Sentiment Analysis Using BERT and Multi-Instance Learning

Iryna Burak

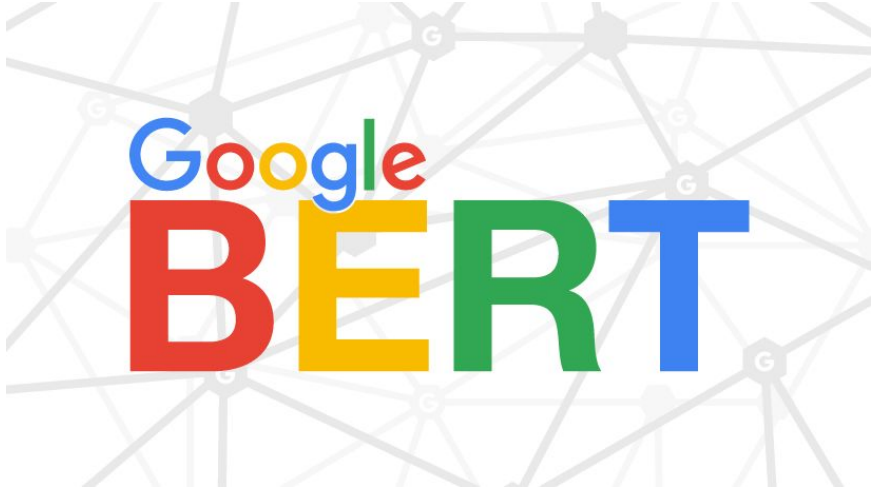
Pablo Restrepo

May 11, 2020

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

1. **Getting Embeddings for our data using BERT:**
   1.1 Analyzing and preparing our data.
   1.2. Exploring BERT options and hyperparameters.
   1.2. Getting embeddings for our data.
2. **Implementing our first MILNET (Multi Instance Learning Network) using PyTorch**
3. **Plan for the next two weeks.**
4. **References.**

1. Getting Embeddings for our data using BERT:



1.1 Analyzing and preparing our data

Amazon reviews Dataset:

Categories:

- Grocery and Gourmet Food -> 198.502 rows
- Health and Personal Care -> 151.254 rows
- Beauty -> 346.355 rows

After filtering those containing the word: “organic”:

- Grocery and Gourmet Food -> 9.962 rows
- Health and Personal Care -> 3.272 rows
- Beauty -> 2.227 rows

TOTAL: 15.561 rows

1.1 Analyzing and preparing our data

How to asses the sentiment of a comment?

- if rating < 3 -> Negative Sentiment
- if rating = 3 -> Neutral Sentiment
- if rating > 3 -> Positive Sentiment

overall	0
1.0	676
2.0	704
3.0	1537
4.0	3074
5.0	9570

- Negative sentiment: 1.380 -> 8,86%
- Neutral sentiment: 1.537 -> 9,87%
- Positive sentiment: 12.644 -> 81,25%

1.1 Analyzing and preparing our data

Processing our data:

Splitting the comments into sentences using NLTK:

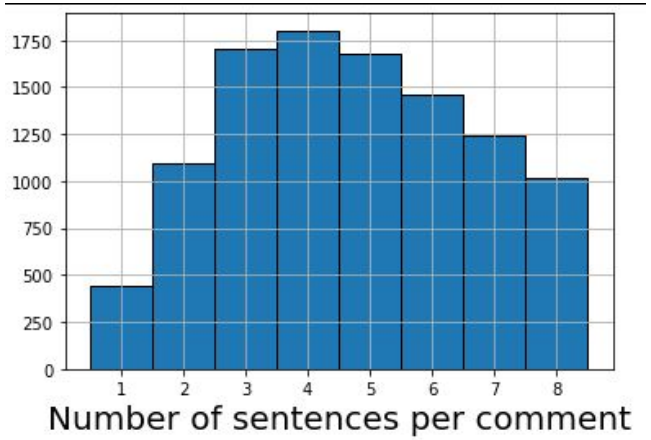
Index	comment_id	sentence_id	comment_rating	comment_sentiment	sentence_text	number_of_tokens	dataset_name
0	0	0	1	n	I have thin fine hair and I am always searching for the "miracle" to a beautiful head of hair.	22	beauty
1	0	1	1	n	This is not it.	5	beauty
2	0	2	1	n	It may be organic and have less surfactants but, it's the thinnest shampoo I've ever used.	20	beauty
3	0	3	1	n	Can't get it from bottle to hand to head without losing it down the drain.	17	beauty
4	0	4	1	n	Stupid purchase.	3	beauty
5	0	5	1	n	Oh well, next time.	6	beauty
6	1	0	2	n	I BOUGHT THIS Avalon Organics: Biotin B Comple…	39	beauty
7	2	0	4	p	So i have been using the Avalon Organics both Shampoo and Conditioner for a few months now.	18	beauty
8	2	1	4	p	I have seen a pretty good result.	8	beauty
9	2	2	4	p	My hair has definitely gotten a bit thicker and stronger.	11	beauty
10	2	3	4	p	I notice less hair falling out.	7	beauty

Total number of sentences: 129.867

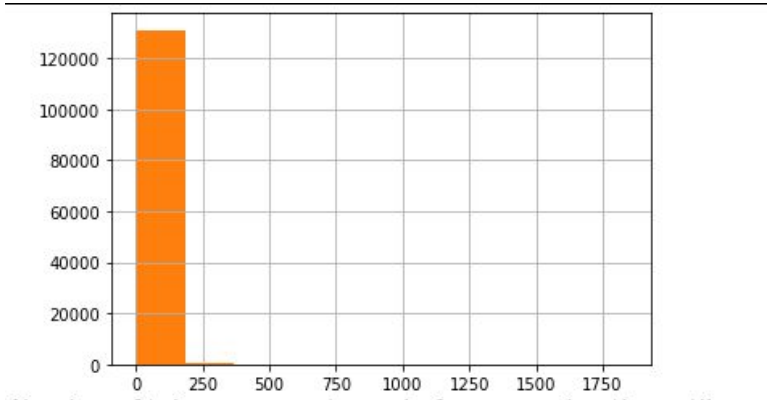
1.1 Analyzing and preparing our data

Tokens per sentence:

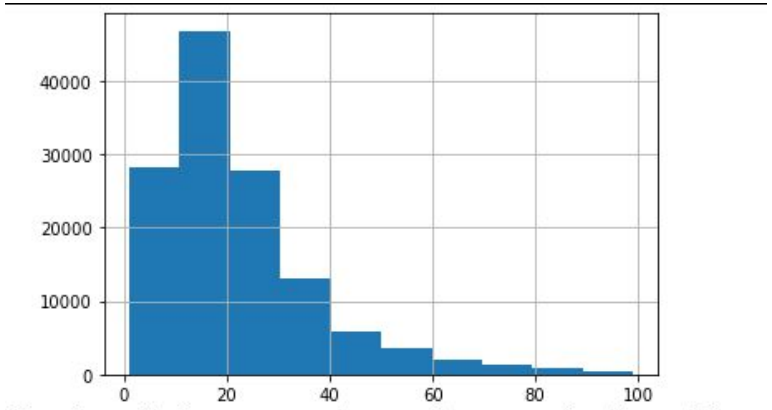
- BERT can only handle sentences with a maximum of 512 tokens!
- We have sentences with more than 1000 tokens
- 0.2% of the dataset has more than 100 tokens.



Number of sentences per comment



Number of tokens per sentence before removing the outliers



Number of tokens per sentence after removing the outliers

1.2. Exploring BERT options and hyperparameters.

- What Tool to use?
- What Bert model to use?
- Pooling layer and pooling strategy?

1.2. Exploring BERT options and hyperparameters.

What Tool to use?

- Using BERT library.
- BERT as a Service.
- Other Tools for getting BERT embeddings: SBERT (rich embeddings for BERT).

We decided to use “BERT as a Service”.

1.2. Exploring BERT options and hyperparameters.

What Bert model to use?

- Many pretrained models available
- Huggingface Transformers library has 24 different BERT pretrained models.

We decided to use “bert-base-uncased”.

1.2. Exploring BERT options and hyperparameters.

Pooling Layer and Pooling Strategy?

- For each token of we have 12 separate vectors.
- Different layers of BERT encode very different kinds of information

- To get individual vectors, we need to combine some of the layer vectors, which layer or combination of layers provides the best representation?
- **This is application dependent!.**

We decided to use BERT as a Service default parameters: pooling strategy: Reduce Mean Pooling layer = -2.

1.3. Getting embeddings for our data.

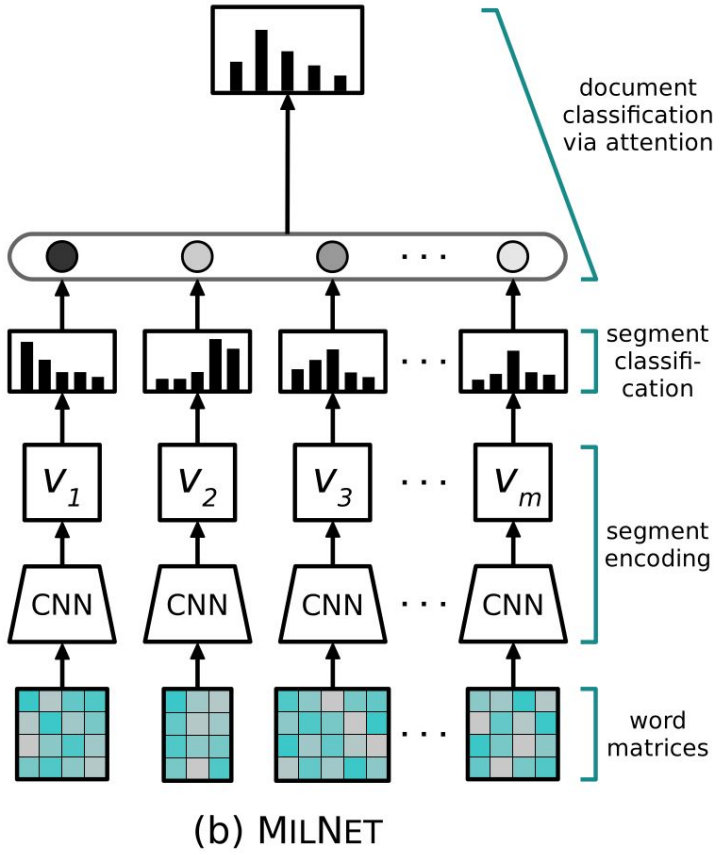
Final Configuration:

- Tool: Bert as a Service
- Model: “bert-base-uncased”
- Pooling Strategy: Reduce Mean
- Pooling Layer = -2
- 767 values per sentence
- file size: 2.1 gb

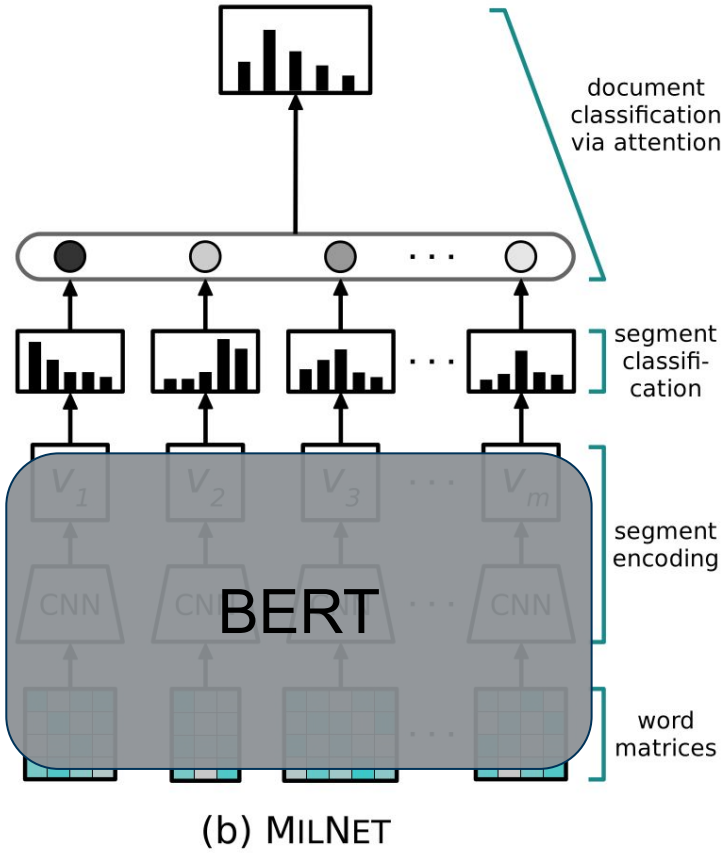
2. Implementing MilNet

- The authors of the article implemented MilNet in Lua using torch.
- We decided to use PyTorch.
- Lua implementation is still a good reference.

2. Implementing MilNet



(b) MILNET



(b) MILNET

2. Implementing MilNet

- Our implementation doesn’t crash.
- It overfits on a small dataset.

3. Plan for the next two weeks

- Implement and run our baselines.
- Train the model on the whole dataset.
- More data analysis!

4. References

- Reimers N. & Gurevych I (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks, https://arxiv.org/abs/1908.10084
- Pretrained models,Tansformers 2.0.9 documentation, https://huggingface.co/transformers/pretrained_models.html
- McCormick C. (2019), BERT word Embeddings Tutorial, https://mccormickml.com/2019/05/14/BERT-word-embeddings-tutorial/
- Xiao, H. BERT as a Service, https://github.com/hanxiao/bert-as-service