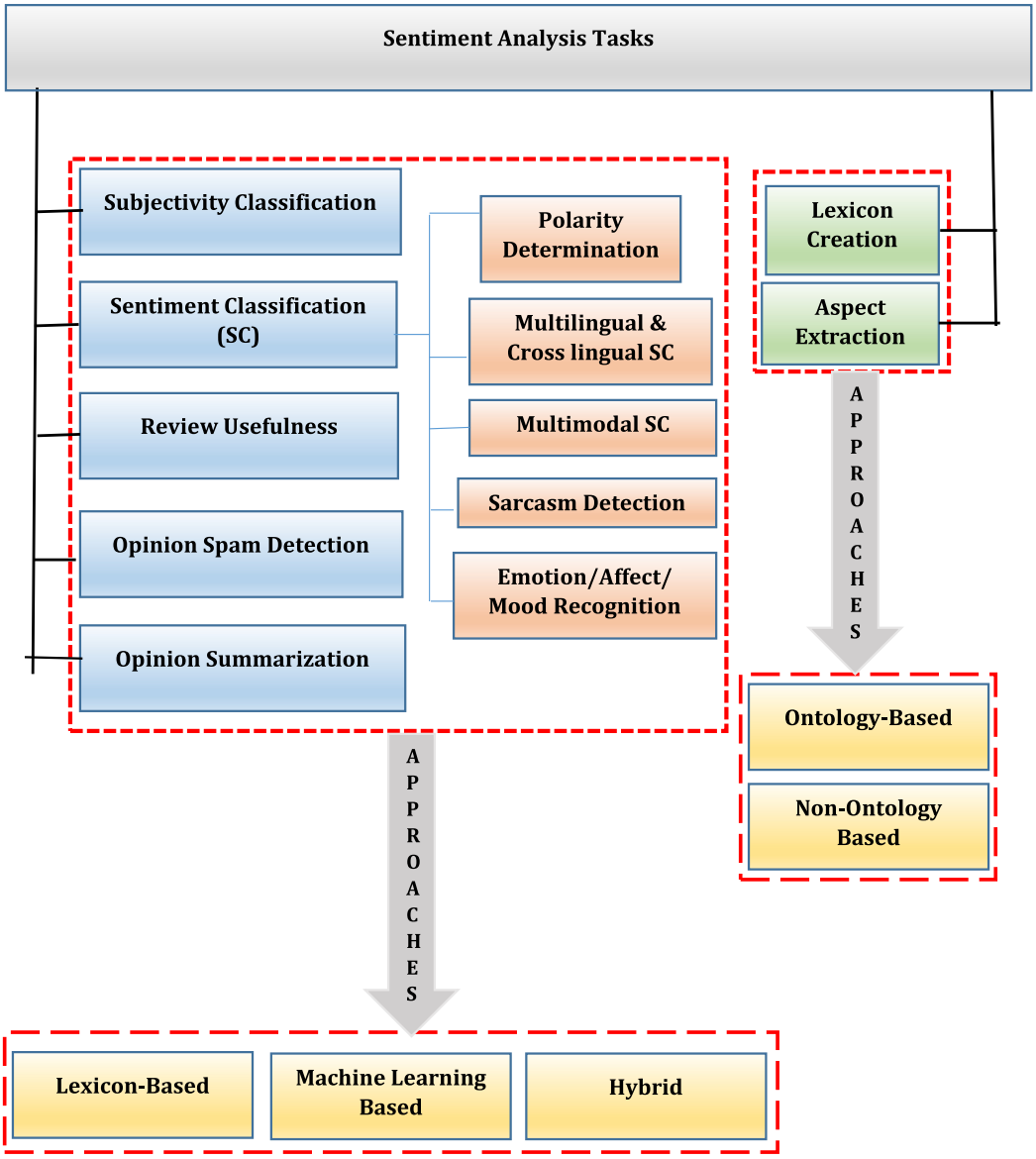

Fig. 2. Sentiment analysis tasks.

3 METHODOLOGY

3.1 Datasets

The SemEval 2017 Task 4A dataset,[3] which consists of English tweets annotated for positive, negative, or neutral categories, is used to train the language model on the sentiment analysis task on English, i.e., the training dataset. The IIT-Patna Movie and Product review datasets, which contains Hindi texts, are used to experiment transfer learning and accomplish sentence-level

[3]https://alt.qcri.org/semeval2017/task4/.

Table 1. Dataset Statistics

Dataset	Language	Positive	Negative	Neutral	Conflict	Total
SemEval-2017 Task 4A*	English	2,352	3,811	5,742	F	11,905
Movie Review	Hindi	823	530	598	201	2,152
Product Review	Hindi	2,290	712	2,226	189	5,417

*Only Tweet IDs provided by organizers and therefore some tweets were not available for download due to edited privacy or removed.

sentiment analysis task on Hindi language. Though the IIT-Patna Movie review set has been annotated with four labels (positive, negative, neutral, or conflict), we choose to ignore the conflict class to ensure the same number of classes in the target dataset. The dataset statistics are given in Table 1.

3.2 Pre-processing

Deep learning models accept certain kinds of inputs, which is vectors of integers, each value representing a token. Each string of text must first be converted to a list of indices to be fed to the model. Also, the data is converted into BERT specific format. That is, before performing tokenization, the text is converted to lower case. This conversion to lower-case is performed through the BERT tokenizer and the TFIDF vectorizer. Long sentences, if any, were split into multiple samples.

3.3 XLM-R Transformer Architecture for Sentiment Classification

XLM-R is a transformer-based multilingual masked language model released by the Facebook AI team in November 2019 as an update of its original XLM-100 model. It is a scaled cross-lingual sentence encoder trained on 2.5 Tb of data across 100 languages data filtered from CommonCrawl texts. Compared to the original version, the biggest update of XLM-Roberta is a significant increase in the amount of training data [18].

XLM-R achieves state-of-the-arts results on multiple cross-lingual benchmarks. Based on masked language model, it works as a successful alternative for non-English NLP. The remarkable fact about XLM-R is that it is compatible with both monolingual as well as cross-lingual benchmarks and alleviates the curse of multilinguality. In this research, we first fine-tune pre-trained XLM-RoBERTa (XML-R) with the SemEval-2017 Task 4 A dataset and then use zero-shot cross lingual transfer learning to test Hindi sentiment dataset. That is, the classification model is trained on a resource rich English language using XML-R cross-lingual transformer model and transfer learning is done on the resource-poor Hindi language. The model consists of two components: (1) Sentiment classification component for both languages and (2) cross-lingual transfer learning component that uses English sentiment analysis data to predict positive, negative and neutral sentiment class in resource-poor Hindi.

We use the XLM-R_{large} model. The model contains approximately 355M parameters with 24-layers, 1,027 hidden-states, 4,096 feed-forward hidden-states and 16-heads. The maximum sequence length can be set to default, 512, that is, it takes an input of a sequence of no more than 512 tokens and outputs the representation of the sequence. The first token of the sequence is always [CLS], which contains the special classification embedding [39]. The first token of every sequence is always a special classification token ([CLS]). The final hidden state, h, corresponding to this token is used as the aggregate sequence representation for classification tasks. A simple softmax classifier is added to the top of XLM-R to predict the probability of label c, as shown in Equation (1), where W is the task-specific parameter matrix:

$$(c|h) = softmax(W\mathrm{h}). \tag{1}$$

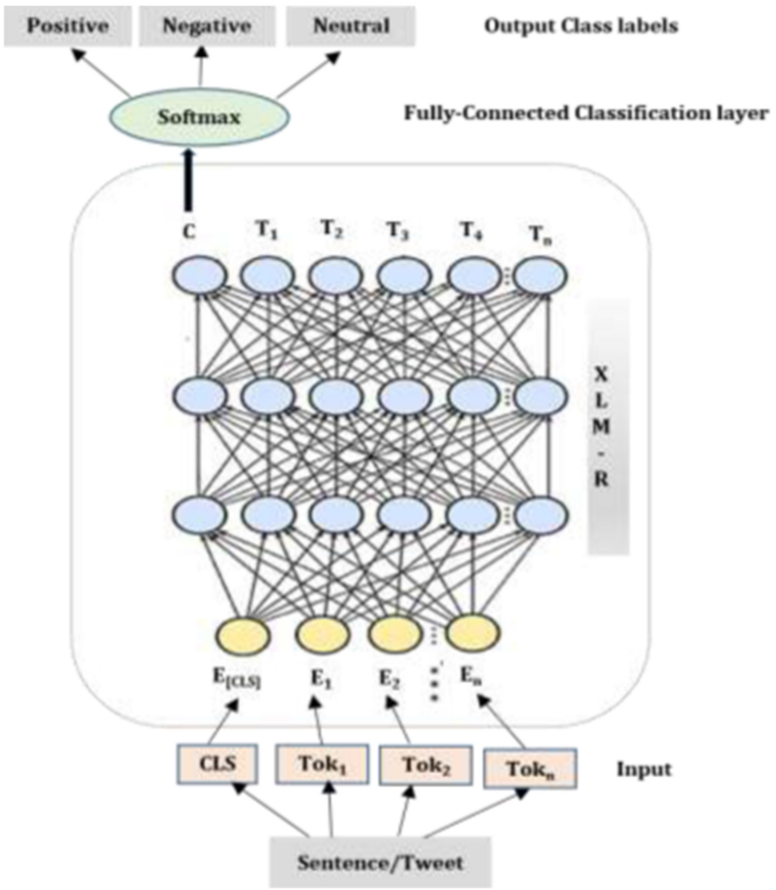

Fig. 3. Sentiment classification architecture using XLM-R.

We fine-tune all the parameters from XLM-R as well as W jointly by maximizing the log-probability of the correct label. The architecture of the sentiment classification is shown in Figure 3.

3.4 Zero-shot Cross-lingual Transfer Learning

From a psychological point of view, transfer of learning is the "study of dependency of human conduct, learning, or performance on prior experience." Analogously, in machine learning, transfer learning is the ability of a system to recognize and apply knowledge and skills learned in previous domains/tasks to novel tasks/domains, which share some commonality. Building every model from scratch is time consuming and expensive in terms of data collection/labeling, privacy and training time whereas reusing common knowledge extracted from existing systems is a viable solution. Transfer learning differs from traditional machine learning, because it involves using a pre-trained model as a springboard to start a secondary task. A pre-trained model is a saved network that was previously trained on a large dataset, typically on a large-scale sentiment classification task. Advantages of using transfer learning include less training data with better generalization capabilities make deep learning more accessible, faster convergence and higher asymptomatic accuracy. In transfer learning, we reuse by freezing or fine tuning the layers of a model. Domain adaption, domain confusion, multi-task learning, one-shot learning and zero-shot learning are various

types of transfer learning. For training and testing a learning model, we need data and the number of samples of a class required in data for learning are shots for that class. In zero-shot the machine is capable of describing what class an unlabeled sample belongs to when it does not fall into the category of any of the trained categories, that is, zero shots for the data point. In this work, we utilize zero-shot cross-lingual learning [40] to solve a sentiment analysis task in Hindi despite not having received any training examples of the task. Zero-shot means that the cross-lingual XLM-R is fine-tuned on one language and then evaluated on the foreign language test. In this case, machine translation was not involved at all in either the pre-training or fine-tuning.

The concept of transfer learning allows a novel algorithm to inherit the knowledge (features, weights) of the existing algorithm [41, 42]. That is, we use what well-trained, well-constructed networks have learned over large sets, and apply them to boost the performance of a detector on a smaller data set. Deep neural models are centered on network weights that come from training them on huge amount of data. That is, if the weights in the existing model are obtained and transferred to the new neural model, then it implies the "transfer" of learned features without building the model from scratch and retraining it. The object in the transfer method has both the network structure and weights. For adopting transfer learning to cross-lingual sentiment analysis two methods describing what to transfer can be used:

- *Transfer learning method 1 (TLM1):* Network is pre-trained on source task and the full structure (all layers), and all weights of network are transferred to the second network. That is, it starts with all the previously trained, fine-tuned weights from source, and then test it for the target. (Freeze the base.)
- *Transfer learning method 2 (TLM 2):* Network is pre-trained on source task and the partial structure (some layers), and their weights are transferred to the second network. The transferred weights are frozen, whereas the non-transferred weights are randomly initialized for a second training phase for the target task. Fine-tuning is done. (Train some layers, leave some frozen, i.e., selectively unfreeze.)

In this work, we use TLM1, where the XLM-R classification model is trained using the English language Benchmark SemEval 2017 dataset Task 4A. The complete network structure including the fully-connected softmax layer is adopted along with all weights to initialize the weights for the Hindi language sentiment classification model. Basically, we freeze the weights of pre-trained model in one language and use them for extracting the features in the other. In Keras, each layer has a parameter called "trainable". For freezing the weights of a particular layer, we should set this parameter to False, indicating that this layer should not be trained. To examine TLM1, we experiment on both the IIT-Patna Movie and Product review datasets, which has four sentiment classes, namely, positive, negative, and neutral or conflict, but as discussed, we choose to ignore the conflict class to ensure the same number of classes in the target dataset. If we consider different number of output classes to predict in the target language, then TLM2 is a better strategy where the last softmax layer is detached from the pre-trained base XLM-R. Therefore, in this case new fully-connected layers according to the number of classes in the target dataset are added with randomized weights (with frozen weights of re-trained model) and updated by training the classifier layers on training data available for task. It involves hyperparameter tuning and may require unfreezing more layers as needed.

4 MODEL PERFORMANCE

4.1 Experimental Setup

The pre-trained model that is used were made available online using the transformers library (Wolf et al., 2019) and the same have been fine-tuned for the training of the model. To examine

Table 2. Hyperparameters

Hyperparameter	Value
Epochs	3
Batch Size	32
Learning Rate	9e-06
Max Sequence Length	64
Warmup Proportion	0.1
Optimizer	Adam

Table 3. Performance of XLM-R_{large} on SemEval2017-Task4A

Model	AvgRec	$F1^{PN}$	Acc
BB_twtr [44] (Leader SemEval-2017, TaskA)	68.1	68.5	65.8
DataStories [45] (Leader SemEval-2017, TaskA)	68.1	67.7	65.1
State-of-the-art [43]	73.2	72.8	71.7
XLM-R_{large}	**73.5**	72.3	**71.8**

the effectiveness of our classification model, we tested it on two datasets, namely, the IIT-Patna Movie and Product review datasets. The training dataset was SemEval-2017 Task A. We divided the SemEval 2017 Task A English dataset into a training set and a validation set using 80:20 split and primarily fine-tuned the learning rate and number of epochs of the classification model manually to obtain the best results for the validation set. Training XLM-R on the SemEval-17 data required nearly 30 min. Nvidia Tesla K80 GPU to train the models. Early stopping was done if the evaluation loss did not improve over ten evaluation rounds. Table 2 depicts the hyperparameter setting.

4.2 Baselines and Evaluation

To ensure that the XLM-R achieves state-of-the-art performance, we compared its effectiveness to the existing best reported performance on both the Hindi language datasets. The metrics used for evaluation were average recall, $F1^{PN}$ and Macro F-score. The best performing system on the SemEval 2017 task (English language) dataset is the one described in Reference [43], which achieved a macro-average recall of 73.2. There were two winners [44, 45] of the SemEval 2017 Task A with a macro-average recall of 68.1. The focus of this research was to implement XML-R for sentiment analysis in Hindi language (non-English data) and not to outperform the SemEval-2017 state-of-the-art (SOTA). Interestingly, we achieved superlative results than the winners of task and also better results to the SOTA. Table 3 depicts the comparison of SemEval-2017 sentiment analysis task on English language.

Akhtar et al. [38] reported an average performance accuracy of 51.55 on both the Hindi datasets. They proposed a hybrid deep learning CNN-SVM (W+X) model trained and evaluated using word embeddings and extra hand-crafted features. The SOTA was recently given by Kunchukuttan et al. [16], reporting a 54.64 average accuracy on both datasets. The transfer learning with XLM-R cross-lingual embeddings provided the best results achieving an average performance accuracy of 60.93 on both the Hindi datasets. Table 4 depicts the comparison of IIT-P: Movie and product sentiment analysis datasets on Hindi language.

Figures 4 and 5 depict the precision-recall bar graph for both the Hindi datasets.

We take a closer look at the predictions made by XLM-R with zero-shot TL with the help of heatmap visualization for both the Hindi datasets as shown in Figures 6 and 7. The rows show the

Table 4. Accuracy of XLM-R_{large} with Zero-shot TL on IIT-P, Movie and Product Hindi Dataset

Model	IITP-Movie_{Acc}	$\text{IITP-Product}_{Acc}$
Pre-trained word embeddings using the IndicNLP corpus (SOTA) [16]	45.81	63.48
Review_{SH} and Movie_{H} [38]	44.88	57.34
XLM-R_{large} with Zero-shot TL	**51.74**	**70.12**

Fig. 4. Precision-recall of IITP-movie dataset.

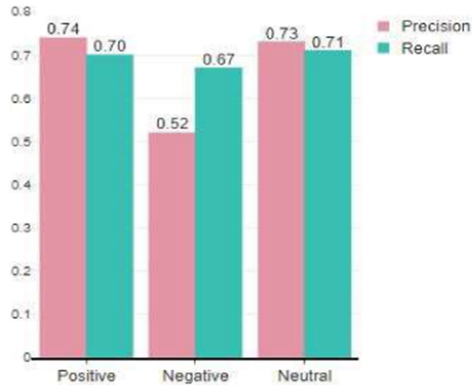

Fig. 5. Precision-recall of IITP-product dataset.

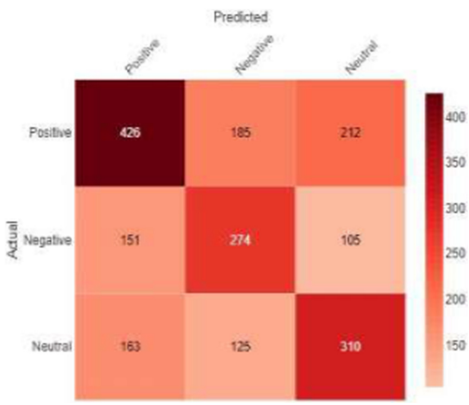

Fig. 6. HeatMap confusion matrix of IITP-movie dataset.

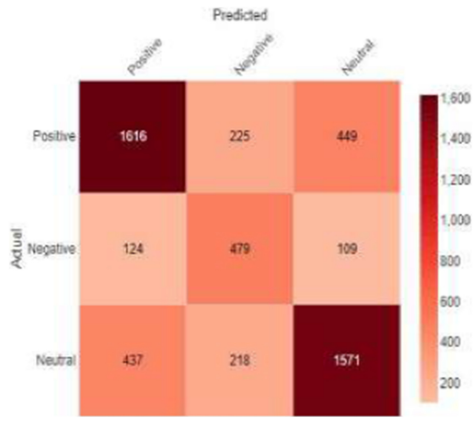

Fig. 7. HeatMap confusion matrix of IITP-product dataset.

actual class of a repetition and columns show the classifier's prediction of the respective confusion matrix.

5 CONCLUSION

Huge volumes of unstructured, multi-lingual, and multi-modal data is created on social media every day. Undeniably, social intelligence on these platforms is continuously increasing but the heterogeneity of content makes it hard to analyze and understand sentiments. This work uses the

existing cross-lingual transformer model, XLM-R as the pre-training model to provide an effective solution to sentence-level (tweet-level) analysis of sentiments in a resource-poor scenario with the help of zero-shot transfer learning. Both the network structure and weights are fine-tuned on benchmark English dataset and tested on two benchmark Hindi datasets. Interestingly, XLM-R achieves an average recall of 73.5 and accuracy of 71.8 for the English benchmark dataset, which is superlative to the best performing models. XLR-M outperforms the SOTA on Hindi datasets too. Future work included considering more Hindi test datasets for classification. Also, currently the training and test dataset are for sentence-level sentiment analysis, evaluation of XLM-R for aspect-based sentiment analysis is also a prominent direction of future research. Last, the model can be tested on other regional Indian languages such as Kannada, Marathi, or Sindhi.

REFERENCES

[1] Bing Liu. 2012. Sentiment analysis and opinion mining. *Synth. Lect. Hum. Lang. Technol.* 5, 1 (2012), 1–167.

[2] Akshi Kumar, Kathiravan Srinivasan, Wen-Huang Cheng, and Albert Y. Zomaya. 2020. Hybrid context enriched deep learning model for fine-grained sentiment analysis in textual and visual semiotic modality social data. *Info. Process. Manage.* 57, 1 (2020), 102141.

[3] Akshi Kumar and Arunima Jaiswal. 2020. A deep swarm-optimized model for leveraging industrial data analytics in cognitive manufacturing. *IEEE Trans. Industr. Info.* 17, 4 (2020), 2938–2946. doi: 10.1109/TII.2020.3005532

[4] Santwana Chimalamarri, Dinkar Sitaram, and Ashritha Jain. 2020. Morphological segmentation to improve crosslingual word embeddings for low resource languages. *ACM Trans. Asian Low-Resource Lang. Info. Process.* 19, 5 (2020), 1–15. https://doi.org/10.1145/3390298

[5] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2018. Meld: A multimodal multi-party dataset for emotion recognition in conversations. Retrieved from https://arXiv:1810.02508.

[6] Md Shad Akhtar, Asif Ekbal, and Erik Cambria. 2020. How intense are you? predicting intensities of emotions and sentiments using stacked ensemble. *IEEE Comput. Intell. Mag.* 15, 1 (2020), 64–75.

[7] Maria Pontiki, Dimitrios Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad Al-Smadi, Mahmoud Al-Ayyoub et al. 2016. Semeval-2016 task 5: Aspect-based sentiment analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval'16)*. 2016.

[8] Ning Liu and Bo Shen. 2020. Aspect-based sentiment analysis with gated alternate neural network. *Knowl.-Based Syst.* 188 (2020), 105010.

[9] Fazel Keshtkar and Diana Inkpen. 2009. Using sentiment orientation features for mood classification in blogs. In *Proceedings of the International Conference on Natural Language Processing and Knowledge Engineering*. IEEE, 1–6.

[10] Mahmoud Al-Ayyoub, Abed Allah Khamaiseh, Yaser Jararweh, and Mohammed N. Al-Kabi. 2019. A comprehensive survey of arabic sentiment analysis. *Info. Process. Manage.* 56, 2 (2019), 320–342.

[11] Majdi Beseiso and Haytham Elmousalami. 2020. Subword attentive model for Arabic sentiment analysis: A deep learning approach. *ACM Trans. Asian Low-Resource Lang. Info. Process.* 19, 2 (2020), 1–17.

[12] Asad Khattak, Muhammad Zubair Asghar, Anam Saeed, Ibrahim A. Hameed, Syed Asif Hassan, and Shakeel Ahmad. 2021. A survey on sentiment analysis in Urdu: A resource-poor language. *Egypt. Info. J.* 22, 1 (2021), 53–74.

[13] Valentin Barriere and Alexandra Balahur. 2020. Improving sentiment analysis over non-english tweets using multilingual transformers and automatic translation for data-augmentation. Retrieved from https://arXiv:2010.03486.

[14] Wenhuan Wang, Bohan Li, Ding Feng, Anman Zhang, and Shuo Wan. 2020. The OL-DAWE Model: Tweet polarity sentiment analysis with data augmentation. *IEEE Access* 8 (2020), 40118–40128.

[15] De Leon, Frances Adriana Laureano, Florimond Guéniat, and Harish Tayyar Madabushi. 2020. CS-embed-francesita at semeval-2020 Task 9: The effectiveness of code-switched word embeddings for sentiment analysis. Retrieved from https://arXiv:2006.04597.

[16] Anoop Kunchukuttan, Divyanshu Kakwani, Satish Golla, Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. AI4Bharat-IndicNLP Corpus: Monolingual corpora and word embeddings for indic languages. Retrieved from https://arXiv:2005.00085.

[17] K Karthikeyan, Zihan Wang, Stephen Mayhew, and Dan Roth. 2020. Cross-lingual ability of multilingual BERT: An empirical study. In *Proceedings of the International Conference on Learning Representations (ICLR'20)*.

[18] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. Retrieved from https://arXiv:1911.02116.

[19] Kumar Akshi and Geetanjali Garg. 2019. Systematic literature review on context-based sentiment analysis in social multimedia. *Multimedia Tools Appl.* 79, 21 (2019), 15349–15380.

[20] Jean-Benoit Delbrouck, Noé Tits, Mathilde Brousmiche, and Stéphane Dupont. 2020. A transformer-based joint-encoding for emotion recognition and sentiment analysis. Retrieved from https://arXiv:2006.15955.

[21] José Ángel González, Lluís-F. Hurtado, and Ferran Pla 2020. Self-attention for Twitter sentiment analysis in Spanish. *J. Intell. Fuzzy Systems* 39, 2 (2020), 2165–2175.

[22] Ahmed Sultan, Mahmoud Salim, Amina Gaber, and Islam El Hosary. 2020. WESSA at SemEval-2020 Task 9: Code-mixed sentiment analysis using transformers. Retrieved from https://arXiv:2009.09879.

[23] Y Kuratov, M. Arkhipov. 2019. Adaptation of deep bidirectional multilingual transformers for Russian language. Retrieved from https://arXiv:1905.07213.

[24] Anindya Sarkar, Sujeeth Reddy, and Raghu Sesha Iyengar. 2019. Zero-shot multilingual sentiment analysis using hierarchical attentive network and BERT. In *Proceedings of the 3rd International Conference on Natural Language Processing and Information Retrieval (NLPIR'19)*. Association for Computing Machinery, New York, NY, 49–56. DOI: https://doi.org/10.1145/3342827.3342850

[25] Avishek Garain, Sainik Kumar Mahata, and Dipankar Das. 2020. JUNLP@ SemEval-2020 Task 9: Sentiment analysis of Hindi-English code mixed data using grid search cross validation. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*. 1276–1280. https://arxiv.org/abs/2007.12561.

[26] Somnath Banerjee, Sahar Ghannay, Sophie Rosset, Anne Vilnat, and Paolo Rosso. 2020. LIMSI_UPV at SemEval-2020 Task 9: Recurrent convolutional neural network for code-mixed sentiment analysis. Retrieved from https://arXiv:2008.13173.

[27] Parth Patwa, Gustavo Aguilar, Sudipta Kar, Suraj Pandey, Srinivas PYKL, Björn Gambäck, Tanmoy Chakraborty, Thamar Solorio, and Amitava Das. 2008. Semeval-2020 task 9: Overview of sentiment analysis of code-mixed tweets. Retrieved from https://arxiv.org/abs/2008.04277.

[28] Yaman Kumar, Debanjan Mahata, Sagar Aggarwal, Anmol Chugh, Rajat Maheshwari, Rajiv Ratn Shah. 2019. BHAAV—A text corpus for emotion analysis from Hindi stories. Retrieved from https://arXiv:1910.04073.

[29] Kanika Garg and D. K. Lobiyal. 2020. Hindi EmotionNet: A scalable emotion lexicon for sentiment classification of Hindi text. *ACM Trans. Asian Low-Resource Lang. Info. Process.* 19, 4 (2020), 1–35.

[30] A. Bakliwal, P. Arora, and V. Varma. 2012. Hindi subjective lexicon: A lexical resource for Hindi polarity classification. *Int. J. Comput. Linguist. Appl. (IJCLA)* 2012

[31] A Balamurali, R. Joshi, A, and P. Bhattacharyya. 2012. Cross-lingual sentiment analysis for Indian languages using linked wordnets. In *Proceedings of the International Conference on Computational Linguistics (COLING'12)*.

[32] Braja Gopal Patra, Dipankar Das, Amitava Das, and Rajendra Prasath. 2015. Shared task on sentiment analysis in Indian languages (SAIL) tweets—An overview. In *Proceedings of the International Conference on Mining Intelligence and Knowledge Exploration (MIKE'15)*. Springer.

[33] Se Shriya, R. Vinaya Kumar, M. Anand Kumar, and K. P. Soman. 2015. AMRITA-CEN@SAIL2015: Sentiment analysis in Indian languages. In *Proceedings of the International Conference on Mining Intelligence and Knowledge Exploration (MIKE'15)*. Springer.

[34] A. Kumar, S. Kohail, A. Ekbal, and C. Biemann. 2015. IIT-TUDA: System for sentiment analysis in Indian languages using lexical acquisition. In *Proceedings of the International Conference on Mining Intelligence and Knowledge Exploration (MIKE'15)*. Springer.

[35] M. S. Akhtar, A. Ekbal, and P. Bhattacharyya. 2016. Aspect-based sentiment analysis in Hindi: Resource creation and sentiment classification. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC'16)*.

[36] Shad Akhtar, Palaash Sawant, Sukanta Sen, Asif Ekbal, and Pushpak Bhattacharyya. 2018. Solving data sparsity for aspect-based sentiment analysis using cross-linguality and multi-linguality. In *Proceedings of the 16th Annual Conference of the NAACL on Human Language Technologies (HLT'18)*. 572–582.

[37] Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. 2015, Bilingual word representations with monolingual quality in mind. In *Proceedings of the NAACL Workshop on Vector Space Modeling*.

[38] M. S. Akhtar, A. Kumar, A. Ekbal, and P. Bhattacharyya. 2016. A hybrid deep learning architecture for sentiment analysis. In *Proceedings of the 26th International Conference on Computational Linguistics (COLING'16)*. 482–493.

[39] Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune BERT for text classification? In *Chinese Computational Linguistics, Maosong Sun*, Xuanjing Huang, Heng Ji, Zhiyuan Liu, and Yang Liu (Eds.). 194–206

[40] Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot cross-lingual transfer with multilingual transformers. Retrieved from https://arXiv:2005.00633.

[41] Jie, Tao and Xing Fang. 2020. Toward multi-label sentiment analysis: A transfer learning-based approach. *J. Big Data* 7, 1 (2020), 1–26.

[42] Sultan Ahmed, Mahmoud Salim, Amina Gaber, and Islam El Hosary. 2020. WESSA at SemEval-2020 Task 9: Code-mixed sentiment analysis using transformers. Retrieved from https://arXiv:2009.09879.

[43] Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English tweets. Retrieved from https://arXiv:2005.10200.

[44] Mathieu Cliche. 2017. BB twtr at SemEval-2017 Task 4: Twitter sentiment analysis with CNNs and LSTMs. *Proceedings of the 11th International Workshop on Semantic Evaluations (SemEval'17)*. 573–580.

[45] C. Baziotis, N. Pelekis, and C. Doulkeridis. 2017. DataStories at SemEval-2017 Task 4: Deep LSTM with attention for message-level and topic-based sentiment analysis. *Proceedings of the 11th International Workshop on Semantic Evaluations (SemEval'17)*. 747–754.