BERT of all trades, master of some

Author1, Author2, Author3
Affiliation1, Affiliation2, Affiliation3
Address1, Address2, Address3
author1@xxx.yy, author2@zzz.edu, author3@hhh.com
{author1, author5, author9}@abc.org

Abstract

This paper describes our results for TRAC 2020 competition held together with the conference LREC 2020. Competition consisted of 2 subtasks in 3 languages (Benghali, English and Hindi) where the participants' task was to classify aggression in short texts from social media and decide if it is gendered or not. We used a single BERT-based system with two outputs for all tasks simultaneously. Our model took the first place in English gendered text classification with 0.87 in F1 score and the second place in Bengali gendered text classification with the F1-score equal to 0.93 .

Keywords: aggression, classification, BERT, neural network, Transformer, NLP

1. Introduction

This paper is devoted to our system's solution for TRAC 2020 competition held together with LREC 2020 conference. TRAC 2020 competition consisted of 2 sub-tasks in 3 languages: Bengali, English and Hindi. In the first sub-task participants needed to make a system that labeled texts into three classes: 'Overtly Aggressive', 'Covertly Aggressive' and 'Non-aggressive'. In the second task the contestants' task was to label the same texts as gendered or not. The dataset contained 18681 texts in total, approximately 6000 texts for each language.

We used a single BERT-based system with two Linear layer outputs for all subtasks and languages simultaneously. Our model took the first place in English gendered text classification and the second place in Bengali gendered text classification.

2. Related Work

Aggression and misogyny detection is a rampant problem nowadays on the Internet. Many research initiatives have been devoted to its investigation. Given the overwhelming amount of information that social media users output every second, it is incomprehensible to monitor and moderate all of it manually. So it becomes useful to make at least semi-automatic predictions about whether a message contains aggression. Shared tasks and competitions are of great utility in this task because they provide data that can be used to research into new ways of aggression expression and allow different methods to be compared in a uniform and impartial way. Among such competitions we can name the previous TRAC competition (Kumar et al., 2018) and Offenseval (Zampieri et al., 2019). The first TRAC shared task on aggression identification was devoted to a 3-way classification in between 'Overtly Aggressive', 'Covertly Aggressive' and 'Non-aggressive' Facebook text data in Hindi and English. Offenseval was very similar in nature but it contained texts only in English. It consisted of 3 subtasks: binary offense identification, binary categorization of offense types and offense target classification.

Private initiatives also do not keep out of this problem. For example, there were held several challenges on machine learning competition platform Kaggle devoted to aggression investigation in social media, among them: Jigsaw Toxic Comment Classification Challenge and Jigsaw Unintended Bias in Toxicity Classification. The best solutions at Kaggle used a bunch of various techniques to improve their score.

There are few competitions that have the data labelled in several languages at the same time. However, we can see the progress in machine translation for inspiration. Single model in machine translation by Google.

3. TRAC-2 dataset

TRAC 2020 competition contained around 18000 texts in 3 languages (see Table 1): Bengali, English and Hindi. Hindi and Bengali texts could be written both in Roman and Bangla or Devanagari script within a single text (see Table 2). Moreover, many texts were written in two languages at the same time.

The authors of the competition split texts in all languages into training, validation and test datasets.

Dataset	English	Hindi	Bengali
Train	4263	3984	3826
Development	1066	997	957
Test	1200	1200	1188
Total	6529	6181	5971

Table 1: Number of texts for each language and dataset

Language	Examples
Bengali	best giris jain a katha
English	no gay gene discovered recently
Hindi	Negative positive दोनो म h sir
Hindi	Please logic mat ghusao

Table 2: Examples for different languages.

Each text had one label for each of the subtasks. The first subtask was a 3-way classification of aggression in social media texts. The classes were 'Overtly Aggressive',

'Covertly Aggressive' and 'Non-aggressive'. The second task was a binary classification between "gendered" and "not gendered" texts.
[Train/dev class distribution]

4. BERT model with multiple outputs

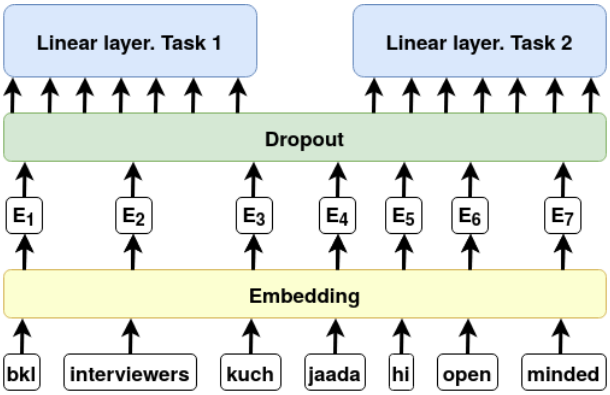

Figure 1: Our multitask model depiction)

In this task we wanted to experiment with a single model that works with multiple languages at once. We could have used an embedding-based approach with Word2Vec (Mikolov et al., 2013) or FastText (Joulin et al., 2016) input and a neural network classifier to classify aggression in texts (Gordeev, 2016). However, pre-trained language models are usually trained for one language at a time and either require augmentation via back-translation (Aroyehun and Gelbukh, 2018) or training a new word embedding model for several languages at once. Fortunately, it is possible to overcome this using multilingual language models such as BERT (Devlin et al., 2018).
BERT is a Transformer-based model (Vaswani et al., 2017). We used a multilingual uncased BERT model provided by Hugging Face (Wolf et al., 2019). We used PyTorch framework to create our model. BERT was trained using Wikipedia texts in more than 100 languages. All texts were tokenized using byte-pair encoding (BPE) which allows to limit the vocabulary size compared to Word2vec and other word vector models. The training consisted in predicting a random masked token in the sentence and a binary next sentence prediction. We did not fine tune the language model using the text data provided by the organizers. Information about the text language was not included into the model. We also did not perform any text augmentation or preprocessing besides standard byte-pair encoding. All texts longer than 510 tokens were truncated. Two tokens marking the beginning and the end of the sequence were added to each input text ("[CLS]" and "[SEP]"). Texts shorter than 510 tokens were padded with zeroes. All tokens excluding special ones were masked with ones, while all other tokens were masked with zeroes.
On top of BERT we added a Dropout layer to fight overfitting. The Dropout probability was equal to 0.1. On top of the Dropout Layer two softmax layers were added for each of the subtasks. Their dimensions were 3 and 2 respectively, equal to the number of classes. Target values were one-hot encoded. All texts were selected randomly out of the training and validation datasets. Cross entropy loss function was used for each of the outputs. The final loss function was calculated just as the sum of these two output losses. Half precision training was used vi Apex library [1]. We used a single Nvidia V100 GPU to train our model. The training batch size was made equal to 16. The model was trained for 10 epochs.
We used the same training, validation and test datasets as they were provided by the organizers. The validation data was applied only to hyperparameter tuning and was not included into the training dataset.

5. Results

Task	F1 (weighted)	Accuracy	Rank
Bengali-A	0.7716	0.7811	4
Bengali-B	0.9297	0.9293	2
English-A	0.7568	0.7683	3
English-B	0.8716	0.8708	1
Hindi-A	0.7761	0.7683	4
Hindi-B	0.8381	0.8392	3

Table 3: Results for all tasks

The results of our system are provided in Table 3. All in all we took the first place in gendered classification for English and the second place for the same task in Bengali. The results of our model were better for binary gendered classification than for 3-way aggression labelling. It might be due to the fact that we did not weight our loss function and both tasks contributed equally to the result. While it might be a better idea to give more emphasis to the target that has more potential values. We also did not use any early stopping or other similar techniques. The model was trained only for 3 epochs. A more challenging task might require more epochs to converge, thus, in future research we will also check the balance for early stopping between two tasks. Moreover, we could have enhanced subtask predictions by using values inferred by our model. We hope to also try it in future research.

6. Conclusion

This paper describes our results for TRAC 2020 competition held together with the conference LREC 2020. Competition consisted of 2 subtasks where participants had to classify aggression in texts and decide if it is gendered or not for 3 languages: Benghali, English and Hindi. We used a single BERT-based system with two outputs for all tasks simultaneously. Our model took the first place in English gendered text classification and the second place in Bengali gendered text classification.

7. References

Aroyehun, S. T. and Gelbukh, A. (2018). Aggression Detection in Social Media: Using Deep Neural Networks, Data Augmentation, and Pseudo Labeling. *Proc. First Work. Trolling, Aggress. Cyberbullying*, pages 90–97.

[1] https://github.com/NVIDIA/apex

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. oct.

Gordeev, D. (2016). Detecting state of aggression in sentences using cnn. In *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, volume 9811 LNCS, pages 240–245.

Joulin, A., Grave, E., Bojanowski, P., and Mikolov, T. (2016). Bag of Tricks for Efficient Text Classification. jul.

Kumar, R., Ojha, A. K., Malmasi, S., and Zampieri, M. (2018). Benchmarking Aggression Identification in Social Media. In *Proc. First Work. Trolling, Aggress. Cyberbulling*, Santa Fe, USA.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. In *Proc. Int. Conf. Learn. Represent. (ICLR 2013)*, pages 1–12, jan.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Adv. Neural Inf. Process. Syst.*, volume 2017-Decem, pages 5999–6009.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., and Brew, J. (2019). HuggingFace's Transformers: State-of-the-art Natural Language Processing. *ArXiv*, abs/1910.0.

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., and Kumar, R. (2019). SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval). In *Proc. 13th Int. Work. Semant. Eval.*