The Augmented Social Scientist

Tips & Tricks

SICSS-Paris, June 2023

Using an LLM to Annotate Text

Using a supervised machine learning algorithm to automatically annotate text is easy

But there are pitfalls you'd better avoid, and best practices

⇒ Last moment of this tutorial = a few tips and tricks

1. **Defining Categories to Annotate**
2. **Creating an appropriate test set**
3. **Designing a Training Strategy**
4. **Active Learning**
5. **"What can I do if I have bad validation scores?"**

1. Defining Categories

1. Defining Categories

You want to train an LLM to annotate automatically some text for you.

How do you decide on the categories you want to use?

- **Theory**: any category you deem relevant
- **Practice:** more complex, and depends on a set of parameters

1. Defining Categories

1. Design a temporary coding scheme

1. Defining Categories

1. Design a temporary coding scheme
2. Annotate a small (10-100) units of text, sampled at random
3. Revise coding scheme and annotate again

1. Defining Categories

1. Design a temporary coding scheme
2. Annotate a small (10-100) units of text, sampled at random
3. Revise coding scheme and annotate again
    a. Pause: Is it working? Note down issues, hesitations (w/ examples)

1. Defining Categories

1. Design a temporary coding scheme
2. Annotate a small (10-100) units of text, sampled <u>at random</u>
3. Revise coding scheme and annotate again

Are you happy with your coding scheme?

- **No**: Start over
- **Yes**: Continue annotating + write detailed guidelines

1. Defining Categories

Question: How refined can my categories be?

1. Defining Categories

Question: How refined can my categories be?

- "It depends on the data", but
  - Do not limit yourself to the obvious (go for semantics, not lexicon)
  - Do not expect an algorithm to do better than a skilled human

  ⇒ TRY!

  (And ask yourself: could you easily convey the idea to a colleague?)

1. Defining Categories

Question: What is the correct unit of analysis?

1. Defining Categories

Question: What is the correct unit of analysis?

- “It depends on the data”, but
  - A sentence is an obvious candidate, but you will lose lots of context

1. Defining Categories

Question: What is the correct unit of analysis?

- “It depends on the data”, but
  - A sentence is an obvious candidate, but you will lose context
  - A paragraph is a second obvious candidate, but
    - You do not always have paragraphs clearly delimited
    - It will take more examples to train a model

1. Defining Categories

Question: What is the correct unit of analysis?

- "It depends on the data", but
  - A sentence is an obvious candidate, but you will lose context
  - A paragraph is a second obvious candidate, but
  - A longer text then seems great, but
    - It will take many more examples to train a model
    - The models "stop" reading after a few hundreds tokens

1. Defining Categories

Question: One multi-class classifier, or several binary classifiers?

1. Defining Categories

Question: One multi-class classifier, or several binary classifiers?

- "It depends on the data", but
  - Binary models will be easier to train
  - If you do a lot of binary classifiers instead of a multiclass,
    - You will get more refined…
    - Or more ambivalent results

1. Defining Categories

Question: How much should you annotate?

- "It depends on the data", but
  - A binary classifier could only need a few dozen examples

1. Defining Categories

Question: How much should you annotate?

- "It depends on the data", but
  - A binary classifier could only need a few dozen examples
  - Annotation is intellectually healthy

1. Defining Categories

Question: How much should you annotate?

- "It depends on the data", but
    - A binary classifier could only need a few dozen examples
    - Annotation is intellectually healthy
    - There are shortcuts to save massive amounts of time ("active learning")
        - If your simple classifier does not work after 8h of annotation, reconsider

2. Creating an appropriate test set

- Should be representative of the corpus
- No intersection with the training set
- Double check it!

3. Training strategy

- Training parameters

3. Training strategy

- Training parameters
  - Learning rate (lr)

3. Training strategy

- Training parameters
  - Learning rate (lr)
  - Number of epochs (n_epochs)

3. Training strategy

- Training parameters
  - Learning rate (lr)
  - Number of epochs (n_epochs)

- Training set (downsampling/oversampling)

3. Training strategy

What to remember

1. **No general rules, you need to try**
2. **No need to spend too much time**

4. Active learning

A very common problem: unbalanced dataset

e.g. 10% positive, 90% negative

Objective: obtain 300 positive sentences for training

- If random sampling -> manual annotation of 3000 sentences
- How to obtain more positive samples with less manual annotation?

4. Active learning

Active learning: use intermediate models to find more positive samples

- How does it work?

4. Active learning

Active learning: use intermediate models to find more positive samples

- How does it work?

2 strategies of active learning

- Most probable (max probability)
- Most ambiguous (max entropy)

4. Active learning

Active learning: use intermediate models to find more positive samples

- How does it work?

2 strategies of active learning

- Most probable (max probability)
- Most ambiguous (max entropy)

⚠️⚠️⚠️ **Hold-out (representative) test set**

5. What to do if you have bad validation scores

- Double check (again) the test set
- Add training data
- Read some model predictions to understand

Conclusion

- Diving into new sources of data
- With your own research questions + tagging scheme

Resources:

- Package: https://github.com/rubingshen/Replication_Augmented
- Article (*SMR*, 2022)
- Google Colab tutorial