Available online at: http://www.ijmtst.com/vol6issue04.html

International Journal for Modern Trends in Science and Technology

ISSN: 2455-3778 :: Volume: 06, Issue No: 04, April 2020

Perlustration on Hindi News Classification using Transfer Learning

Desai Eesha C[1] | Mosin I Hasan[1] | Hemant D Vasava[1]

[1]Computer Engineering, Birla Vishvakarma Mahavidyalaya, Anand, Gujarat, India

To Cite this Article

Desai Eesha C, Mosin I Hasan and Hemant D Vasava, "Perlustration on Hindi News Classification using Transfer Learning", International Journal for Modern Trends in Science and Technology, Vol. 06, Issue 04, April 2020, pp.:175-181.

Article Info

Received on 11-March-2020, Revised on 09-April-2020, Accepted on 12-April-2020, Published on 14-April-2020.

ABSTRACT

There is an enormous amount of data accessible in a virtual space, to make that data expedient; there is an urge for text classification. Due to technological expansions, there have been many classification models for text. There are certain monotonous situations where building a model data is not enough available or for that matter to build a model from scratch is time-consuming. To overcome that problem, transfer learning is proposed. Transfer learning is a machine learning method that is useful when we have a less amount of data for a given task for training and testing data. The knowledge gained while training a source dataset is used to train a target dataset. This article contains a summary of transfer learning and its instance cross-lingual learning which can be helpful for Hindi news classification.

KEYWORDS: *Text classification, Transfer Learning, Hindi News, Natural Language Processing (NLP), Word Embedding, Cross-lingual learning*

I. INTRODUCTION

In today's era, when everything is going digital there is a large amount of data available in data space. One of the versatile forms of data is news. The news is spread widely due to technological advancements and it influences people to a great extent. The news classification is an important aspect when there is a processing of news information, it can help distinguish news according to its category and even help to organize. It is even helpful for preference or relevance. With the help of NLP, text classification can be categorized based on its content. There are many factors that need to be kept in mind while applying NLP an introduction about the factors and under what conditions it can be used is explained in this section.

A. Natural Language Processing (NLP)

Humans are understood to be one of the most developed species because of their ability to communicate. There are 6500 dialects that we use for communication whereas the machine uses zeroes and ones to communicate. As we speak, write or tweet a huge amount of data is produced which is in its unstructured form. To make the acquired data useful it is important to get acquainted with text mining and NLP techniques. Text mining helps to derive significant meaning from natural language text. NLP is a part of computer science and artificial intelligence, which helps in communicating with human languages. NLP can be used in sentiment analysis, chat bot, speech recognition, Machine translation, spell checking and keyword searching, information

extraction, advertisement matching, etc. We humans can understand what words are of importance and what is not but machines don't understand the importance of words. To remove the words which cannot be helpful for classification or give the words it priority there are certain steps of NLP. The basic steps for NLP include tokenization, stemming, lemmatization, POS tags, Named Entity relationship, chunking[1].

B. *Word Embedding*

To make text classification a possibility after removing all the not so useful words and keeping all useful words now we want to relate them with each other. But many deep learning and machine learning algorithms are not capable of processing strings or texts in their real form. They require numeric value as input to perform any operation. But there is a huge amount of data that is available in text format, thus word embeddings come into the picture. Word embedding maps a word using a dictionary to vector. It also helps in dimensionality reduction and contextual similarity organization. It can be done based on prediction or frequency[2]. There are pre-trained word embedding models available like word2vec, GloVe (Global Vector for Word Representation), fast Text. The pre-trained model uses a combination of techniques from the below categories or similar to them to get a better result.

There are generally three types that come under the frequency category:

1. Count Vector: Consider a corpus C, of multiple documents D {d1, d2...dD}, and N unique tokens are drawn out of the C that will form the dictionary. Let, the size of the Count Vector Matrix M be D*N. The row in the matrix will have the frequency of token in documents D(i).[3]

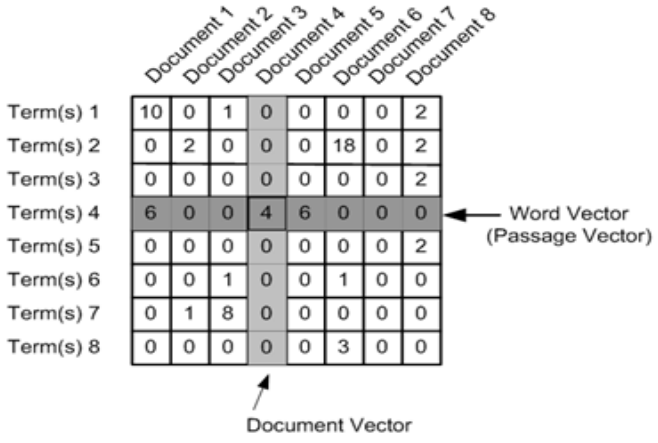

Figure 1: Matrix M representation[3]

2. TF-IDF Vector: The TF-IDF (Term Frequency-Inverse Document Frequency) uses the occurrence of the word not just in a single document but in the entire corpus to know that the so the frequent, which is not important for classification can be counted out. Where TF = (Number of times term t appears in a document)/ (Number of terms in the document) and IDF = log(N/n), where, N is the number of documents and n is the number of documents a term t has appeared[3].

3. Co-occurrence Vector: It describes how words appear together, which helps to capture the relationship between the words[3].

There are generally two types that come under the prediction category:

1. CBOW (Continuous Bag of Words): It learns to predict words from the set of words or from a word for a given target word[3].
2. Skip-gram Model: It learns to predict a word or set of words from a single word[3].

Word2Vec uses CBOW and Skip-gram model combination.

C. *Transfer Learning*

Humans have the ability to use our gained knowledge of different tasks. Whatever knowledge we acquire from one task we use it for the relevant task. For instance, if we know how to ride a bicycle we can easily learn how to ride a motorbike. Machine Learning and Deep learning algorithms have been very useful when classification, clustering or regression is considered. It works very well when we use it to train a specific task. But when there is a change in a feature the model is not useful[4]. A trained neural network gains knowledge from the data, which is the weight of the network. These weights can be extracted and then transferred to any neural network.

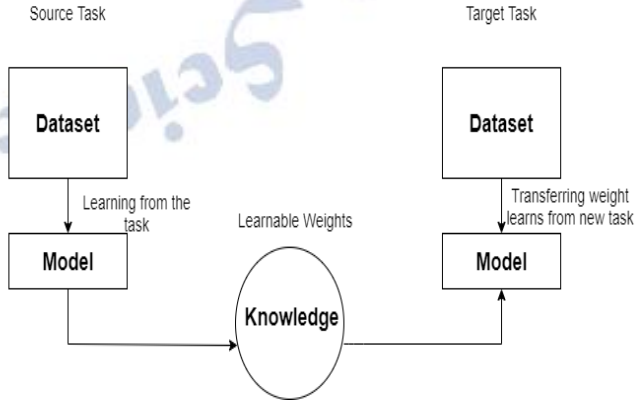

Figure 2: Transfer Learning[5]

D. Cross Lingual Learning

Transfer learning is used when the source task and target task has some difference in their base scenarios[6]. To train a particular dataset for a good outcome there is a requirement of a large amount of dataset. In some scenarios, a large amount of dataset is available. But in some cases, the dataset is not large enough for training and testing. In NLP with the help of Cross-Lingual transfer, we can use one or more similar languages with good resources to improve the performance of the less resource task.

II. Existing Research on NLP and it's Technique

Text classification is useful because of the large amount of information available on the webspace. To make the data useful it is important to classify them in categories. Traditional methods use models that need to be created from scratch but with the help of transfer learning the problem of creating a model from scratch can be solved. Similarly, Cross-lingual transfer is capable of transfer language to improve the accuracy of a low dataset task language, which is a tool for improving the performance of NLP on low dataset languages.

A. Natural Language Processing

JingJing Cai et al.[7]in the year 2018, proposed the idea of using deep learning and its advantages over the traditional method. There are three models proposed in the paper Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), RNN+Attention for classification. The dataset used in the experiment is Sohu news data, the results show that the RNN model is suitable for short text processing because of its dependent nature whereas the CNN model is used for long text processing because of its parallel running property. However, according to the type of task and the dataset of the task the model should be used. The comparison result is shown in the prediction result of the test set table.

Chenbin Li et al.[8]in the year 2018, proposed the uses of the Bi-LSTM model which is used to obtain the representation of two directions, and then the two directions representation through the convolution neural network. Each word expression is added by itself to the left text vector and the right text vector to indicate. For the left and right texts, a loop structure is used, which is a non-linear transformation of the previous word and a text on the left side. This approach preserves contextual information and a wider range of word order better. The paper uses the CBOW model for the training of word vectors. The data set used in this experiment is a subset of THUC News for training and testing. The classification method TF-IDF, SVM, CNN, and LSTM are compared. The result is shown in the comparison experiment result table.

Mazhar Iqbal Rana et al.[9]in the year 2014, discuss the ways in which news can be classified based on their headlines. There are some existing classifiers whose workings, advantages, disadvantages, and results are discussed in the news headline classification section. The process includes data collection; document indexing, pre-processing, feature selection, classification techniques, application and evaluating performance measures. Text pre-processing is done by removing all kinds of noisy and useless information, headlines tokenization, removal from the existing list, removal on the basis of word frequency, word stemming. It is observed that the classification process remains the same only minor changes have been recorded in feature selection and pre-processing. The accuracy factor and statistical calculation have been reduced.

Sandeep Kaur et al.[10]in the year 2016,discusses the implementation ofthe neural network in predicting the popularity of online news. The classification method used in the paper here is the neural network. First, all the documents are classified as on what categories they belong to after that the dataset is split into two. One is the training and the other is validation. Then the accuracy on the validation set of the classifier built on the training set is checked. The experiment result indicates that the neural network is better than the traditional method.

B. Word Embedding

Jeffrey Pennington et al.[11]in the year 2014, discussed the GloVe model (Global Vector for Word Representation) for word embedding which produces a vector space with meaningful substructure (Figure 3). The aim of the model is to capture the meaning in vector space by creating word vectors and using global count statics rather than any local information. The model learns on the co-occurrence matrix and predicts the co-occurrence ratio by training word vectors. The result shows that it is a better approach to word embeddings and is more focused on the way word2vec and embedding works.

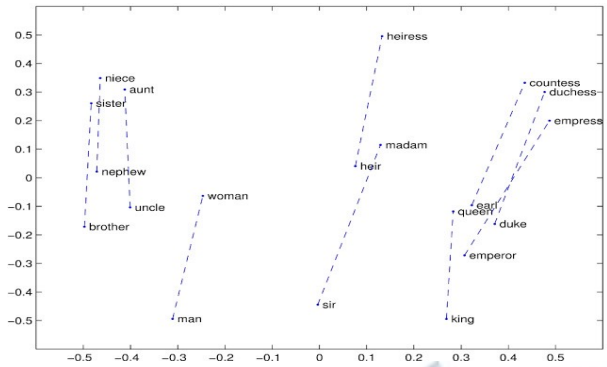

Figure 3: Linear Substructure[12]

Jacob Devlin et al.[13]in the year 2018, introduced a language representation model BERT which stands for Bidirectional Encoder Representation from Transformers. BERT is based on the transformer architecture.It is intended to pre-train deep bidirectional representations from the unlabelled text by training together on both left and right context in all layers. As a result, the pre-trained BERT model can be fine-tuned with just one further output layer to create state-of-the-art models for a wide range of tasks, such as question answering and language interpretation, without considerable task-specific architecture modifications. It is pre-trained on a large corpus of unlabelled text. BERT is currently available in two types: BERT Base and BERT Large The base has 12 layers of transformer blocks, 12 attention heads, and 110 million parameters whereas the Large has 24 layers of transformer blocks, 16 attention heads, and 340 million parameters.

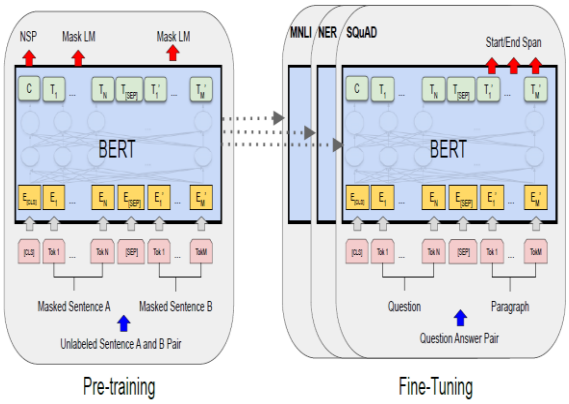

Figure 4: BERT[13]

Figure 4 represents the pre-training and fine-tuning procedures for BERT. Everything is the same in the architecture of pre-training and fine-tuning only there is an output layer difference. The pre-trained model parameters are the same to initialize models for different down-stream tasks. During fine-tuning, all parameters are fine-tuned. CLS(Classification) is a special symbol added in front of every input example, and SEP is a special separator token.

C. *Transfer Learning*

Abdul Kawsar Tushar et al.[5]in the year 2017, discussed the model for the recognition of handwritten numerals in Indic Language. It uses Convolution Neural Network with backpropagation for error reduction and dropout for data overfitting. The model uses CMATERDB 3.1.1, CMATERDB 3.2.1, CMATERDB 3.3.1 dataset of Bangla, Hindi and Urdu respective samples of handwritten numerals from various writers. The result shows that independent model-specific training helps to reduce the re-training time in target tasks, and by reducing the overfitting problem better accuracy is achieved with the help of experiments because of which there is a performance gain that can be very path-breaking in transfer learning.

Sinno Jialin Pan et al.[14] in the year 2010, have done a survey on Transfer Learning which reviews and categorizes the progress of Transfer Learning for classification, clustering, and regression problems and the relationship between machine learning techniques and transfer learning is discussed. Transfer learning is classified into three main categories Inductive Transfer Learning, Transductive Transfer Learning and Unsupervised Transfer Learning as shown in Figure 5. They are further classified on based of what need to be transfer. The paper discusses the disadvantage of Transfer Learning which is Negative Transfer Learning why it can create a problem and how to solve it. It also hints that in the future, transfer learning techniques can be widely used to solve other challenging applications, such as video classification,social network analysis, and logical inference.

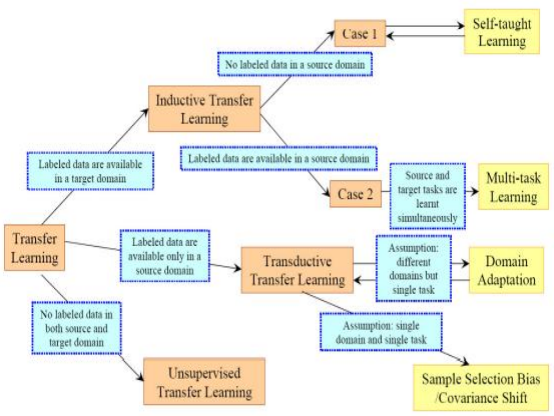

Figure 5: Transfer Learning Strategies[14]

Feng Yu, et al.[15]in the year 2010, discussed the transfer learning approach for text categorization. The proposed model will learn transfer knowledge from different category data and then different classifiers will be constructed which will categorize tasks with the help of transfer knowledge. They compared with two baseline systems, one is multiple classifiers fusion methods, and another baseline system is a flat classification under a top-down level-based approach. The experiment shows that transfer learning gets a better result.

D. Cross Lingual Learning

Yu-Hsiang Lin et al.[16]in the year 2019, discussed that Cross-lingual transfer is a high-resourcetransfer language helpful for improving the accuracy of a low dataset task language, which is a helpful tool for improving the performance of NLP on lower dataset languages. Theproposed model considers the aforementioned featuresto perform the prediction of language. In experiments on representative NLP tasks, the model predicts good transfer languagesmuch better than ad hoc baselines consideringsingle features in isolation, and glean insightson what features are most informativefor each different NLP tasks, which may informfuture ad hoc selection even without the use of the proposed method. Figure 6 represents the workflow of how to learn to select the transfer languages for an NLP task. $L_{tf,1}$is used to train a set of NLPmodels with all available transfer languages and collect evaluation scores and $L_{tf,2}$ train a ranking model to predictthe top transfer languages.

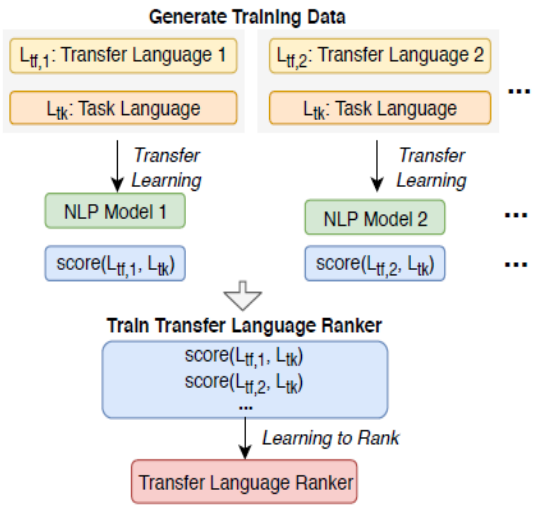

Figure 6: Workflow of learning to select the transfer Language for an NLP Task[16]

Takashi Wada et al.[17]in the year 2018, proposed a model called multi-lingual neural language model, which produces cross-lingual word embeddings in an unsupervised way. Figure 6 represents the multilingual model that contains bidirectional LSTMs that perform as forward and backward language models, and the networks are shared between all the Languages and Word embeddings and linear transformation among hidden states and outputs are definite to every single language. The shared LSTMs capture the common sentence structure among all languages and accordingly, word embeddings of each language are mapped into a common latent space, making it possible to measure the similarity of words across multiple languages. On a word, alignment task evaluation of the quality of the cross-lingual word embeddings is done. The experiments demonstrate that the model can obtain cross-lingual embeddings of much higher quality than existing unsupervised models when only a small amount of monolingual data is available, or the domains of monolingual data are diverse through languages. Shared parameters are $\vec{f}$ (forward) and $\overleftarrow{f}$ (backward) LSTMs, E^{BOS} -initial input to the language model, W^{EOS}-calculates how likely the next word is the end of a sentence.Separate parameters are E^l – word embeddings of language l, W^l – used to calculate the probability distribution of the next word.

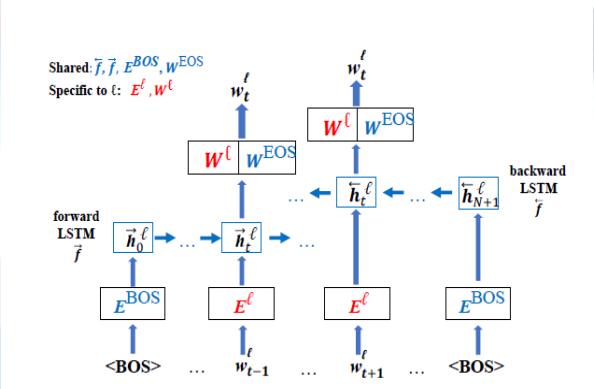

Figure 7: Multilingual neural network model[17]

Multilingual neural network formula

$$\sum_{l=1}^{L}\sum_{i=1}^{S^l}\sum_{t=1}^{N^i} log\, p(w_{i,t}^l | w_{i,1}^l, w_{i,2}^l \dots w_{i,t-1}^l ; \overrightarrow{\theta}) + \log p(w_{i,t}^l | w_{i,t+1}^l, w_{i,t+2}^l \dots w_{i,N^i}^l ; \overleftarrow{\theta})$$ [17] **(1)**

Here L= number of language, S^l= Sentences of languages, $\overrightarrow{\theta}$ and $\overleftarrow{\theta}$ denotes the parameter of forwarding and backward LSTMs $\vec{f}$ and $\overleftarrow{f}$ respectively.

Guillaume Lample et al.[18]in the year 2019, discusses the approach to use generative pre-training for multiple languages and shows the effectiveness of cross-lingual pre-training. Two

methods have been proposed to learn cross-lingual language models (XLMs): one is unsupervised which relies only on monolingual data, and another one is supervised that leverages parallel data with a new cross-lingual language model objective. The results have shown the strong impact of cross-lingual language model (XLM) pre-training. Two unsupervised training objectives that require only monolingual corpora: Causal Language Modelling (CLM) and Masked Language Modelling (MLM) approaches have provided strong cross-lingual features that can be used for pre-training models. On unsupervised machine translation, MLM pre-training is extremely effective.

Shahnawaz Khan et al.[19]in the year 2019, proposed a multilingual machine translation system that translates from English to Urdu and Hindi and uses a translation rules-based approach with an artificial neural network. There are various methods used in the system like n-gram bleu score, F-measure, Meteor and precision and recall is used for Machine translation and evaluation score for translation output obtained from the system for around 500 Hindi test sentences. Translation Results obtained from the system evaluated using machine translation evaluation methods and manually indicates that the system works efficiently on the trained linguistic (translation) rules and bilingual dictionaries which further show that the efficient and accurate machine translation system can be achieved by enhancing the grammar rules and size of bilingual dictionary. They even pointed out that if case marking is improved then the results will be more efficient.

III. Conclusion

There are many dialects around the globe. In a country like India, around 720 dialects are used. One of the major languages used other than English is Hindi, thus there are many major enlargements that can be done in Hindi Languages that can be useful in NLP. Classification can be helpful in news cataloguing, Fake news detection, summary, unnecessary reverberation of news, etc. Appropriate news cataloguing solves the predicamentof news information which is widely available now because of digitization. There are outmoded methods that require the training on a large dataset and then use it for classification which is a time-consuming task. In this paper, cross-lingual learning is offered for the Hindi news classification. The paper reconnoitred a literature survey on the existing methods in NLP, word embeddings, transfer learning, and cross-lingual language and explores the advantages of transfer learning over traditional language.

References

[1] Singh M (2018) Word embedding. In: Medium. https://medium.com/data-science-group-iitr/word-embedding-2d05d270b285. Accessed 29 Nov 2019

[2] (2017) Understanding Word Embeddings: From Word2Vec to Count Vectors. In: Anal. Vidhya. https://www.analyticsvidhya.com/blog/2017/06/word-embeddings-count-word2veec/. Accessed 6 Jan 2020

[3] Sarkar D (DJ) (2018) A Comprehensive Hands-on Guide to Transfer Learning with Real-World Applications in Deep Learning. In: Medium. https://towardsdatascience.com/a-comprehensive-hands-on-guide-to-transfer-learning-with-real-world-applications-in-deep-learning-212bf3b2f27a. Accessed 4 Jan 2020

[4] Tushar AK, Ashiquzzaman A, Afrin A, Islam MdR (2018) A Novel Transfer Learning Approach upon Hindi, Arabic, and Bangla Numerals Using Convolutional Neural Networks. In: Hemanth DJ, Smys S (eds) Computational Vision and Bio Inspired Computing. Springer International Publishing, Cham, pp 972–981

[5] (2019) Unsupervised Cross-lingual Representation Learning. In: Sebastian Ruder. https://ruder.io/unsupervised-cross-lingual-learning/. Accessed 4 Jan 2020

[6] Cai J Deeplearning Model Used in Text Classification. 4

[7] Li C, Zhan G, Li Z (2018) News Text Classification Based on Improved Bi-LSTM-CNN. In: 2018 9th International Conference on Information Technology in Medicine and Education (ITME). IEEE, Hangzhou, pp 890–893

[8] Rana MI, Khalid S, Akbar MU (2014) News classification based on their headlines: A review. In: 17th IEEE International Multi Topic Conference 2014. IEEE, Karachi, Pakistan, pp 211–216

[9] Kaur S, Khiva NK Online news classification using Deep Learning Technique. 03:6

[10] Pennington J, Socher R, Manning C (2014) Glove: Global Vectors for Word Representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, Doha, Qatar, pp 1532–1543

[11] What is GloVe? - Japneet Singh Chawla - Medium. https://medium.com/@japneet121/word-vectorization-using-glove-76919685ee0b. Accessed 5 Jan 2020

[12] Devlin J BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 35

[13] Pan SJ, Yang Q (2010) A Survey on Transfer Learning. IEEE Trans Knowl Data Eng 22:1345–1359. https://doi.org/10.1109/TKDE.2009.191

[14] Yu F, Wang H, Zheng D, Fei G (2010) Research on Transfer Learning Approach for Text Categorization. In: 2010 International Conference on Artificial Intelligence and Computational Intelligence. IEEE, Sanya, China, pp 418–422

[15] Lin Y-H, Chen C-Y, Lee J, et al (2019) Choosing Transfer Languages for Cross-Lingual Learning. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Florence, Italy, pp 3125–3135

[16] Wada T, Iwata T (2018) Unsupervised Cross-lingual Word Embedding by Multilingual Neural Language Models. ArXiv180902306 Cs

[17] Lample G, Conneau A (2019) Cross-lingual Language Model Pretraining. ArXiv190107291 Cs

[18] Khan S, Usman I (2019) AModel for English to Urdu and Hindi Machine Translation System using Translation Rules and Artificial Neural Network. 16:7

[19] Z. Li, W. Shang, and M. Yan, "1HZV_7H[W_&ODVVLILFDWLRQ_0RGHO_%DVHG_RQ_7RSLF_0RGHO_," p. 5.

[20] A review of BERT based models - Towards Data Science." [Online]. Available: https://towardsdatascience.com/a-review-of-bert-based-models-4ffdc0f15d58. [Accessed: 05-Jan-2020].

[21] "A review of BERT based models - Towards Data Science." [Online]. Available: https://towardsdatascience.com/a-review-of-bert-based-models-4ffdc0f15d58. [Accessed: 05-Jan-2020].

[22] A. Patra and D. Singh, "A Survey Report on Text Classification with Different Term Weighing Methods and Comparison between Classification Algorithms," *IJCA*, vol. 75, no. 7, pp. 14–18, Aug. 2013, doi: 10.5120/13122-0472.

[23] V. U. Suryawanshi, P. Bogawar, P. Patil, P. Meshram, K. Yadav, and N. S. Sakhare, "Automatic Text Classification System," vol. 4, no. 2, p. 5, 2015.

[24] J. Devlin, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," p. 35.

[25] F. Miao, P. Zhang, L. Jin, and H. Wu, "Chinese News Text Classification Based on Machine Learning Algorithm," in *2018 10th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC)*, Hangzhou, 2018, pp. 48–51, doi: 10.1109/IHMSC.2018.10117.

[26] G. Lample and A. Conneau, "Cross-lingual Language Model Pretraining," *arXiv:1901.07291 [cs]*, Jan. 2019.

[27] A. Søgaard, Ž. Agić, H. Martínez Alonso, B. Plank, B. Bohnet, and A. Johannsen, "Inverted indexing for cross-lingual NLP," 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, 2015, pp. 1713–1722, doi: 10.3115/v1/P15-1165.

[28] A. Maheshwari, "Report on Text Classification using CNN, RNN & HAN," *Medium*, 17-Jul-2018. [Online].Available:https://medium.com/jatana/report-on-text-classification-using-cnn-rnn-han-f0e887214d5f. [Accessed: 29-Nov-2019].

[29] Z. Wang and B. Song, "Research on hot news classification algorithm based on deep learning," in *2019 IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*, Chengdu, China, 2019, pp. 2376–2380, doi: 10.1109/ITNEC.2019.8729020.

[30] "Text Classification using CNN." [Online]. Available: https://kaggle.com/au1206/text-classification-using-cnn. [Accessed: 29-Nov-2019].

[31] "Transfer Learning Introduction Tutorials & Notes | Machine Learning | HackerEarth." [Online].Available:https://www.hackerearth.com/practice/machine-learning/transfer-learning/transfer-learning-intro/tutorial/. [Accessed: 29-Nov-2019].

[32] "Using Transfer Learning and Pre-trained Language Models to Classify Spam." [Online]. Available:https://heartbeat.fritz.ai/using-transfer-learning-and-pre-trained-language-models-to-classify-spam-549fc0f56c20.[Accessed: 11-Dec-2019].

[33] A. Mandelbaum and A. Shalev, "Word Embeddings and Their Use In Sentence Classification Tasks," *arXiv:1610.08229 [cs]*, Oct. 2016.

[34] J. S. Chawla, "Word Vectorization using GloVe," *Medium*, 24-Apr-2018. [Online]. Available:https://medium.com/@japneet121/word-vectorization-using-glove-76919685ee0b. [Accessed: 29-Nov-2019].