BERT FOR RUSSIAN NEWS CLUSTERING

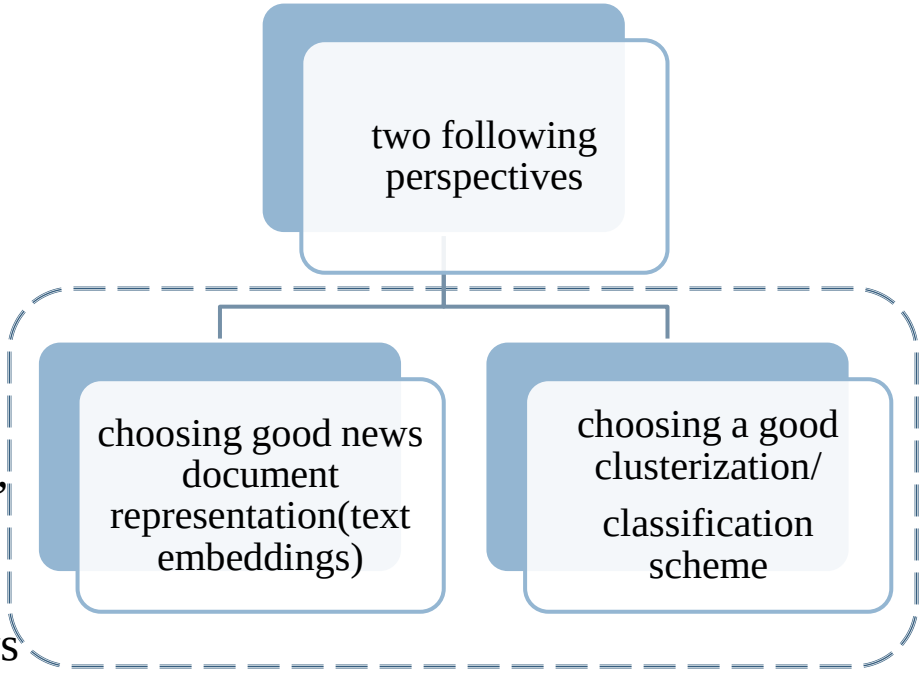
Khaustov Sergei, Gorlova Nadezhda, Kalmykov Andrey,
Kabaev Anton
MTS AI
Moscow, Russia
2021

Task

News clustering is a common task in the industry, and its purpose is to group news by events.

- News aggregators
- News skill for smart devices and assistants

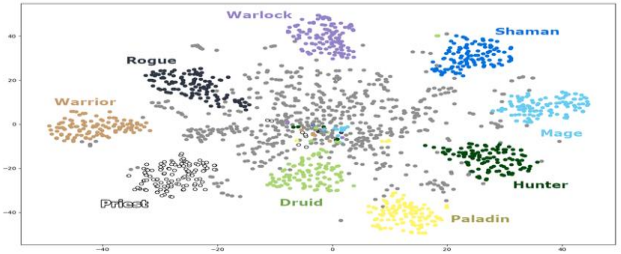
Related work

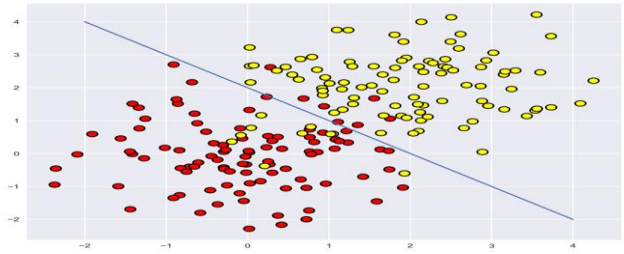


- BERT outperforms Word2vec, GloVe, fastText, ELMo
- RuBERT shows the best score
- Sentence-BERT

- K-Means Clustering algorithm.
- Metrcic algorithms
- Agglomerative clustering

Method Description





trying to learn good news text representation embeddings for

subsequent clustering.

use binary classification to classify if two news texts are from

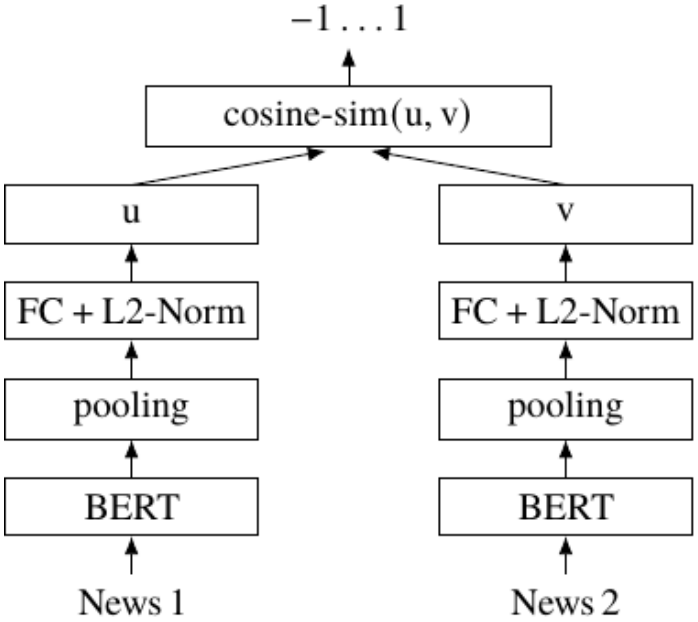
the same group or not.

Model : BERT embeddings

- First, tokenized news text goes to the pre-trained BERT layers.
- Then, BERT output embeddings for each token are averaged by the pooling layer.
- These averaged vectors then proceed to the fully connected layer size of 768, followed by L2 normalization layer. For model training, hard triplet loss is used.

Model 1 inference

used for agglomerative clustering with average linkage and cosine distance.



token embeddings are averaged by the pooling layer

Figure 1: Model inference

Results: model 1

For model training, hard triplet loss is used.

BERT weights do not freeze while training and initialized from pre-trained model.

We trained the model on GPU Tesla V100 for ten epochs with a learning rate of 1e-6 and a batch size of 16.

	Model	F1-score public LB
without finetune	USE	89.4%
without finetune	SBERT	88.1%
	OUR	91.7%

Table 1: Embeddings comparison on public leaderboard.

Model 2: BERT classifier

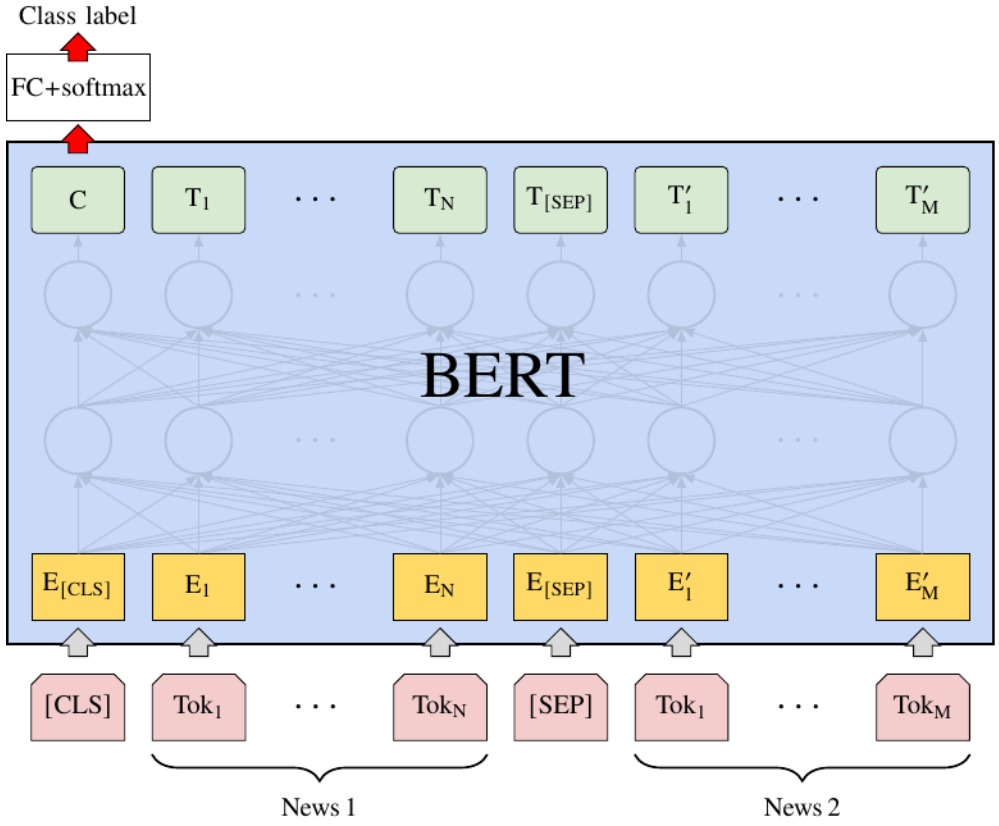


Figure 2: BERT NSP task architecture, depicted from the original paper.

Results: Model 2

The model is trained with cross-entropy loss. BERT weights do not freeze while training.

We trained a model on GPU Tesla V100 with the batch size of 8 for 6 epochs with a learning rate of 1e-5.

Model	F1-score public LB
RuBERT	96.7%
bert-base-multilingual	96.1%
sbert-large-nlu-ru	96.2%

Table 2: Initial weights comparison on the public leaderboard.

Results

Model	F1-score public LB	F1-score private LB
BERT Embeddings + clusterization	91.7%	91.27%
BERT Classifier	96.7%	95.98%
Competition winner	96.9%	96.04%

Table 3: F1 score on public/private leaderboards and comparison with the competition leader.

Our model gives best result without using ensembles.

BERT embeddings model requires O(n) inferences and BERT classifier – $O(n^2)$.

Confusion matrix

The confusion matrix shows there were 263 misclassifications: 148 false positive and 115 false negative.

tp	3803	148	fp
fn	115	4427	tn

Confusion matrix

Error analysis

false positive:

- 45% of false positive samples to have doubtful labels
- Another 15% of false positive also relates to the same news main event, but with the addition of continuation or person's comments
- The rest 40% of false positive samples definitely belong to different news stories but shares similar topics or context

false negative:

- 33% have doubtful labels
- 48% are supported with additional recap/continuation/person's comment
- 19% are surely the model's errors.

Error Samples

error type	first news fragment	second news fragment
false positive, addition of continuation	Житель вологды дмитрий губин подал в суд на кадырова из-за комендантского часа в чечне.	Верховный суд чечни не стал рассматривать иск вологжанина дмитрия губина, в котором он пытался оспорить спецмеры из-за пандемии коронавируса, введенные рамза-ном кадыровым.
false positive, same topic but different place/time	В туле покупатели устроили давку в очереди за дешевыми кастрюлями	В башкирии устроили давку из-за кастрюль за 99 рублей

Table 4: Error analysis.

Conclusion

- presented and compared two approaches for news clustering based on BERT
- The second method has shown promising results, but it can hardly be applied without modifications in real life due to performance.
- However, the advantage of our first approach is comparative computational efficiency.

Thank You for your attention!

Questions?