Sentiment Analysis Using BERT and Multi-Instance Learning

Iryna Burak

Pablo Restrepo

June 8, 2020

Presentation Overview: What did we do on the last 2 weeks?

1. **Cross-lingual zero shot learning:**
   1.1 Exploring our Amazon reviews dataset in German.
   1.2 Generating multilingual BERT embeddings for our data.
   1.3 Exploring other multilingual embedding options.
2. **Trying to fine-tune our network.**
   2.1 Running our MilNet with BERT base embeddings.
   2.2 Running our MilNet with BERT multilingual embeddings.
3. **Plan for the next two weeks.**
4. **References.**

1. Cross-lingual zero shot learning:

1. Cross-lingual zero shot learning:

1.1 Exploring our Amazon reviews dataset in German:

Amazon reviews dataset in German.

- 678.993 reviews.
- 34 categories including:

➔ Home Improvement
➔ Books
➔ Digital_Video_Download,
➔ Watches
➔ Mobile_Apps
➔ ...

1. Cross-lingual zero shot learning:

1.1 Exploring our Amazon reviews dataset in German:

As training data, we are using amazon reviews in English from the following categories:

- Grocery and Gourmet Food
- Health and Personal Care
- Beauty

From the 34 categories in our dataset in German, the most relevant for us are:

- Health & Personal Care
- Personal_Care_Appliances
- Beauty
- Grocery

1. Cross-lingual zero shot learning:

1.1 Exploring our Amazon reviews dataset in German:

Number of instances per selected category:

- Personal_Care_Appliances 411
- Health & Personal Care 37
- Grocery 2
- Beauty 1

Sentiment	Instances	Percentage
negative	60	13.3%
neutral	40	8.8%
positive	351	77.8%

➔ Total number of instances: 451

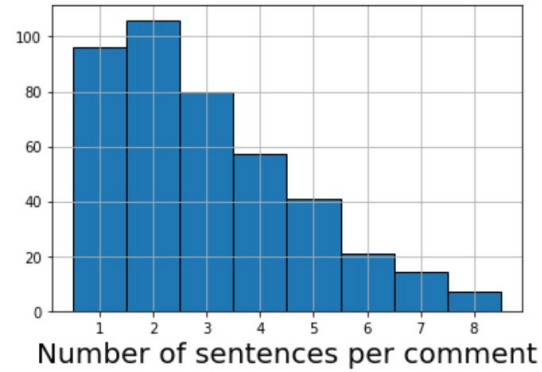
➔ Total number of instances after splitting the comments in sentences (using NLTK): 1.762 (1.35% of our total data)

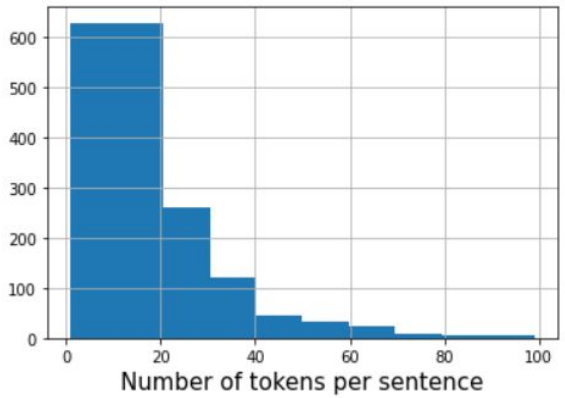
➔ Number of sentences in our dataset in English: 129.867

1. Cross-lingual zero shot learning:

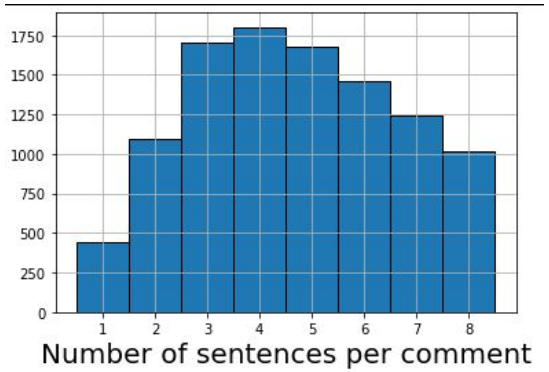
1.1 Exploring our Amazon reviews dataset in German:

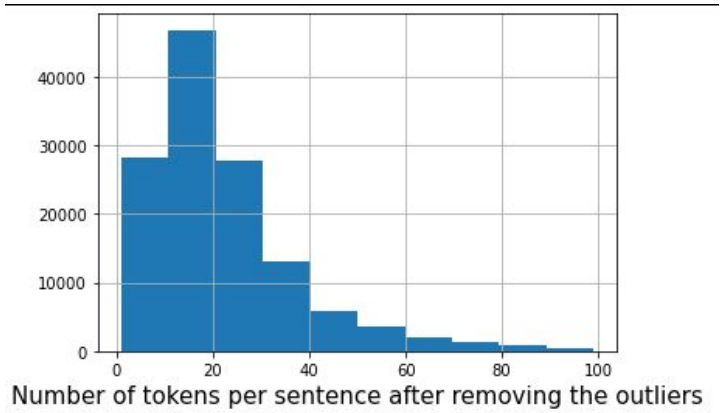
Dataset in German:





Dataset in English (training data):





1. Cross-lingual zero shot learning:

1.2 Generating multilingual BERT embeddings for our data.

Number of instances per selected category:

- Model: “BERT-Base, Multilingual Cased” (recommended model for multilingual models)
- Pooling Strategy: Reduce Mean
- Pooling Layer = -2
- 767 values per sentence
- file size: 2.1 gb

Google BERT

1. Cross-lingual zero shot learning:

1.3 Exploring other multilingual embedding options.

- **Bert Multilingual.**
- **XLM:** improved version of BERT to achieve state-of-the-art results in classification and translation tasks.
- **XLM-RoBERTa (Facebook AI):** provides strong gains over previously released multi-lingual models like mBERT or XLM on downstream tasks like classification, sequence labeling and question answering
- **XLING:** Concatenated Power Mean Embeddings as Universal Cross-Lingual Sentence Representations
- **XLNet:** Generalized Autoregressive Pretraining for Language Understanding.

2. Trying to fine-tune the model

2.1 Running our MilNet with BERT base embeddings.

- Added one additional linear layer to convert 768-dim embeddings to 300-dim embeddings.
- Dropout makes the training process unstable.
- Downsampling factor: 0.15

sentiment	before	after
negative	1372	1372
neutral	1533	1533
positive	12605	1951

2. Trying to fine-tune the model

2.1 Running our MilNet with BERT base embeddings.

metric	amazon	organic
F1 (micro)	0.662	0.464
F1 (macro)	0.643	0.449

Amazon

True	-	0	+
-	58	54	13
0	22	109	16
+	7	50	150

Previous results

metric	amazon	organic
F1 (micro)	0.672	0.450
F1 (macro)	0.530	0.447

Organic dataset

True	-	0	+
-	540	876	42
0	361	1295	317
+	178	909	526

2. Trying to fine-tune the model

2.2 Running our MilNet with BERT multilingual embeddings.

Base BERT

metric	amazon-german	amazon-english	organic
F1 (micro)	-	0.662	0.464
F1 (macro)	-	0.643	0.449

Multilingual BERT

metric	amazon-german	amazon-english	organic
F1 (micro)	0.47	0.585	0.430
F1 (macro)	0.40	0.557	0.430

3. Plan for the next two weeks:

- Update our baselines to work with data in German
- Generate embeddings for our data using different techniques (XLM-RoBERTa,XLING,XLNet).
- Train our MilNet using different embeddings and compare results.
- Try to improve our numbers as much as possible.

4. References

- Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova: “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”, 2018; [http://arxiv.org/abs/1810.04805 arXiv:1810.04805].
- Multi-lingual models, Huggingface, https://huggingface.co/transformers/multilingual.html
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, Veselin Stoyanov: “Unsupervised Cross-lingual Representation Learning at Scale”, 2019; [http://arxiv.org/abs/1911.02116 arXiv:1911.02116].
- Andreas Rücklé, Steffen Eger, Maxime Peyrard, Iryna Gurevych: “Concatenated Power Mean Word Embeddings as Universal Cross-Lingual Sentence Representations”, 2018; [http://arxiv.org/abs/1803.01400 arXiv:1803.01400].
- Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, Quoc V. Le: “XLNet: Generalized Autoregressive Pretraining for Language Understanding”, 2019; [http://arxiv.org/abs/1906.08237 arXiv:1906.08237].