One Million Posts Corpus

Seminar Deep Learning for Language and Speech

Jens Becker, Julius Plehn, Oliver Pola

Language Technology Group
Fachbereich Informatik
Fakultät für Mathematik, Informatik und Naturwissenschaften
Universität Hamburg

25.02.2020

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Agenda

Corpus

- One Million Posts Corpus
- User posts from website of Austrian daily newspaper DER STANDARD
- Taken over 12 months 2015-2016
- 1,000,000 unlabeled posts
- 11,773 labeled posts
- Available at `https://ofai.github.io/million-post-corpus/`

[Schabus et al., 2017, Schabus and Skowron, 2018]

Categories

	Labeled	Does apply		We apply	
Sentiment Negative	3599	1691	47%		
Sentiment Neutral	3599	1865	52%		
Sentiment Positive	3599	43	1%		
Off Topic	3599	580	16%		
Inappropriate	3599	303	8%		
Discriminating	3599	282	8%		
Possibly Feedback	6038	1301	22%	72	2%
Personal Stories	9336	1625	17%	47	1%
Arguments Used	3599	1022	28%		

[Schabus et al., 2017]

- We use only posts, that are annotated as 0 or 1 for each category

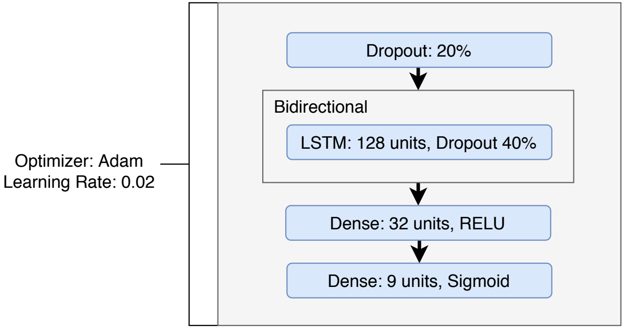
Word2Vec Embedding

- Using Word2Vec embedding [Mikolov et al., 2013]
- Applied by Gensim [Řehůřek and Sojka, 2010]
- Loading pretrained german model [depset.ai, www]
- Vocabulary size = 1,309,281
- Embedding dim = 300
- Padded sequence length = 80

Embedding Method

	Method 1	Method 2
Preprocessing	posts to lists of word indices	posts to vectors
Embedding matrix	feed matrix to training model	not needed, discard all unseen words
Training	repeat feeding lists of word indices	repeat feeding vectors
Memory usage (GPU)	high	lower
Delete embedding model after preprocessing	no	yes
Memory usage (CPU)	high	lower
Applicable to low-end systems	no	yes

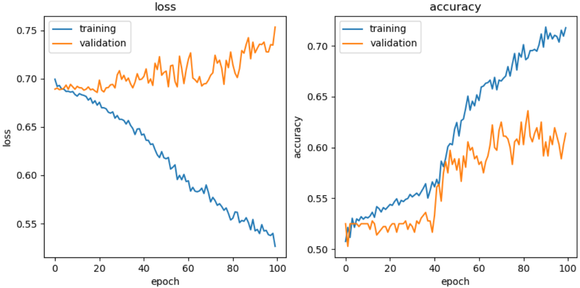
Model



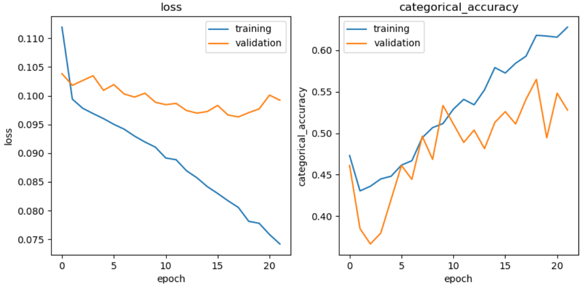
- Implemented using Tensorflow 2 and Keras
- Supervised using automatic learning rate adaption (*ReduceLROnPlateau*) and *EarlyStopping*

Training

SentimentNegative



All



- Left two: Single-Model for Sentiment Negative before implementing Early Stopping
  - Although validation loss increases early, accuracy (precision, recall) still improve
- Right two: Multi-Model with Early Stopping

Results Single-/Multi-Model

	True Pos	True Neg	False Pos	False Neg	Accuracy	Precision	Recall	F_1
Sentiment Negative	145	159	99	137	0.56	0.59	0.51	0.55
	112	236	59	133	0.64	0.65	0.46	0.54
Sentiment Neutral	190	133	149	68	0.60	0.56	0.74	0.64
	216	124	126	74	0.63	0.63	0.75	0.68
Sentiment Positive	0	533	0	7	0.99	0	0	0
	0	535	0	5	0.99	0	0	0
Off Topic	0	452	0	88	0.84	0	0	0
	11	423	14	92	0.80	0.44	0.11	0.17
Inappropriate	0	504	0	36	0.93	0	0	0
	1	483	0	56	0.90	1.00	0.02	0.03
Discriminating	0	497	0	43	0.92	0	0	0
	1	492	3	44	0.91	0.25	0.02	0.04
Possibly Feedback	0	531	0	9	0.98	0	0	0
	0	527	0	13	0.98	0	0	0
Personal Stories	0	532	0	8	0.99	0	0	0
	0	534	0	6	0.99	0	0	0
Arguments Used	78	350	51	61	0.79	0.60	0.56	0.58
	99	339	41	61	0.81	0.71	0.62	0.66

Comparison: Sentiment Negative

	Accuracy	Precision	Recall	F_1
[Schabus et al., 2017] (best)		0.5842	0.7197	0.6137
[Schabus et al., 2017] (LSTM)		0.5349	0.7197	0.6137
Our Single-Model	0.5630	0.5943	0.5142	0.5513
Our Multi-Model	0.6444	0.6550	0.4571	0.5384

Comparison: Sentiment Positive

	Accuracy	Precision	Recall	F_1
[Schabus et al., 2017] (best)		0.2353	0.4651	0.1333
[Schabus et al., 2017] (LSTM)		0	0	0
Our Single-Model	0.9870	0	0	0
Our Multi-Model	0.9907	0	0	0

- Model learns to predict always 0 (true pos = 0, false pos = 0)

Comparison: Arguments Used

	Accuracy	Precision	Recall	F_1
[Schabus et al., 2017] (best)		0.6105	0.6614	0.6098
[Schabus et al., 2017] (LSTM)		0.5685	0.6458	0.6047
Our Single-Model	0.7926	0.6047	0.5612	0.5821
Our Multi-Model	0.8111	0.7071	0.6188	0.6600

- Our Multi-Model is an improvement to the original paper
- Good result although category only applies 28%

References

[depset.ai, www] depset.ai (www). German Word Embeddings. https://deepset.ai/german-word-embeddings. (02/2020).

[Mikolov et al., 2013] Mikolov, T., Corrado, G., Chen, K., and Dean, J. (2013). Efficient estimation of word representations in vector space. pages 1–12.

[Řehůřek and Sojka, 2010] Řehůřek, R. and Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, pages 45–50, Valletta, Malta. ELRA. http://is.muni.cz/publication/884893/en.

[Schabus and Skowron, 2018] Schabus, D. and Skowron, M. (2018). Academic-industrial perspective on the development and deployment of a moderation system for a newspaper website. In Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC), pages 1602–1605, Miyazaki, Japan.

[Schabus et al., 2017] Schabus, D., Skowron, M., and Trapp, M. (2017). One million posts: A data set of german online discussions. In Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), pages 1241–1244, Tokyo, Japan.

Code available at `https://github.com/oliver-pola/OneMillionPostsCorpus`