How to Cite:
Zade, N., & Ajani, S. (2022). Multilingual text classification using deep learning. *International Journal of Health Sciences*, *6*(S1), 10528–10536. https://doi.org/10.53730/ijhs.v6nS1.7539

Multilingual text classification using deep learning

Neha Zade
Department of Computer Science and Engineering, Shri amdeobaba College of Engineering and Management, Nagpur, India
Corresponding author email: zadenj@rknec.edu

Sameer Ajani
Department of Computer Science and Engineering, Shri Ramdeobaba College of Engineering and Management, Nagpur, India

***Abstract*---**Every living organism shares a common property to interact whether it is other animal or human being. To interact everyone use some kind of language. It might be a direct language like English, Marathi, Hindi or it can be signed language which includes some symbols or signs or gestures. Moral of the discussion is that for any type of communication the most important channel is language and for interchanging the information the most important thing recognition and understands of the language. Today with the help of technology human have made the non-living things like machines able to interact and speak. Mostly, the languages which are universally acceptable like English can be considered for normal mode of recognition by the machine which we can call it as Monolingual understanding or recognition but when we consider the diversity of languages that is used in entire world, we can imagine the complexity of the implementation. The main concern of proposed title is to focus on exploration of multilingual classification for regional languages like Marathi, Hindi; English to a better extend by using one of the most powerful ways of deep learning.

***Keywords*---**text classification, deep learning, Monolingual, Multilingual, corpora, Natural Language Processing.

Introduction

Deep learning can be referred as the one of the branch of machine learning which works on the data with more level of abstraction as compared to machine learning. While dealing with the classification of multiple languages, the data analysis may lead to more complexity in order to implement. If we look back into the history we can found that the linguistic research has started many years

back. The terms like corpus were present for research in linguistic domain. Actually Corpus (plural Corpora) is a collection of linguistic data present in the form of computer database. First systematically organized computer corpus is said to be the Brawn University Standard Corpus for Present day American English which is also commonly referred as Brawn Corpus compiled in 1960s. We can get many things available on internet regarding corpus. The tool like concerdancing is used to find the every occurrence of word or phrase.

Deep learning can be referred as the efficient way to implement the complex concepts. As far as the preference given to deep learning strategy in task like multilingual text classification it can be proved very convenient way to improve the accuracy of classification to a better extend by using the somewhat more explored approach of calculating the terms.

Related work

The work that we found as the preliminary starting of this domain was done in 1996 which was proposed by MichkleJardino LIMSI-CNRS, ORSAY, France. They worked on multilingual classification focusing mainly on some specific languages like French, German and English. The key term that was used in their proposed system was n-gram model in which stochastic approach was used for implementing classification.[1]

We can find many proofs in 19's which shows that at that time also research was going on lingual area. In 1997, TONY McENERY and ANDREW WILSON and team worked on a resource development for European Languages which was done for contributing in project of CRATER. The project was based on Spanish, French and English.[2]

Followed by this Michel Simard Universite de Montreal, Canada has written a book chapter on alignment of multilingual text in 2000. In his writing he had specifically heighted the various queries regarding multilingual text means a text which is represented using more than one language. When he analysed this domain at that time bilingual alignment was in more use as compared to trilingual or multilingual. He had also referred many previous projects in previous time like CRATER which was introduced around 1997, MULTEXT (1994) followed by MULTEXT-EAST in 1998. [3]

The research and curiosity in researchers in this field did not end there. Roberto Basili and Alessandro Moschitti came up with a new approach by proposing a robust feature in the process of natural language processing in 2001. They used language driven text classification which was in turn focuses on retrieval of text and enrichment of important information. They had also used concept of corpus driven extraction in their work. [4]

In immediate next year i.e. 2002 Rowena Chau and Chung-HsingYeh add a new direction to view towards the multilingual text classification by proposing the Fuzzy technique or knowledge discovery. The system proposed was based on concept based classification. The following diagram visualizes the overall structure of their work.[5]

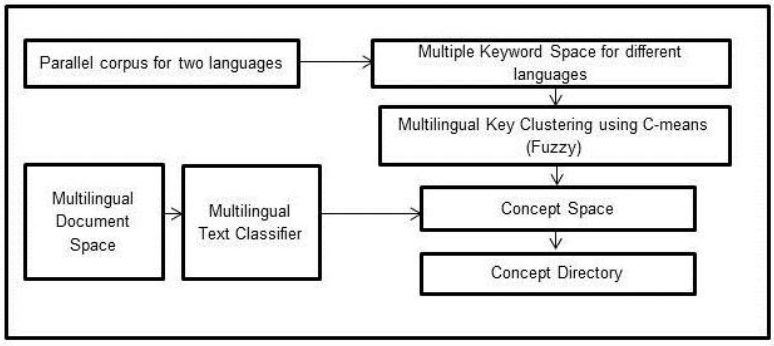


Fig. 1.1 concept-based multilingual text categorisation system

Some researchers like Barbara Plank worked on the type of model for multilingual text classification which can be all in one solution for multiple languages in 2017. They assured that it was a simple model which was developed specifically classification small text. The key term was support vector machine classifier for exploiting multilingual word embedding as well as character n-gram.[6] The proposed approach can be analysed in diagram 1.2

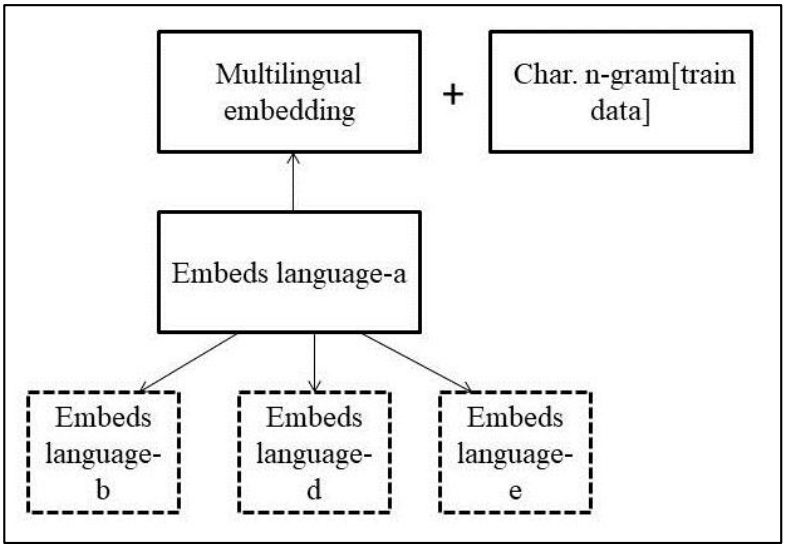


Fig. 1.2 All in one model

Nurendra Choudhary of International Institute of Information technology, Hyderabad worked on this domain of multilingual text classification and published a thesis on it in 2019. He has given a detailed analysis and terminologies like morphology analyser, clustering of word, experimental setup for its testing etc. [7]

Year of 2020 was witness of multilingual text classification method which was transformer based proposed by Sophie Groenwold and team. These people came

to a conclusion that difference in language modelling can also affect typological difference in language model. [8]

Multilingual classification is not limited to one use and it was proved by the work of Stephen Mutuvi and team who gave an comprehensive study focusing on text classification of different languages for specifically epidemiological study.[8]

Recently in 2021, SumanthDoddapaneni along with team from IIT madras done as survey on multilingual text classification models that already exist and gave some guidelines that what should be the focus of new work or future work in the domain of multiligual classification.[9]

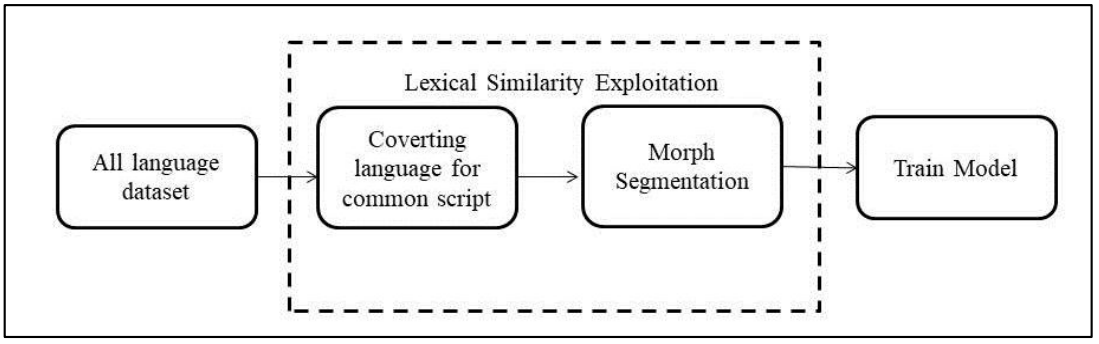


Fig. 1.3 Multiple Text Classification

As far as specifically Indian languages are concern, there is too much diversity of language in India. Some of the IIITiansSalil Aggarwal,Sourav Kumar, and Radhika Mamidi from Hyderabad have worked on this last year. They have proposed a method for text classification of Some Indian Languages based on their lexical similarity. They have also worked on zero shot classification. [10] The following diagram gives an overview of their system.

Corpus and dataset samples

A corpus is a representative sample of actual language production within a meaningful context and with a general purpose. A dataset is a representative sample of a specific linguistic phenomenon in a restricted context

Marathi WordNet:
Marathi WordNet is a database similar to the idea of English WordNet. It cannot be directly referred as dictionary. It specifies relations between synonym sets which represent unique concepts.

00000001 02 01 अजन्मा 0001 0400 00000101 | ज्यासजन्मनाहीअसा:"ईश्वरअजन्माआहे"
00000002 02 02 अशुभ:अमंगळ 0001 0400 00000101 | शुभनाहीअसा:"यायोगामूळेकुंडलीतीलइतरअशुभयोगांचानाशहोतो."
00000003 02 01 अप्रविष्ट 0001 0400 00000111 | ज्यानेप्रवेशकेलानाहीअसा:"अप्रविष्टव्यक्तींनाताबडतोबआतप्रवेशकरूद्या."

00000004 01 05 पुण्यभूमी:पवित्रभूमी:पुण्यस्थान:पवित्रस्थान:पावनस्थान 0001 0400 00000037 | पवित्रमानलेगेलेलेस्थान:"हिंदूंसाठीकाशीहीपुण्यभूमीआहे."
00000005 01 02 शिवालय:शिवमंदिर 0001 0400 00000037 | जिथेशंकराच्यापिंडीचीस्थापनाकेलीअसूनत्याचीपूजाहोतेतीजागा:"तीदरसोमवारीशिवालयातजाते."
00000006 01 01 अपवित्र_स्थान 0001 0400 00000037 | असेस्थानजेपवित्रमानलेजातनाही:"धार्मिकविचारांनुसारभुतेखेतेअपवित्रस्थानीचआढळतात. "
00000007 02 02 आलेला:आगत 0001 0400 00000101 | दाखलझालेला:"आलेल्यापाहुण्यांचेस्वागतआहे."
00000008 02 03 जन्मलेला:जलमलेला:जल्मलेला 0001 0400 00000101 | ज्यानेजन्मघेतलाआहेअसा:"जन्मलेल्याप्राणीचामृत्यूनिश्चितआहे."
00000009 01 05 सत्कर्म:सुकृत:नैतिक_कार्य:सुकृत्य:चांगले_कार्य 0001 0400 00000058 | असेकार्यजेनीतिलाधरूनकेलेलेअसतेआणिज्यातसमाजाचेहीतअसते:"सत्कर्मानेसमाजाचाखराविकासहोतो."
00000010 02 02 उत्पादित:उत्पन्न 0001 0400 00000101 | ज्याचीउत्पत्तिझालीआहेअसा:"आसाममध्येउत्पादितचहाजगभरप्रसिद्धआहे."
00000011 01 02 चारित्र्यहीनता:दुश्चरित्रता 0001 0400 00000079 | चारित्र्यवाईटअसण्याचीअवस्था:"चारित्र्यहीनतेमुळेत्यालाखूपनिंदासहनकरावीलागली."
00000012 01 01 शिष्टपणा 0002 0400 00000062 0400 00000155 | शिष्टअसण्याचीस्थिती:"ह्यासूचनेमागेशिष्टपणानसूनवास्तवाचीजाणीवआहेहेरोजप्रवासकरणाऱ्यांनाकळूनयेईल."
00000013 01 03 अत्याचार:जुलूम:जुलूमजबरी 0001 0400 00000062 | एखाद्यालाअत्यंतत्रासदेण्याचीक्रिया:"इंग्रजांनीभारतीयक्रांतिकारकांवरअनेकअत्याचारकेले."
00000014 01 08 घोटाळा:अफरातफर:भानगड:घोळ:गोंधळ:गडबड:गोलमाल:हेराफेरी 0001 0400 00000040 | लबाडीनेकेलेलाअपहारकिंवाकळतनकळतझालेल्याचुकीमुळेनिर्माणझालेलीमोठीसमस्या:"बँकेच्याहिशेबातघोटाळाकरूनतोपळूनगेला."
00000015 01 02 जीवनसत्त्व:व्हिटॅमिन 0001 0400 00000026 | खाद्यपदार्थांतआढळणाराएकप्रकारचापोषकतत्त्व:"जीवनसत्त्वेसहाप्रकारचीआहेत."
00000016 01 02 चारित्र्यसंपन्नता:सच्छीलता 0001 0400 00000079 | चारित्र्यचांगलेअसण्याचीअवस्था:"चारित्र्यसंपन्नतामाणसालामहानबनवते."
00000017 01 01 मनमानी 0001 0400 00000052 | मनालावाटेलतेकरण्याचीकिंवाकरवूनघेण्याचीक्रिया:"उद्योगधंद्यातकेवळभांडवलदारांचीमनमानीनसूनश्रमिकांचेहीविचारलक्षातघेतलेपाहिजेत."
00000018 01 03 अनैतिक_कार्य:दुष्कर्म:कुकर्म 0001 0400 00000052 | नीतीविरुद्धकेलेलेकाम:"अनैतिककार्यालायशमिळतनाही."
00000019 01 03 धान्य:अन्नधान्य:दाणागोटा 0001 0400 00000029 | शेतातउत्पन्नहोणारेवजेवणबनवण्यासाठीउपयोगीपडणारेगहू, तांदुळ, तूरइत्यादीखाद्यरूपीदाणे:"श्यामधान्याचाव्यापारीआहे."
00000020 01 06 खाद्य:खाद्यपदार्थ:खाद्य_पदार्थ:खाद्यवस्तू:खाद्य_वस्तू:अन्नपदार्थ 0001 0400 00000029 | खाण्याजोगीवस्तू:"दूधआणिदुधाचेइतरपदार्थउदा. दही, लोणी, तूप, खवाखाद्यम्हणूनउपयोगातयेतात."
00000021 02 04 अननुभवी:नवखा:नवा:कच्चा 0001 0400 00000101 | अनुभवनसलेला:"तोह्याबाबतीतअजूनअननुभवीआहे"
00000022 02 01 अकुशल 0001 0400 00000101 | कुशलनसलेला:"अकुशलवअर्धकुशलकामगारकनिष्ठस्तरावरमानलेजाते."

Hindi WordNet:

It is a Lexical Database for Hindi gives semantic relations between Hindi words.

00000001 02 11 अजन्मा:अजात:अनुत्पन्न:अनुद्भूत:अप्रादुर्भूत:अज:अजन:अजन्म:अनन्यभव:अनागत:अयोनि 0005 2224 0000 00000008 2224 0000 00000008 2224 0000 00000008 2111 00000047 0400 00000101 | जिसनेजन्मनलियाहो:"देवकीकेअजन्मेबालकोंकेविषयमेंभविष्यवाणीहुईथी।"

00000002 02 09 अशुभ:अमांगलिक:अमाङ्गलिक:मनहूस:अमंगल:अमङ्गल:अक्षेम:अरिष्ट:दग्ध 0003 2224 0000 00011763 2111 00000183 0400 00000101 | जोशुभनहो:"बिल्लीकेद्वारारास्ताकाटाजानाअशुभमानाजाताहै।"

00000003 02 01 अप्रविष्ट 0004 2224 0000 00002475 2111 00000196 2111 00000923 0400 00000111 | जोप्रविष्टनहुआहो:"अप्रविष्टअतिथियोंकोशीघ्रहीभीतरप्रवेशकरनेदियाजाय।"

00000004 01 08 पवित्र_स्थान:चैत्य_स्थान:पुण्य_भूमि:पुण्य-स्थल:पुण्य_स्थल:चैत्य_स्थल:पवित्रभूमि:पवित्र_भूमि 0014 1223 0000 00000006 1102 00001973 1103 00003616 1103 00000452 1103 00006007 1103 00003619 1103 00006852 1103 00008218 1103 00008457 1103 00023798 1103 00030213 1103 00037934 1103 00039305 0400 00000037 | वहस्थानजोपवित्रमानाजाताहो:"हिंदुओंकेलिएकाशीएकपवित्रस्थानहै।"

00000005 01 07 शिवालय:शिव_मंदिर:शिव_मन्दिर:शिवाला:सिवाला:सौधाल:शिवायतन 0005 1102 00000451 1103 00039683 1141 00005665 1141 00000454 0400 00000037 | वहमंदिरजिसमेंभगवानशिवकीमूर्तिस्थापितकीगईहोऔरजहाँशिवकीआराधनाकीजातीहो:"वहप्रत्येकसोमवारकोशिवालयजाताहै।"

00000006 01 03 अपवित्र_स्थान:अपुण्यभूमि:अपवित्रस्थली 0004 1223 0000 00000004 1223 0000 00000004 1102 00001973 0400 00000037 | वहस्थानजोपवित्रनहो:"धार्मिकमान्यताहैकिभूत-प्रेतअपवित्रस्थानोंपरहीनिवासकरतेहैं।"

Methodology

The main focus of proposed study is to give a method should classify the multilingual text specifically which include regional languages like Hindi, Marathi, and English. The regional language like Marathi is having many variations of using a same word or phrase which makes them quite complex to classify as compared to the other universally used languages. To implement this complex idea the deep learning strategies can be more considered as it works on complex machine learning using algorithms like neural networks.

Neural networks were has been used recently to classify the multilingual Text classification for disaster related information. For proposed work we have referred the architecture given by the researchers. [13] The overview of the system that can be considered for analysing and implementing the work can be visualized as in fig 2.1

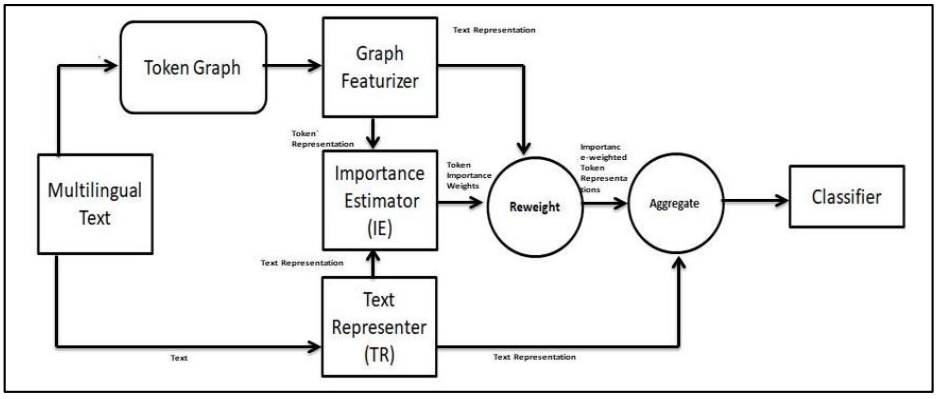


Fig. 2.1 Multiple Text Classification using neural Network

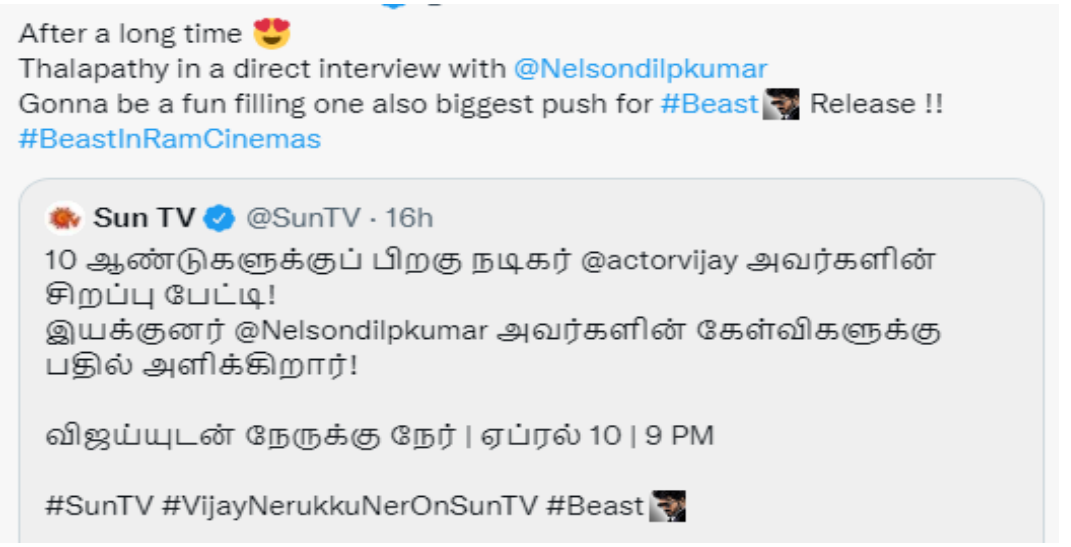
From the Overview of proposed GNoM framework we can analyse that multilingual text is given as an input which is given to text representor module and it generates different representations for it. Here word level and example level representations are used. Tokens are created to utilize it in graph of word by using a module graph featurizer. Other section which is used to find out the significance of the word with respect to the example level representation is importance estimator. The outcome of the all process these parts are combined and forwarded to the classifier.

Pseudocode

Step 1: Input the multilingual text to text representor.
Step 2: Analyse the text and convert it into the example level representation and word level representation
Step 3: Convert the text into tokens
Step 4: Input the Token to graph creator
Step 5: Create word graph
Step 6: Input graph vector and example level representation to importance estimator
Step 7: Output of importance estimator: weighted vector
Step 8: aggregate all weighted node vectors.
Step 9: input all aggregated node vectors to classifier
Step 10: Output: classified text

Application of proposed work

In technological world data analysis is one of the booming areas. The term like sentimental analysis is being used in a large way as the social media platforms are generating a vast data every minute that can be utilized in many ways in many field like media and so on. The input given to process like sentimental analysis is not always in a single language. Sometimes a single sentence may consist of multilingual text which can minimize the accuracy of result. The proposed work can be used to increase the result of accuracy to a better extend by introducing the methodologies like deep learning. The sample multilingual data can be considered in given example taken from twitter where more than one language has been used in single post.



Conclusions

Multilingual analysis can be utilized in many applications for classification of multiple languages used in same text. In this paper we have given overview of the work done in previous years on multilingual classification along which there methodologies. We have proposed a new way to look at this concept by using deep learning. The implementation of proposed work has not yet done but main focus is to implement the deep learning in order to provide multilingual classification for Indian regional languages which are rarely used as parameters for implementation of multilingual classification.

References

1. MichkleJardino , LIMSI-CNRS, ORSAY, France, “Multilingual Stochastic N-Gram Class Language Models”, IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings, 1996
2. Tony Mcenery And Andrew Wilson, Fernando Sanchez-Leon, Amalio Nieto-Serrano, “Multilingual Resources for European Languages:Contributions of the CRATER Project”, Literary and Linguistic Computing, Volume 12, Issue 4, November 1997
3. Michel Simard, “Multilingual text alignment”, Simard, M. (2000)., J. (eds) Parallel Text Processing. Text, Speech and Language Technology, vol 13. Springer, Dordrecht. https://doi.org/10.1007/978-94-017-2535-4_3
4. Roberto Basili and Alessandro Moschitti, “A robust model for intelligent text classification”, Proceedings 13th IEEE International Conference on Tools with Artificial Intelligence. ICTAI 2001.
5. Rowena Chau, Chung-HsingYeh, “Multilingual Text Categorisation for Global Knowledge Discovery Using Fuzzy Techniques”, Proceedings of the 2002 IEEE International Conference on Artificial Intelligence Systems (ICAIS’02)
6. Barbara Plank, “ALL-IN-1 at IJCNLP-2017 Task 4: Short Text Classification with One Model for All Languages”, Proceedings of the 8th International Joint Conference on Natural Language Processing, Shared Tasks, pages 143–148,

Taipei, Taiwan, November 27 – December 1, 2017. c 2017 AFNLP
7. NurendraChoudhary, “Learning Representations for Text Classification of Indian Languages”, DOI: 10.13140/RG.2.2.18818.94403
8. Sophie Groenwold, Samhita Honnavalli, Lily Ou, AeshaParekh,Sharon Levy, Diba Mirza, William YangWang , “Evaluating Transformer-BasedMultilingual Text Classification”, arXiv:2004 13939V2[cs.CL] 3 April,2020
9. Stephen Mutuvi, EmanuelaBoros, Antoine Doucet, Ga¨el Lejeune, Adam Jatowt, Moses Odeo, “Multilingual Epidemiological Text Classification: A Comparative Study”, Proceedings of the 28th International Conference on Computational Linguistics, pages 6172–6183 Barcelona, Spain (Online), December 8-13, 2020
10. SumanthDoddapaneni, Gowtham Ramesh, Mitesh M. Khapra1, AnoopKunchukuttan, Pratyush Kumar, “A Primer on Pretrained Multilingual Language Models”, arXiv:2107.00676v2 [cs.CL] 23 Dec 2021
11. Salil Aggarwal, Sourav Kumar, Radhika Mamidi, “Efficient Multilingual Text Classification for Indian Languages”, Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021) September 2021
12. Xiaotian Lin1, Nankai Lin, Kanoksak Wattanachote1, Shengyi Jiang , Lianxi Wang, “Multilingual Text Classification for Dravidian Languages”, arXiv:2112.01705,2021
13. Samujjwal Ghosh, SubhadeepMaji, MaunendraSankarDesarkar, “GNoM: Graph Neural Network Enhanced Language Models for Disaster Related Multilingual Text Classification”, In Proceedings of ACM Conference (Conference’17). ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn