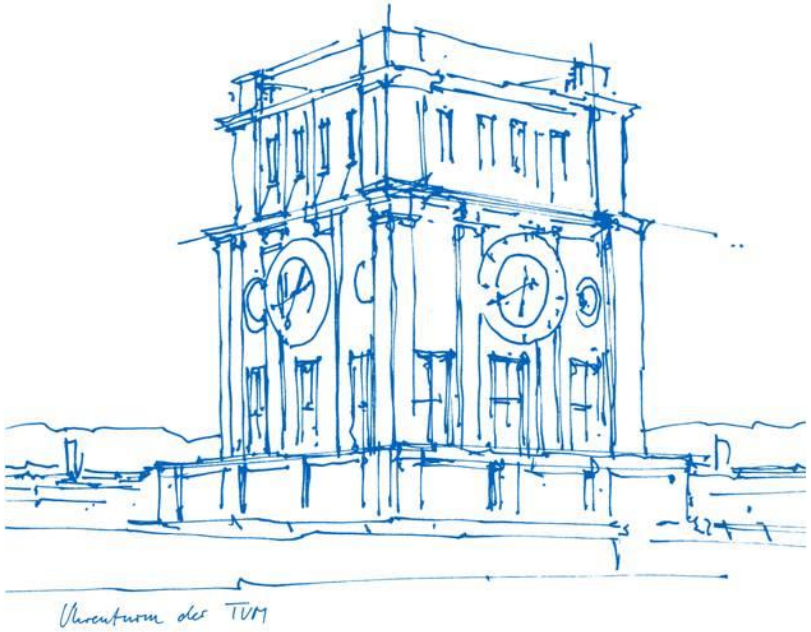
Multi-Lingual Theme Prediction of Customer Reviews Using Deep Pre-Trained Embeddings

Team 06

Michael Sorg

08.05.2019



Task

- Train a **category prediction model** on the Amazon product review dataset **based on XLING embeddings** per review
- Evaluate on German without training on German data
- Fine-tune and evaluate on our Organic Dataset for relevance, entity, and attribute classification

Word embeddings

BERT

„BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”

- Birectional transformer network
- Multilingual version exists

XLING

„Learning cross-lingual sentence representations via a multi-task dual-encoder model“

- Compute dense word vectors (512 dimensions) from sentences
- Embeddings can be fine-tuned (transfer learning)
- Multiple languages

Datasets

- **Kaggle Amazon Fine Food**

(568454, 10)

	Id	ProductId	UserId	ProfileName	HelpfulnessNumerator	HelpfulnessDenominator	Score	Time	Summary	Text
0	1	B001E4KFG0	A3SGXH7AUHU8GW	delmartian	1	1	5	1303862400	Good Quality Dog Food	I have bought several of the Vitality canned d...
1	2	B00813GRG4	A1D87F6ZCVE5NK	dll pa	0	0	1	1346976000	Not as Advertised	Product arrived labeled as Jumbo Salted Peanut...

- **Amazon Review Multilingual**

(678993, 15)

	marketplace	customer_id	review_id	product_id	product_parent	product_title	product_category	star_rating	helpful_votes	total_votes	vine	verified_purchase	review_headline	review_body	review_date
0	DE	10133	RVOG49N0H1FB6	B004TACMZ8	569741360	Bosch GMS120 Ortungsgerät digital multi-Scanner	Home Improvement	5	0	0	N	Y	Super	Delivery took a little bit more then i expecte...	2014-08-01
1	DE	19612	RNCMD6OLTP4HM	1846071224	785505948	The Wheels On The Bus: Favourite Nursery Rhyme...	Books	5	1	1	N	Y	Great compilation	We enjoy listening to the song as preparation ...	2014-12-04

- **Organic dataset**

(8823, 12)

	Author_ID	Author_name	Comment_number	Sentence_number	Domain_Relevance	Sentiment	Entity	Attribute	Sentence	Source_file	Annotator	Aspect
0	Justin-Ma	Justin Ma	521	1	0	NaN	NaN	NaN	Thanks for the thoughtful response.	quora.json	sumit	nan-nan
1	Justin-Ma	Justin Ma	521	2	0	NaN	NaN	NaN	I think we actually have a lot of common groun...	quora.json	sumit	nan-nan
2	Justin-Ma	Justin Ma	521	3	0	NaN	NaN	NaN	All I want to emphasize are my main points: Pr...	quora.json	sumit	nan-nan
3	Justin-Ma	Justin Ma	521	4	9	p	cg	pp	Industrialization is everything about producti...	quora.json	sumit	cg-pp

Questions

- Predict which categories / features ?
- Filter out English reviews in the German dataset?

- Also train XLING embeddings or leave them untouched?

To-do until 22.05

- Implement training and evaluation pipeline