International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 06 | June 2021 www.irjet.net p-ISSN: 2395-0072

Toxic Text Classification for Globalisation of Software Products

Jainil Viren Parikh[1], Prajwal YR[2], Rajashree Shettar[3]

[1,2,3]*Department of Computer Science and Engineering, RV College of Engineering, Bangalore, India*

Abstract - *An Globalization or G11N ensures that products have relevant features and functionalities that help them connect with geographically dispersed users. Localization, which is a part of globalization, deals with how a product can be modified to serve a regional market. During localization of a product one key step is translating the product document from one language to another. Translation can be erroneous and can lead to generation of toxic sentences in the translated language. So such sentences need to be classified and flagged. Performing this classification manually is time consuming and exhausting so we propose a NLP based approach to automate this multi label text classification task. It is observed that many multi label classification tasks have dependencies or correlations among labels. Existing methods generally ignore this relationship between the labels. Hence, we propose two methods one using Multi channel CNN network and the other using BERT to solve this classification problem.*

Key Words: **Multi Channel LSTM, BERT, multi label text classification, Natural language processing, Globalisation**

1.INTRODUCTION

With the increase in the non-EN speaking software markets it has become important for companies to adapt to globalization. Globalization planning and implementation affects every area of product and service commercialization. One important aspect of it is localization which ensures that a product or service aligns with a regional or language market, it considers factors such as local culture, social background and the accessibility of the target audience. One key goal of localization is to translate product documents from one language to another. During this translation there might be some errors that can lead to some toxic sentences. It is crucial to identify and flag these errors as it can have a direct impact on the customers experience. NLP techniques have evolved quickly in the last few years due to the increasing use of transfer learning in NLP tasks. Recent breakthroughs in transformer models have allowed us to generate language models that can be used for classification tasks. In this paper we aim to perform a multi label classification of the sentences from a product document into 6 categories i.e. severe toxic, insult, obscene, threat, hate, toxic. The classified sentences will be sent to a Linguist for a review and will be replaced by the correctly translated sentences.

This paper discusses two approaches to multi label text classification and offers a comparison between the two, carefully bringing out the use cases that each of these techniques excel in.

2. Literature Review

The idea of globalization and building an understanding of the different techniques to build products that can be scaled globally has been an area of interest since the beginning of 2010s. The sudden growth in the non-EN markets as discussed in [1] have also contributed to the increasing research interests in this field. As explained in [2] localization which is a part of G11N can be used to understand and modify the product to fit the regional culture and societal backgrounds. Localization uses many NLP techniques to reduce the manual effort in performing these activities. One such activity is translating document from one language to another.

The recent understanding of language models using attention models as described in [3] has paved the way for transformers. Transformer networks have opened the way to allow transfer learning [4] in NLP. Previously transfer learning was used in vision tasks and has achieved state of the art results. Transfer learning is a technique by which a network can be pretrained on a larger dataset from the same domain and then fine-tuned on the specific dataset for the task in hand. In NLP transfer learning started with the introduction of ULMFit which can be used for multiple NLP tasks. Transformers can also be transfer learnt by training them on huge datasets and later fine tuning then on the problem set in hand as discussed in [5]. BERT is the encoder section of a transformer that is used for language classification tasks. Roberta is a smaller version of the BERT base model presented in the original paper which can be transfer learnt for text classification tasks.

One of our proposed models uses Roberta(Robustly Optimized BERT) for multi label text classification of the sentences from a translated product document to detect toxic sentences. This classification task considerably reduces the manual effort involved in identifying these errors manually. We also train a multi-channel CNN model as described in [6] for the same task.

3. Methodology

The paper proposes to solve a multi label text classification problem of classifying document strings into these 6 categories: severe_toxic, insult, obscene, threat, hate, toxic. This paper proposes two models to solve this problem, namely, Multi-channel CNN and BERT.

3.1 Dataset

The dataset that we used was the google toxic comment classification dataset . This dataset contains 159,000 labeled samples which are comments from the wikipedia comments section and were labeled by humans for any toxic behaviour. This dataset is labeled on the same classes that we use to classify our document strings. The dataset will be split into train, test and validation sets in a 60-20-20 split.

3.2 Multi-channel CNN network

In this section we explain about multi-channel CNN. The data pre-processing and model architecture are explained in this section.

3.2.1 Data-Processing

Each sentence in the dataset is first cleaned by removing any punctuations, other symbols, and stop words that do not add meaningful context. We also remove product specific words that do not have real world meaning and so do not add value to the context. The words are then tokenized by splitting them at each space. The tokenized words are lemmatized to consider only the root word. Each sentence is formed into a vector which is padded to a fixed length which according to our the graph in Fig 1 is fixed at 200.

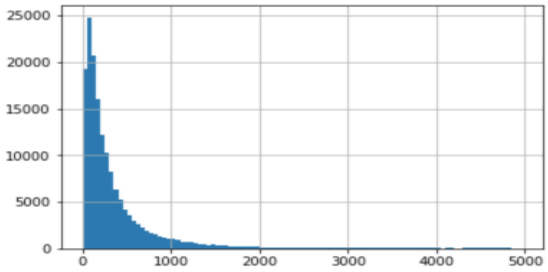

Fig-1: Length of sentences vs No on words

3.2.2 Model Architecture

LSTMs or Long short term memory are a type of Recurrent Neural Network which are capable of learning long-term dependencies. CNN are convolutional neural networks that can compress large amounts of information into smaller pools.We propose a multi channel CNN method that provides a different branch for each class as shown in the figure. Compared to conventional methods this method treats the problem as multiple binary classifiers by training a separate branch for each class.It is to be noted that each branch gets the same input vectors but they do not share any model parameters. One such branch of this model is made of a sequence of LSTM and CNN layers that are followed with a max pool layer and a linear layer. The final layer has just one node and has a sigmoid activation function. The loss function that is used for each branch is binary cross entropy as each branch solves a small binary classification problem.

This model is trained from scratch on the pre-processed dataset and result metrics are calculated on the test set.

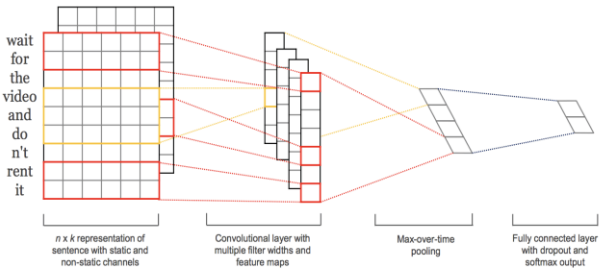

Fig 2: Multi Channel CNN Architecture

3.3 BERT

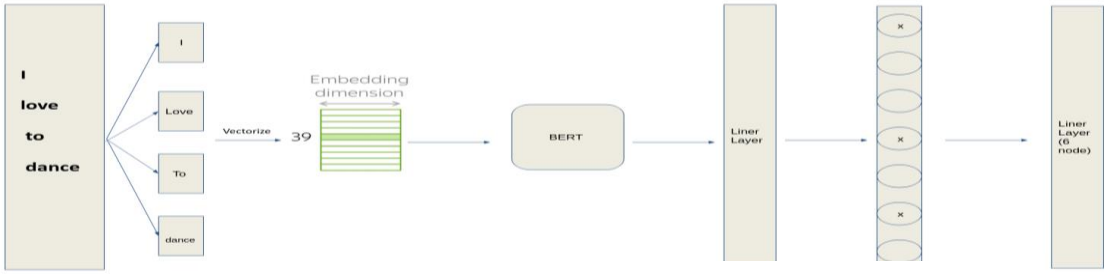

Fig 3: BERT Model Architecture

We propose a transfer learning approach which uses a pre-trained BERT base model which is later fine-tuned with an in-house dataset. A unique feature about BERT is that it has a unified architecture. BERT is a state of the art language model that can be used for multiple NLP tasks. We use a pre-trained BERT model for our implementation which is taken from hugging face[7]. The architecture followed to build the model is as shown in fig 2. Each sentence in the dataset is first cleaned by removing any punctuations, other symbols and stop words that do not add meaningful context. The sentences are sent to the specialized tokenizer that is provided by [8]. The output from the tokenizer is passed to BERT that passes it through the encoder and attention layers to obtain a single vector which is passed through a linear layer followed by a dropout as shown in the figure. The resultant is classified into six categories by passing it through a dense layer with six nodes and a sigmoid activation on each node.

3.1 Model Architecture

The BERT is a deep bidirectional transformer that is used for language understanding tasks. This network follows two steps: first is pre-training followed by fine-tuning. During the pre-training phase the model is trained on an unlabelled dataset over different pre-training tasks. Later for enhancing the network it is fine-tuned and the model is assigned the pre trained parameters, and all the parameters are fine-tuned using the labelled data from the task in hand. BERT base has the same model size as the OPENAI GPT [9] model. The BERT model[10] uses bidirectional self-attention whereas the latter uses constrained self-attention. As BERT handles multiple down-stream tasks, the input representation is able to unambiguously represent both single sentences and a pair of sentences in a single token. In this paper our task of classification deals with only one sentence input.

The BERT model from hugging face which uses the Word Piece embeddings which has a token vocabulary of nearly 30,000 tokens. The first token on each sentence is given a special token which is used for classification called CLS. The hidden state acquired from these tokens are used to collect the sequence representation for classification tasks.

In this paper we add a Dense network to the end of the BERT network to perform the classification task. The output from the BERT model is passed to the first dense layer which has number of nodes equal to the number of values in the output vector. Then the result is passed through a Dropout layer to reduce model complexity and hence overfitting. Then the results are finally passed to a dense layer that has six nodes and a sigmoid is used at each node.

4. Results and Observations

The models are tested on the test set and multiple metrics are calculated. For this multi label text classification task we achieved a best performance score of 0.875 label accuracy for our Multi channel CNN model. The model also performed well with regards to sentence accuracy of 0.72.The BERT model achieved an F1 score of 0.842 and a sentence accuracy of 0.92

For both models in our multi label text classification task we noted that label accuracy was an inaccurate indicator performance as the dataset had many zero labels compared to one's. We tried to resolve this problem by adding class weights by giving a higher weight to the least represented class. We also tried different proportions of our dataset, but these changes provided little relief. Hence we included the F1 score metric to better understand the results.

Table-1 : F1 score and accuracy metric for both models

Model	F1 Score	Accuracy
Multichannel LSTM	0.702	87.5
Pre-trained BERT	0.842	92.4

For training both networks we used 100k examples from the training dataset and 50k examples for testing the results. The pre-trained BERT model has also been trained on a larger corpus before training on these 100k samples. The results have been compared in the following table.

As shown in the table Table-1 the pre-trained BERT model outperforms the multi-channel CNN model in terms of both accuracy and F1 score. This is mainly due to the fact that the model has been trained on a much larger dataset and later fine- tuned on our dataset. This allows the model to learn details of sentence structure that is generally not feasible from a limited dataset like the one on which the multichannel CNN model was trained on.

CONCLUSION

In this work we proposed two methods for classification of text into 7 different categories to fulfil the localization objective. One that used the multichannel CNN approach to have a separate branch for each class to understand the intricacies of those classes. The other method used the concept of a pre-trained model to transfer and learn the parameters from a larger dataset and use it for our use case. By comparing the results of both the models with a F1 score and accuracy metric we come to a conclusion that the BERT model outperforms the multichannel CNN model.. However, the multi-channel CNN approach provides us with a unique way of modifying the complexity of the network for each class. Our research shows that localization can use the pre trained classification models to considerably reduce

human effort.. This research topic is very unique and at this stage numerous other relevant models should be further explored in future work.

REFERENCES

[1] William Aspray, Frank Mayadas, Moshe Y. Vardi "Globalization and Offshoring of Software" ACM 0001-0782/06/0200 2006.

[2] Peter Sandrini "Localization and Translation" LSP Translation Scenarios. Selected Contributions of the EU Marie Curie Conference Vienna 2007

[3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones "Attention Is All You Need" arXiv:1706.03762, https://arxiv.org/abs/1706.03762

[4] [4] Valeriya Slovikovskaya "Transfer Learning from Transformers"Proceedings of the 12th Language Resources and Evaluation Conference May 2020

[5] Kaichao You, Zhi Kou, Mingsheng Long, " Co-Tuning for transfer learning" NeurIPS 2020

[6] Ashok Kumar, Abiram S, Tina Esther, Erik Cambria "Comment toxicity detection via a multichannel convolutional bidirectional gated recurrent unit "Neurocomputing Volume 441, 21 June 2021

[7] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac "Hugging Face's Transformers: State-of-the-art Natural Language Processing" Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations October 2020

[8] Albert Au Yeung "BERT - Tokenization and Encoding arXiv:1706.03712, https://arxiv.org/abs/1706.03712

[9] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhaliwal, Arvind Neelakandan, Pranav Shyam "Language Models are Few-Shot Learners"34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding "Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 June 2019.