Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, pages 93–98
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

BERT of all trades, master of some

Denis Gordeev [1], Olga Lykova[2]
[1]Russian Presidential Academy of National Economy and Public Administration,
[2]National Research Nuclear University MEPhI
Moscow, Russia
gordeev-di@ranepa.ru, OVLykova@mephi.ru

Abstract

This paper describes our results for TRAC 2020 competition held together with the conference LREC 2020. Our team name was **Ms8qQxMbnjJMgYcw**. The competition consisted of 2 subtasks in 3 languages (Bengali, English and Hindi) where the participants' task was to classify aggression in short texts from social media and decide whether it is gendered or not. We used a single BERT-based system with two outputs for all tasks simultaneously. Our model placed first in English and second in Bengali gendered text classification competition tasks with 0.87 and 0.93 in F1-score respectively.

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

1. Introduction

Aggression, hate speech and misogyny detection is a rampant problem nowadays on the Internet. Thousands of people of all ages and nations face it every day. However, the problem is far from being solved. 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 problem because they provide data that can be used to research new ways of aggression expression and allow different methods to be compared in a uniform and impartial way. TRAC 2020 is one of such initiatives (Ritesh Kumar and Zampieri, 2020).

This paper is devoted to our system's solution for TRAC 2020 competition held together with LREC 2020 conference [1]. 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 would label texts into three classes: 'Overtly Aggressive', 'Covertly Aggressive' and 'Non-aggressive'. In the second task the contestants' aim 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 first place in English gendered text classification and second place in Bengali gendered text classification.

2. Related Work

Many researchers have paid attention to the problem of aggression detection on the Internet. However, hate and offensive speech are not homogeneous. There are various types of it that are aimed at different social groups and that use distinct vocabulary. Davidson et al. collected a hate speech dataset exploring this problem (Davidson et al., 2017). The authors relied on heavy use of crowd-sourcing. First, they used a crowd-sourced hate speech lexicon to collect tweets with hate speech keywords. Then they resorted again to crowd-sourcing to label a sample of these tweets into three categories: those containing hate speech, containing only offensive language, and those with neither. Later analysis showed that hate speech can be reliably separated from other types of offensive language. They find that racist and homophobic tweets are more likely to be classified as hate speech but sexist tweets are generally classified as offensive. Malmasi together with Zampiere explored this dataset even further (Malmasi and Zampieri, 2017). They have found that the main challenge for successful hate speech detection lies in indiscriminating profanity and hate speech from each other.

Many works have been devoted to hate speech detection. Thus, it seems that there should be a lot of available data exploring this problem for various languages. However, as the survey by Fortuna and Nunes (Fortuna and Nunes, 2018) showed most authors do not publish the data they collected and used. Therefore, competitions and shared tasks releasing annotated datasets that let explorers study the problem of hate speech detection carry even greater importance. 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 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 offence identification, binary categorization of offence types and offence target classification.

The best model at the previous TRAC competition used an LSTM-model (Aroyehun and Gelbukh, 2018). They

[1]available at github.com/InstituteForIndustrialEconomics/trac2

used preprocessing techniques to remove non-English characters and various special symbols. They also resorted to back-translation into 4 intermediate languages: French, Spanish, German, and Hindi.

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 [2] and Jigsaw Unintended Bias in Toxicity Classification [3]. The best solutions on Kaggle used a bunch of various techniques to improve the model score. Among such techniques were various types of pseudo-labelling such as back-translation and new language modelling subtasks.

There are few competitions that have the data labelled in more than two languages at the same time. However, the latest advances in machine translation show us that simultaneous multiple language learning may vastly improve the scores of the models (Arivazhagan et al., 2019). The researchers trained a single neural machine translation model on more than one billion sentence pairs, from more than 100 languages to and from English. The resulting massively multilingual, massive neural machine translation model demonstrated large quality improvements on both low- and high-resource languages and showed great efficacy on cross-lingual downstream transfer tasks.

Unsupervised cross-lingual language model learning also shows promising results. Some researchers have shown that pretraining of multilingual language models at scale leads to significant performance gains for a wide range of cross-lingual transfer tasks (Conneau et al.,). The authors trained a Transformer-based masked language model on one hundred languages, using more than two terabytes of filtered CommonCrawl data. Their model outperformed previous state-of-the-art solutions in a variety of cross-lingual tasks without hurting single-language performance.

However, even the most modern and sophisticated solutions are far from solving this problem. According to the survey by Fortuna and Nunes (Fortuna and Nunes, 2018) even human annotators have a tendency to disagree while labelling hate speech datasets. Detecting hate speech requires knowledge about social structure and culture. Even some websites may vary in what can be considered hate speech. Moreover, social phenomena and language are in constant evolution especially among young users which makes it challenging to track all racial and minority insults. Hate speech may also be very subtle and contain no offensive vocabulary or slurs.

3. TRAC-2 dataset

TRAC 2020 competition dataset contained around 18000 texts in 3 languages (see Table 2): 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 3). Moreover, many texts were written in two languages at the same time. It should also be noted that texts labelled as English contained a lot of names and words from non-English languages (most probably Hindi) and were hard to comprehend without knowledge of Hindi or Bengali (see Table 1).

Language	Class	Example
English	NAG	Best topic for Law Students!
English	CAG	Arundhati Roy has biggest bowls
English	OAG	One word for u bhaad me jaa chudail
English	NGEN	She is wrong.
English	GEN	I love u sakib but opu sotiya
Hindi	NAG	bro house of card ka review karona
Hindi	CAG	”Liberal bhi hai, Tolerant bhi hai!!!” LoL
Hindi	OAG	Feminism ki maa chod dee
Hindi	NGEN	Amrit Anand अबअ तो जउङअए् हइइ हअइ उनअको बोलो जउङअनए्
Hindi	GEN	@Nareshkumar Ravanaboina teri ma ka bhosda
Bengali	NAG	Dada taratari
Bengali	CAG	Basa niye bhore dite habe sali ke
Bengali	OAG	Ei mahila manasika rogi
Bengali	NGEN	Dada taratari
Bengali	GEN	Kena? Ranu mandala apanara bala chirache.

Table 1: Text Examples for all languages and classes.

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

Table 2: Number of texts for each language and dataset

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

Languages differed in their class distributions. In Subtask A Hindi and Bengali had a larger ratio of covertly aggressive texts than English both in the train and development datasets (see Fig. 1). The numbers for Subtask B are similar. English had a much lower ratio of gendered texts than Hindi or Bengali (see Fig. 2). Moreover, it should be noted

[2]https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge

[3]https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification

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 3: Examples of script usage for different languages.

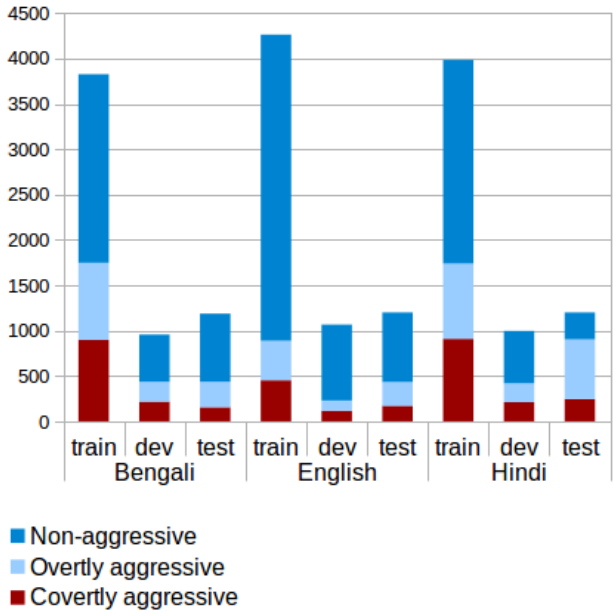

Figure 1: Class distribution (Subtask A)

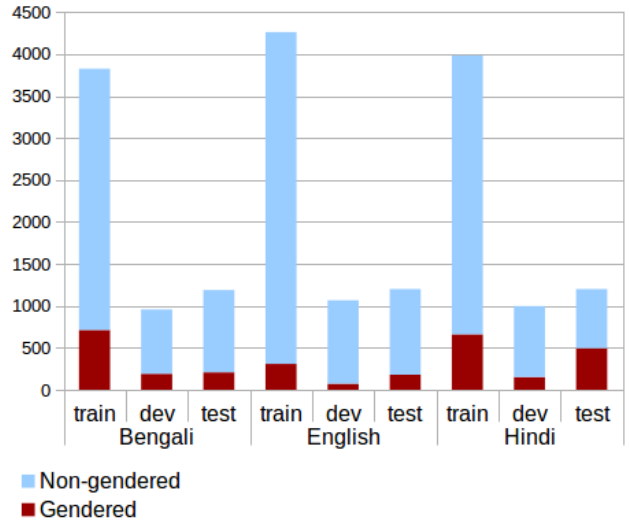

Figure 2: Class distribution (Subtask B)

4. BERT model with multiple outputs

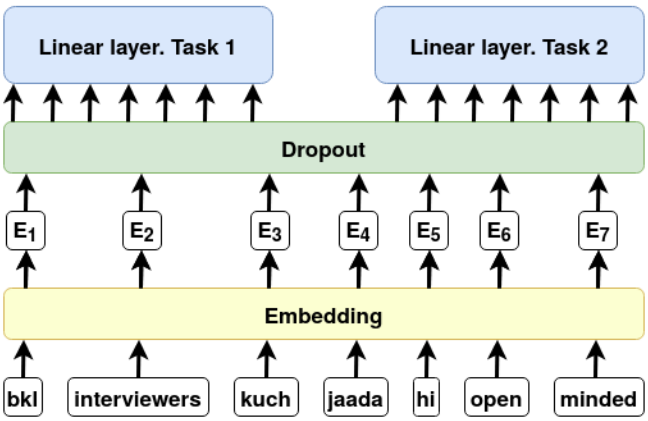

Figure 5: Our multitask model depiction

that the distribution for Subtask B was rather skewed. For example, the number of gendered texts for English in the training dataset was 13 times higher than that of the non-gendered ones (for Bengali and Hindi the numbers are 4.4 and 5 respectively). For all languages class distributions between train and development datasets did not differ much. However, test distributions (which were unknown during the competition) do not look the same as the train dataset. For example, Hindi as well as English had many more gendered texts in the test (0.17 vs 0.70 and 0.07 vs 0.17 ratios respectively). For subtask A, Hindi also had some peculiarities with overtly aggressive texts being the majority in the test dataset while neutral texts dominated the train and development datasets.

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

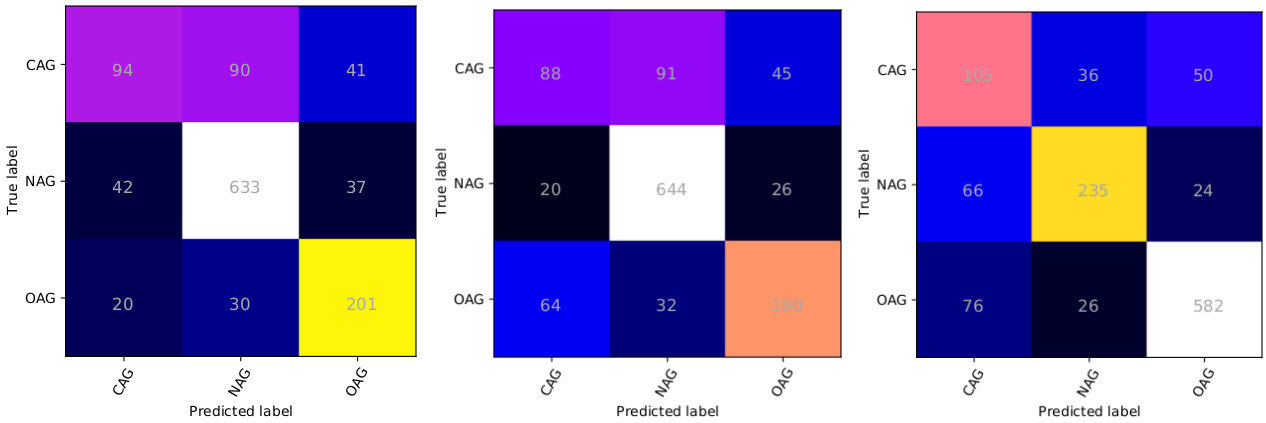

Figure 3: Subtask A. Confusion matrices for the final test dataset. Provided in the following order: Bengali, English, Hindi (the 4th, 3rd and 4th places in the leaderboard respectively)

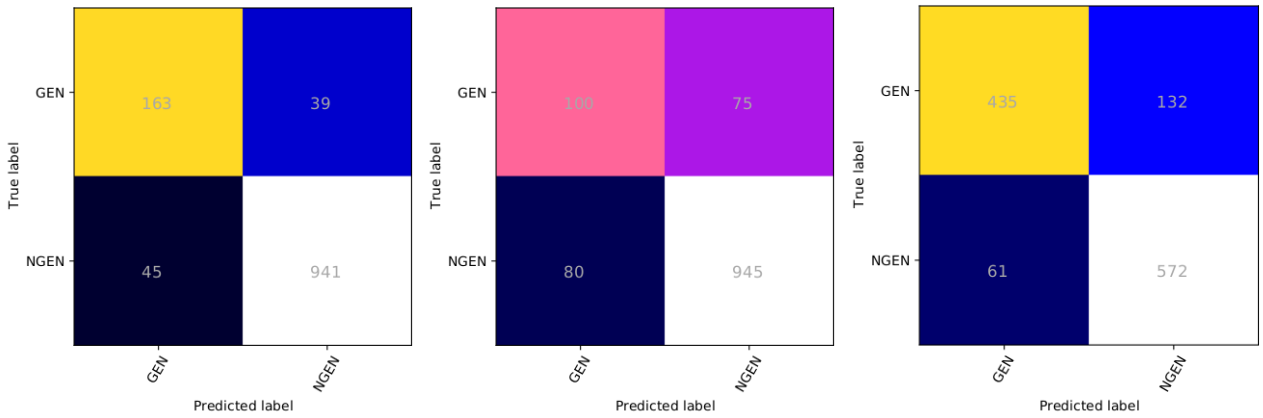

Figure 4: Subtask B. Confusion matrices for the final test dataset. Provided in the following order: Bengali, English, Hindi (the 2nd, 1st and 3rd places in the leaderboard respectively)

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 limiting 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 in the model. We also did not perform any text augmentation or pre-processing 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 via Apex library [4]. 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 in the training dataset.

Our team members have only knowledge of the English language and absolutely no familiarity with Hindi or Bengali.

5. Results

The results of our system are provided in Table 4. All in all we took first place in the 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

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

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 4: Results for all tasks

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. Given that the model was trained for 10 epochs, it might have been not enough for 3-way classification. 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 targets. Moreover, we could have enhanced individual subtask predictions by using values inferred by our model for another target. We hope to also try it in future research.

As can be seen from confusion matrices for subtask A (see Fig. 3) for all languages, our model had difficulties in distinguishing covertly expressed aggression and misclassified it in almost half of the cases. It seems only logical that it should be the most challenging class to predict because in many cases it may be difficult even for humans to correctly recognize subtle aggression, especially on the Internet where there are few non-verbal indicators.

Confusion matrices for the second subtask for all languages can be seen in Figure 4. Our results for the English dataset, where we had almost a half of gendered texts misclassified, were worse than for Bengali. However, given the skewed class distribution for English, this class turned out to be challenging for all of the 15 participants and our model outperformed other solutions. In Bengali all systems including ours had higher results than for all other languages. It may be attributed to the dataset peculiarities or for some features of the Bengali language which make it easy to recognize gendered texts (e.g. for English with its lack of genders and cases in nouns, it might be a more challenging problem given the results of the competition). The lower performance of our model for Hindi might show that our system might have overfitted to the class distributions from the train set.

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: Bengali, 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. Thus, cross-lingual multitask BERT finetuning can be considered a promising approach even for non-IndoEuropean languages. In future work we will check the balance for early stopping between two targets and weighting schemes for simultaneous subtask training which might improve the results of our model.

7. References

Arivazhagan, N., Bapna, A., Firat, O., Lepikhin, D., Johnson, M., Krikun, M., Chen, M. X., Cao, Y., Foster, G., Cherry, C., Macherey, W., Chen, Z., and Wu, Y. (2019). Massively Multilingual Neural Machine Translation in the Wild: Findings and Challenges. *arxiv.org*.

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.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V.). Unsupervised Cross-lingual Representation Learning at Scale. Technical report.

Davidson, T., Warmsley, D., Macy, M., and Weber, I. (2017). Automated Hate Speech Detection and the Problem of Offensive Language. In *Proc. ICWSM*.

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

Fortuna, P. and Nunes, S. (2018). A Survey on Automatic Detection of Hate Speech in Text. *ACM Comput. Surv.*, 51(4):85.

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.

Malmasi, S. and Zampieri, M. (2017). Detecting Hate Speech in Social Media. In *Proc. Int. Conf. Recent Adv. Nat. Lang. Process.*, pages 467–472.

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.

Ritesh Kumar, Atul Kr. Ojha, S. M. and Zampieri, M. (2020). Evaluating aggression identification in social media. In Ritesh Kumar, et al., editors, *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying (TRAC-2020)*, Paris, France, may. European Language Resources Association (ELRA).

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.*