CURRENT LANDSCAPE OF THE RUSSIAN SENTIMENT CORPORA

Evgeny Kotelnikov

Vyatka State University, Kirov, Russia

ITMO University, Saint Petersburg, Russia

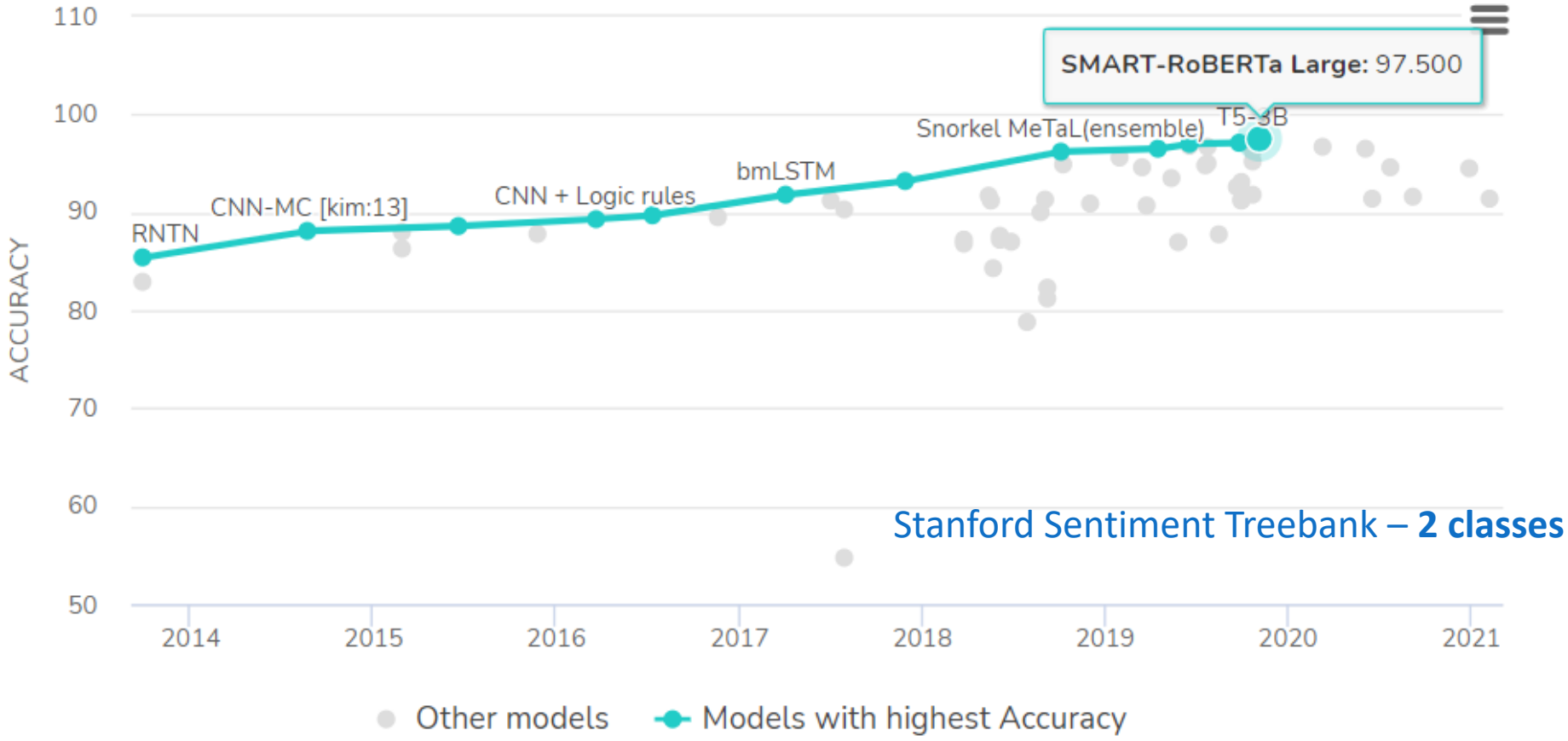
kotelnikov.ev@gmail.com

Sentiment analysis

- Sentiment analysis is still an urgent problem…

Sentiment analysis

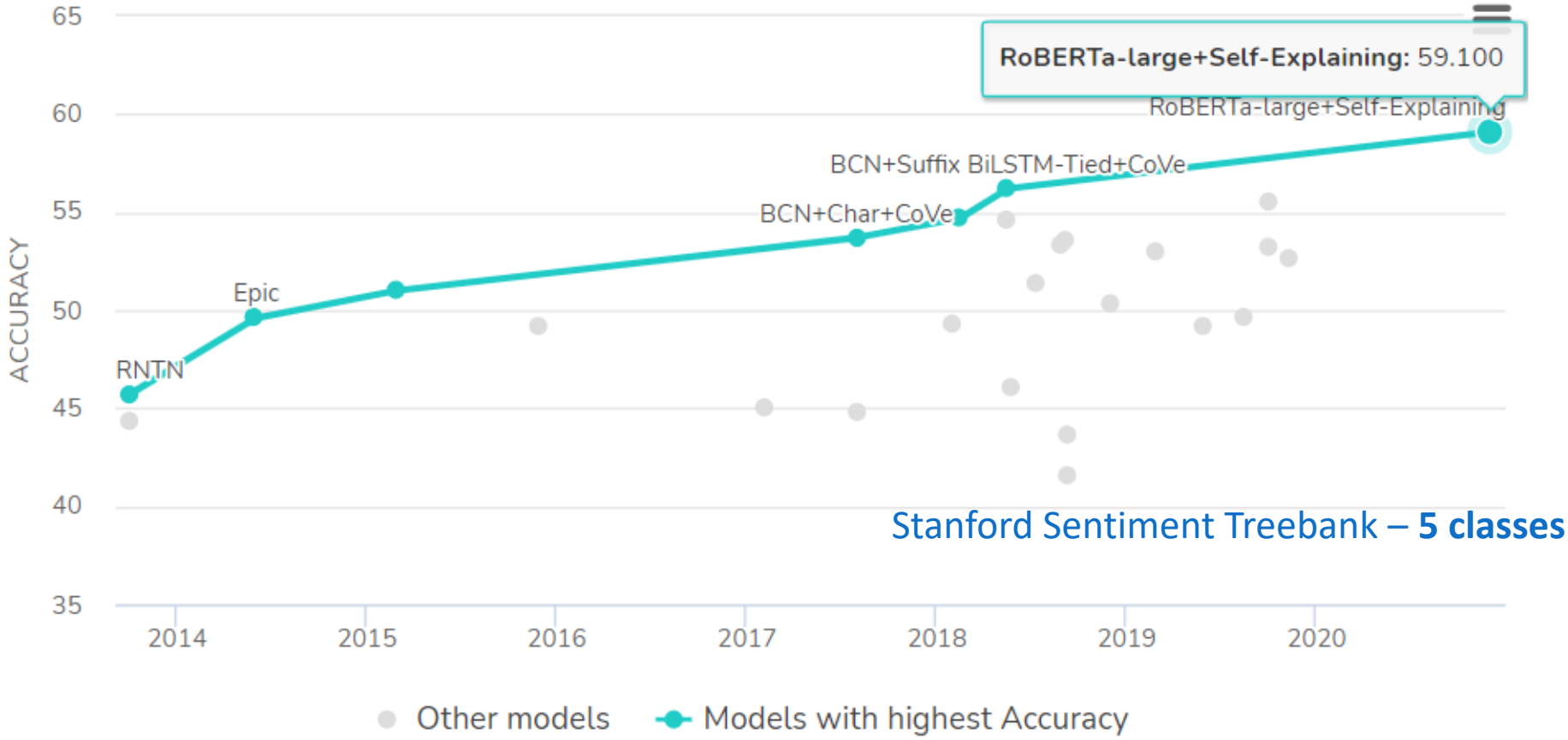
- Sentiment analysis is still an urgent problem…



https://paperswithcode.com/sota/sentiment-analysis-on-sst-2-binary

Sentiment analysis

- Sentiment analysis is still an urgent problem…



https://paperswithcode.com/sota/sentiment-analysis-on-sst-5-fine-grained

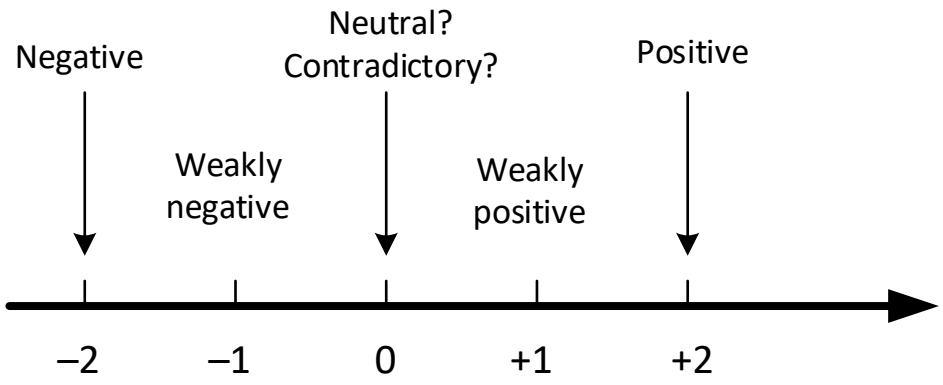
Goals

- Overview of all publicly available Russian corpora
- Ranking of corpora by annotation quality
- New performance scores for the existing Russian corpora of reviews
- Research of the influence of the training dataset expansion on the performance of the sentiment analysis of reviews

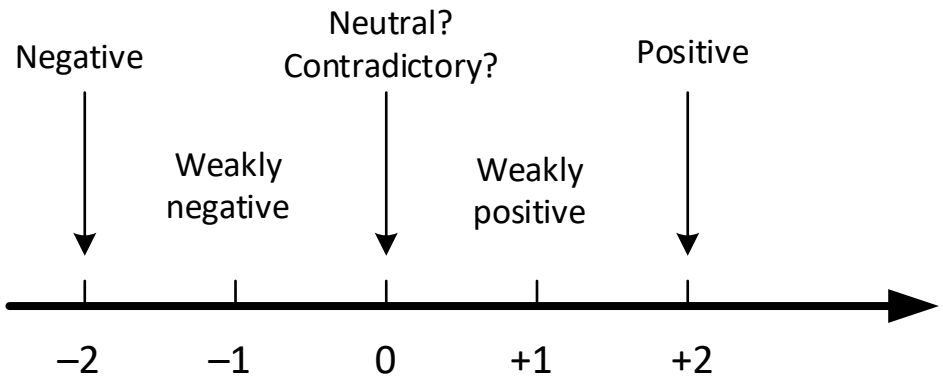
Characteristics of corpora

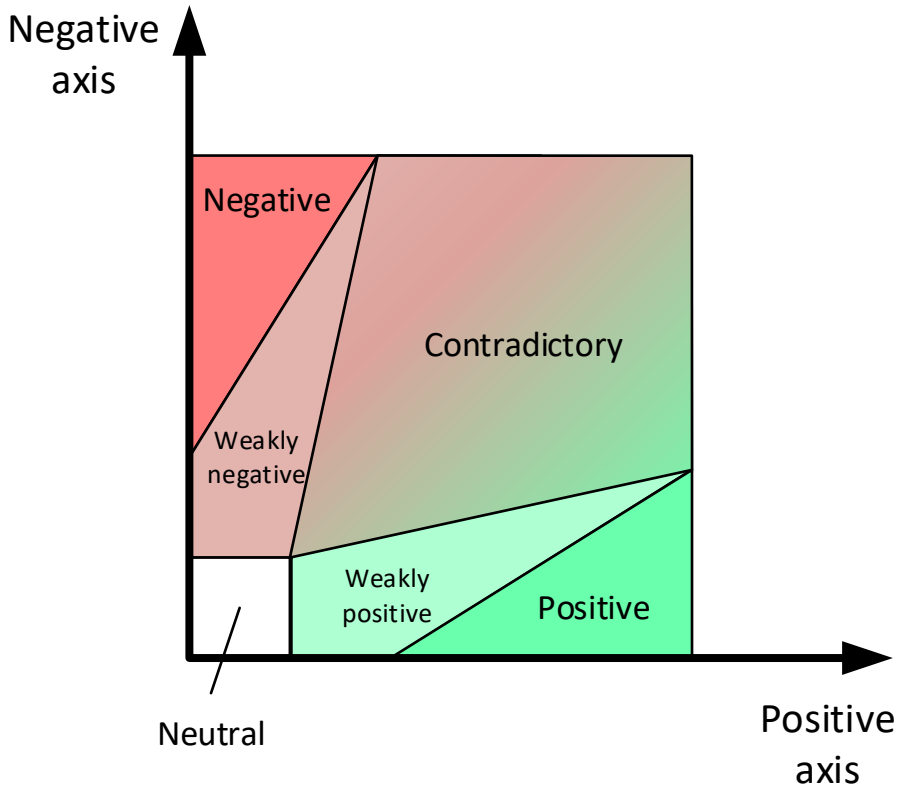
- Source of the texts
- Domain
- Corpus size
- Train/test split
- Annotation method
- Sentiment classes
- The ratio of the sentiment classes

Sentiment spaces



Sentiment spaces





Quality of the annotation

- L1 = High level – at least two annotators
  - including crowdsourcing
- L2 = Middle level – only one annotator
- L3 = Lower middle level – author's score
- L4 = Low level – automatic annotation

Existing corpora

- ROMIP-2011 – competition 2011 of sentiment analysis systems, reviews of books, movies and cameras [Chetviorkin et al., 2012]
- ROMIP-2012 – competition 2012 of sentiment analysis systems, new test data – reviews of books, movies and cameras + news [Chetviorkin & Loukachevitch, 2013]
- RuTweetCorp – large Russian-language tweet corpus [Rubtsova, 2014]
- SentiRuEval-2015 – competition 2015: aspect-based sentiment analysis of reviews and object-oriented sentiment analysis of tweets [Loukachevitch et al., 2015]
- SentiRuEval-2016 – competition 2015: new test data for tweets [Loukachevitch & Rubtsova, 2016]
- Twitter Sentiment for 15 European Languages – corpus with tweet ids and scores (without texts) [Mozetič et al., 2016]

Existing corpora

- SemEval-2016 – international competition 2016: new test data for restaurant reviews [Pontiki et al., 2016]
- LinisCrowd – posts and comments from LiveJournal [Koltsova et al., 2016]
- Russian Hotel Reviews – hotel reviews from tripadvisor.ru with aspect-based annotations [Rybakov & Malafeev, 2018]
- RuSentiment – posts from vk.com [Rogers et al., 2018]
- RuSentRel – news articles from the inosmi.ru: annotated in relation to the named entities [Loukachevitch & Rusnachenko, 2018]
- RuReviews – reviews about women's clothes and accessories [Smetanin & Komarov, 2019]
- Kaggle Russian News Dataset – Kazakhstan news in Russian [Kaggle]

See: [Smetanin, 2020] and this paper

Existing corpora

L1 – at least two annotators
L2 – only one annotator
L3 – author's score
L4 – automatic annotation

Corpus	Source	Annotation quality	Number of texts		Number of classes
			Train	Test	
ROMIP-2011	Reviews	Train: L3, test: L1	40,639	833	2, 3, 5
ROMIP-2012	Reviews	L2		948	2, 3, 5
	News	L2	4,260	4,573	Train: 4, test: 3
SentiRuEval-2015	Reviews	L2	403	403	4
	Tweets	L1	9,722	8,308	Train: 4, test: 3
SentiRuEval-2016	Tweets	L1		5,500	3
SemEval-2016	Reviews	L1	302	103	4
LinisCrowd	Posts	L1, L2	39,419		5
Russian Hotel Rev.	Reviews	L3	50,328	6,876	5
RuSentiment	Posts	L1	24,124	2,621	3
RuSentRel	News	L1	73		2
RuReviews	Reviews	L3	89,999		3
RuTweetCorp	Tweets	L4	226,834		2
Kaggle	News	?	8,263		3

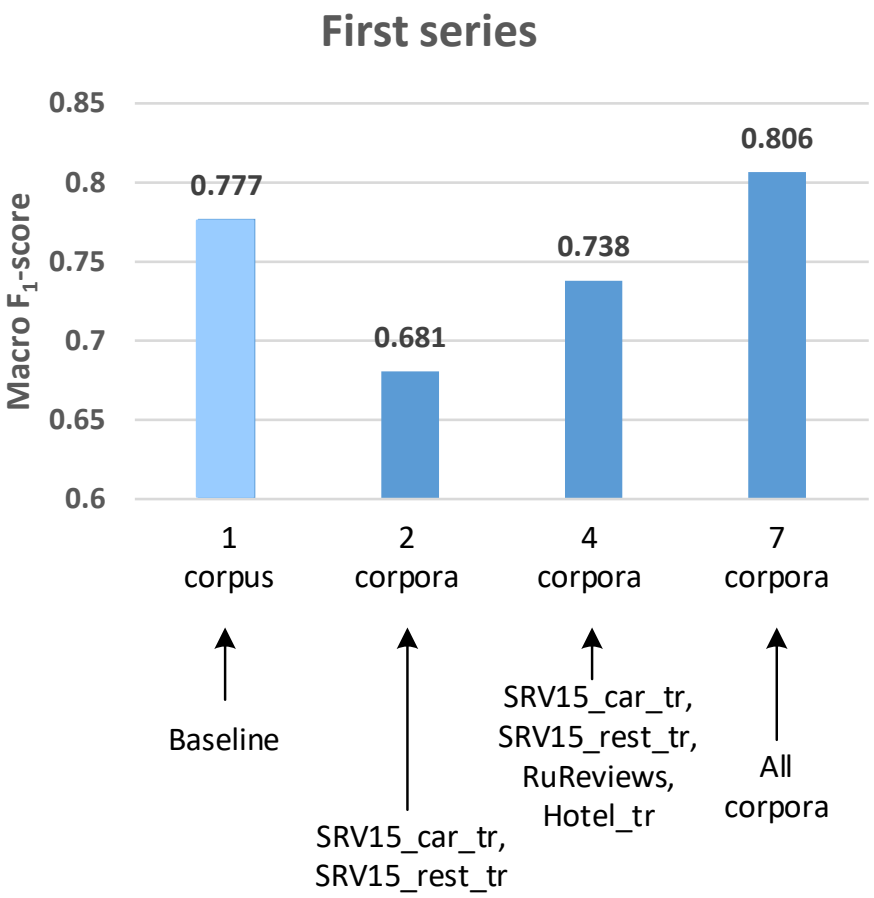
Experiments: corpora

- Train corpora (7):
  - ROMIP-2011: R11_book_tr, R11_mov_tr, R11_cam_tr
  - SentiRuEval-2015: SRV15_car_tr, SRV15_rest_tr
  - RuReviews: RuReviews
  - Russian Hotel Reviews: Hotel_tr
- Test corpora (9):
  - ROMIP-2011: R11_book_te, R11_mov_te, R11_cam_te
  - ROMIP-2012: R12_book_te, R12_mov_te, R12_cam_te
  - SentiRuEval-2015: SRV15_car_te, SRV15_rest_te
  - Russian Hotel Reviews: Hotel_te

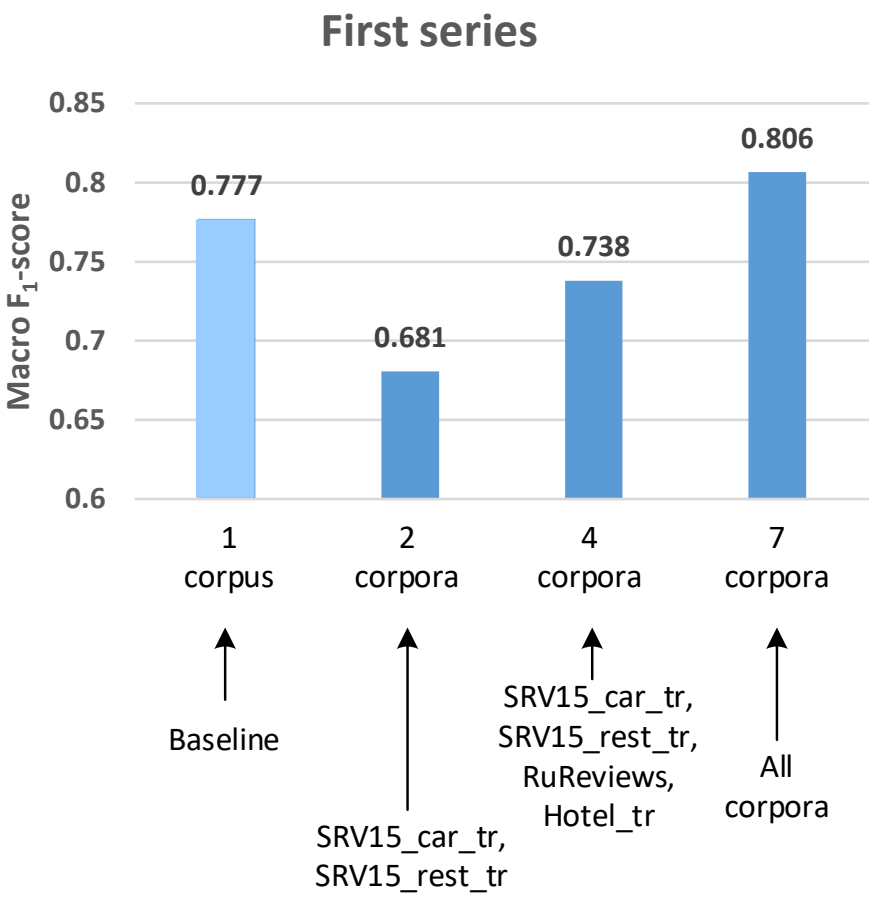
Experiments: model

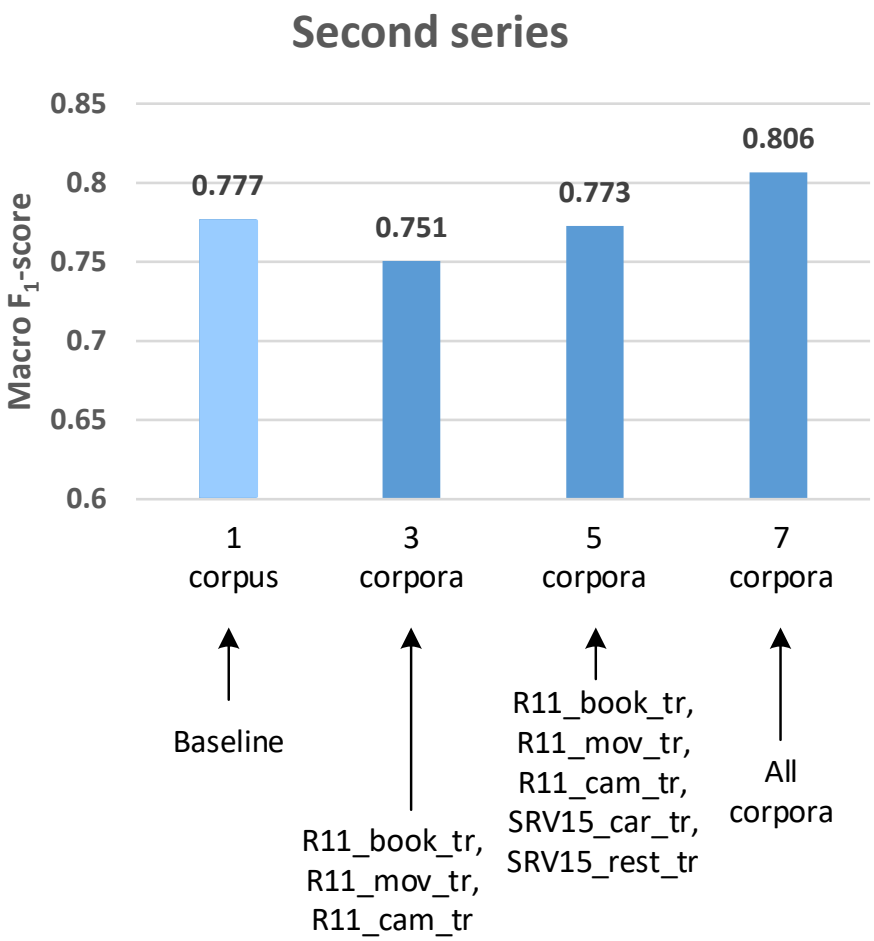
- Model: RuBERT [Kuratov & Arkhipov, 2019]
- Hyperparameters:
  - number of epochs: 5
  - batch size: 8
  - learning rate: $2 \cdot 10^{-5}$.
- Hardware: Google Colab Pro – Tesla V100 and Tesla P100 (16 GB)
- 3 training runs for each experiment due to random initialization of the weights
- Metric: macro-averaged F1-score

Experiments: 2nd series results



Experiments: 2nd series results





Comparison with previous results

Model	ROMIP-2011			ROMIP-2012		
	Book	Movies	Cameras	Book	Movies	Cameras
The best models from ROMIP-2011	0.723	**0.770**	**0.921**	–	–	–
The best models from ROMIP-2012	–	–	–	0.715	0.669	0.707
RuBERT trained on related training corpus	0.745	0.762	0.909	0.648	**0.698**	**0.723**
RuBERT trained on all the corpora	**0.841**	0.756	0.915	**0.724**	0.684	0.665

Conclusion

- There are more than a dozen Russian-language text corpora, annotated by sentiment
  - These corpora differ significantly in sources, domains, sizes, quality of annotation and sentiment scales
- A variety of corpora can be used to build better models, which is confirmed by our experiments
- The performance is (obviously) strongly influenced by the presence of a corpus in a given domain
- Less obvious was the fact that adding corpora in other domains, as a rule, either does not worsen the performance, or improves it

Thank you for your attention!