Erik Körner Ahmad Dawar Hakimi Gerhard Heyer Martin Potthast

erik.koerner@uni-leipzig.de

Casting the Same Sentiment Classification Problem

Abstract

We introduce and study a problem variant of sentiment analysis, namely the "same sentiment classification problem", where, given a pair of texts, the task is to determine if they have the same sentiment, disregarding the actual sentiment polarity. We demonstrate how sentiment data needs to be prepared for this task, and then carry out sequence pair classification using transformer language models.

Code and data: `github.com/webis-de/EMNLP-21`

Motivation

- Focused research on topic-agnosticity, enabling direct observations of the effect of topic and that of agnostic modeling.
- Potentially easing generalization across domains.
- In time, a new paradigm of approaches may emerge (whereas the prevailing one still rules today).
- Distant supervision learning for domains with sparse data, e.g. Same Side Stance Classification. [Stein et al., 2021]

Data

Data requirements:

- Texts with clear stances or sentiments
- Both multiple positive and negative samples about the same topic (e.g. business, ...)
- Multiple topics with enough samples for cross-topic comparisons

Our choice: yelp *business reviews*: contains 6,685,900 user reviews about 192,127 businesses in 22 main categories

Not suitable: Amazon product reviews, IMDb movie reviews.

Training data generation:

- Translate the star rating of 1 to 5 to binary labels, *good* or *bad*; *good* if the rating is above 3 stars
- Filter out businesses that have less than 5 positive and negative reviews
- Sentiment pairs combinations: *good-good*, *good-bad*, *bad-bad*, and *bad-good*.
- Randomly combine pairs of reviews about the same business per pair type

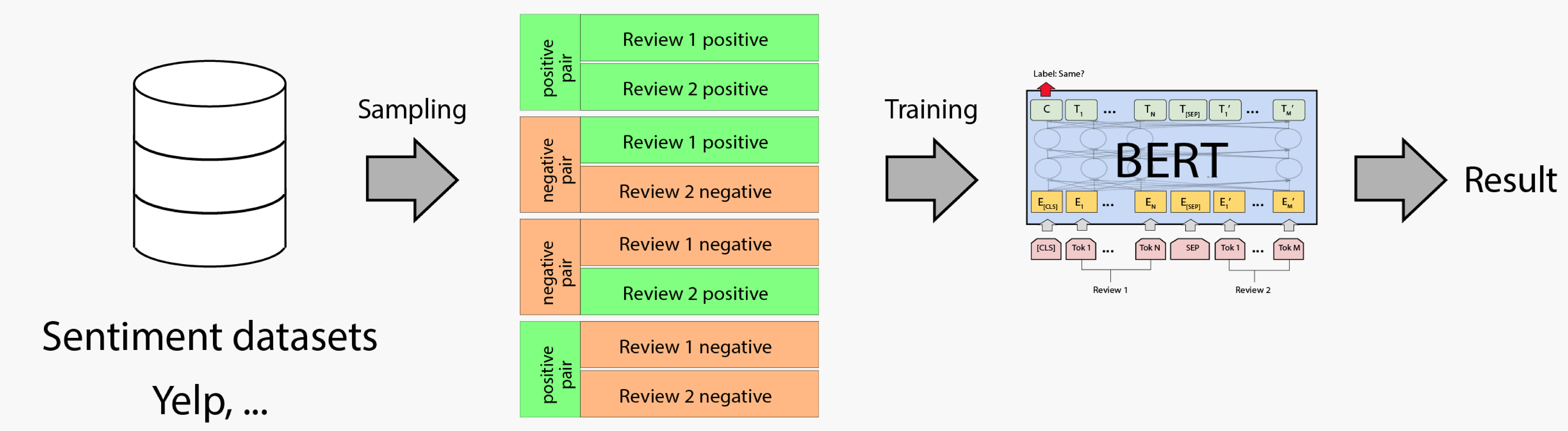
Model

Baselines:

- Count-/TFIDF-vectors not much better than random
- Doc2Vec embeddings & different pooling strategies & different classifiers (SVM, LogReg, ...) → slightly better but only around 57% Acc.
- Good baseline with **Siamese Networks**: 50-dim GloVe embeddings, 256 tokens sequence length [Neculoiu et al. 2016], [Mueller and Thyagarajan 2016]
  → strong baseline, 83% Acc.

Transformer:

- Standard **BERT**-base model [Devlin et al. 2019] for sequence pair classification, default hyper-parameters values
- sequence length of 128 to max. 512 tokens
- fine-tuning for 3 epochs
- *gradient accumulation* to batch small batches (2–6 samples → 64) at 512 sequence length
- newer transformers: DistilBERT, ALBERT performed slightly worse



Evaluation

Evaluation results using model **BERT**-base-uncased, fine-tuning for 3 epochs. Sequence length 128 – 256 tokens (reviews pairs combined, truncated). Bad initial baselines: Count-/TFIDF-vectors, Doc2Vec embeddings and various classifier, never significantly better compared to the random baseline. Strong baseline: **Siamese Networks** [Neculoiu et al. 2016], [Mueller et al. 2016], consistently almost as good as our BERT model in all our experiments.

Pairing	TN	FP	FN	TP	Acc.	Examples
bad-bad	–	–	2,719	14,892	84.6%	17,611
bad-good	15,533	2,098	–	–	88.1%	17,631
good-bad	15,248	2,345	–	–	86.7%	17,593
good-good	–	–	1,537	16,004	**91.2%**	17,541
all*	30,781	4,443	4,256	30,896	87.6%	70,376

Overall performance: 89.1% Accuracy with BERT model, train/valid/test split 80/10/10. Increase of sentiment pairs per business only marginally improved results.

Per-Major Category: 84% to 95% Acc. for evaluations on single categories.

Per-pair type: Siamese baseline achieved best results for *bad-bad* with 86.1%, other pair types at 83%. Our BERT model performed best for *good-good* parings, worse for pair types using *bad* sentiment texts. Decreased variance with increased sequence length, but same ranking.

Category Split		Evaluation Accuracy Per		
	Businesses	(a) Rest	(b) Category split	(c) Single category
Shopping, Local Flavor, Health & Medical, Event Planning & Services, Restaurants, Public Services & Government	279,408	82.4%	79.4% – 85.8%	71.5% – 90.3%
Religious Organizations, Active Life, Arts & Entertainment, Professional Services, Hotels & Travel, Local Services	22,176	84.5%	81.5% – 86.0%	73.6% – 93.0%
Education, Automotive, Bicycles, Mass Media, Home Services	36,624	83.0%	80.9% – 87.6%	72.5% – 95.3%
Pets, Nightlife, Financial Services, Beauty & Spas, Food	89,376	85.2%	84.2% – 92.3%	75.0% – 93.3%

Cross-evaluation results for each fold:
(a) on remaining businesses,
(b) on each other CV fold,
(c) per category not in train fold.
Experiment (c) displays the highest variability as small single categories differ more extremely compared to larger ones or sets of categories.