RUTGERS

PII Detection & Removal From Educational Texts

Master's Project Proposal

Lovedeep Singh

Overview

- Problem Description
- Current Trends / Research
- State-of-the-Art
- Project Work & Structure
- Metrics / Benchmarks
- Key Challenges
- Vision and Future Scope
- References

Problem Statement

Develop automated techniques to detect and remove PII from educational data.

Problem:

- PII is barrier to analyze and create open datasets for educational advancement.
- Hinders anonymous reviews / peer feedback.
- Currently, most reliable method is Manual review and cleaning.

How this solution and tool will help?

- Identify, detect and remove PII from educational texts and documents.
- Safeguarding Student Privacy

Existing Tools

- Enterprise Tools
  - Varonis
  - pi-tools.com
  - Presidio (Microsoft)
  - Amazon Comprehend
  - Google Cloud Data Loss Prevention (DLP)

State-of-the-Art

- Named Entity Recognition (NER)
  - spaCy: https://spacy.io/
  - NLTK: https://www.nltk.org/
  - Stanford NLP: https://stanfordnlp.github.io/CoreNLP/
- BERT (Bidirectional Encoder Representations from Transformers)
- RoBERTa (Robustly Optimized BERT Pre Training Approach)

Problem with existing solutions

- Mostly work on structured data (JSON, CSV's)
- Expensive
- General Purpose
- NER based techniques more effective for PII with common formats (emails, phone numbers)
  - Struggle with labelling names
  - Distinguishing between sensitive names (student's names) and public names (cited Author, etc)

Project Work and Timelines

1.	Data Collection and Cleansing	Week 1
2.	Feature Engineering	Week 2
3.	Establishing Baseline Models	Week 3
4.	Ensemble Methods, NN, use case fit (fine tuning, hyperparameter optimizations)	Week 4-5
5.	BERT Pre-training, Transformers, explore novel approaches	Week 6-8
6.	Model Validation, Testing and deployment	Week 9-10
7.	Open Source / Web GUI	(if time permits)

Datasets

- Kaggle Dataset - 22,000 essays written by students enrolled from massively open online course.
  - All essays written in response to a single assignment prompt
  - JSON format
  - Documents tokenized using the SpaCy English tokenizer.
  - Structure:
    - **(int):** the index of the essay
    - **document** (int): an integer ID of the essay
    - **full_text** (string): a UTF-8 representation of the essay
    - **tokens** (list)
      - (string): a string representation of each token
    - **trailing_whitespace** (list)
      - (bool): a boolean value indicating whether each token is followed by whitespace.
    - **labels** (list)
      - (string): a token label in BIO (Beginning, Inner, Outer) format
        - PII type is prefixed with “B-” when it is the beginning of an entity
        - “I-” if the token is a continuation of an entity
        - “O” for tokens that are not PII

BIO encoding	Michel	Jordan	would	choose	Bush
	B-PER	I-PER	O	O	B-PER

Datasets

Sample Data:

“Design Thinking for innovation reflexion-Avril 2021-Nathalie Sylla Challenge & selection The tool I use to help all stakeholders finding their way through the complexity of a project is the mind map. What exactly is a mind map? According to the definition of **Buzan T. and Buzan B.** (1999, **Dessine-moi l'intelligence**. **Paris: Les Éditions d'Organisation**.), the mind map (or heuristic diagram) is a graphic representation technique that follows the natural functioning of the mind and allows the brain's potential to be released. Cf Annex1 This tool has many advantages: • It is accessible to all and does not require significant material investment and can be done quickly • It is scalable • It allows categorization and linking of information • It can be applied to any type of situation: notetaking, problem solving, analysis, creation of new ideas • It is suitable for all people and is easy to learn • It is fun and encourages exchanges • It makes visible the dimension of projects, opportunities, interconnections • It synthesizes • It makes the project understandable • It allows you to explore ideas The creation of a mind map starts with an idea/problem located at its center. This starting point generates ideas/work areas, incremented around this center in a radial structure, which in turn is completed with as many branches as new ideas. This tool enables creativity and logic to be mobilized, it is a map of the thoughts. Creativity is enhanced because participants feel comfortable with the method. Application & Insight I start the process of the mind map creation with the stakeholders standing around a large board (white or paper board). In the center of the board, I write and highlight the topic to design. Through a series of questions, I guide the stakeholders in modelling the mind map. I adapt the series of questions according to the topic to be addressed. In the type of questions, we can use: who, what, when, where, why, how, how much. The use of the “why” is very interesting to understand the origin. By this way, the interviewed person frees itself from paradigms and thus dares to propose new ideas / ways of functioning. I plan two hours for a workshop. Design Thinking for innovation reflexion-Avril 2021-Nathalie Sylla After modelling the mind map on paper, I propose to the participants a digital visualization of their work with the addition of color codes, images and interconnections. This second workshop also lasts two hours and allows the mind map to evolve. Once familiarized with it, the stakeholders discover the power of the tool. Then, the second workshop brings out even more ideas and constructive exchanges between the stakeholders. Around this new mind map, they have learned to work together and want to make visible the untold ideas. I now present all the projects I manage in this type of format in order to ease rapid understanding for decision-makers. These presentations are the core of my business models. The decision-makers are thus able to identify the opportunities of the projects and can take quick decisions to validate them. They find answers to their questions thank to a schematic representation. Approach What I find amazing with the facilitation of this type of workshop is the participants commitment for the project. This tool helps to give meaning. The participants appropriate the story and want to keep writing it. Then, they easily become actors or sponsors of the project. A trust relationship is built, thus facilitating the implementation of related actions. Design Thinking for innovation reflexion-Avril 2021-Nathalie Sylla Annex 1: Mind Map Shared facilities project”

Datasets

- Synthetic Dataset Generation
  - AI generated datasets.
  - Use Mixtral-8x7B Large Language Model (Open Source LLM) along with others
- External Open Source and standard Datasets
  - CoNLL-2003
    - Eight files covering two languages: **English** (sourced from Reuters Corpus) and **German** (sourced from ECI Multilingual Text Corpus).
    - For each of the languages: we have a training file, a development file, a test file and a large file with unannotated data.
  - OntoNotes 5.0
    - Three languages: English, Chinese, and Arabic
    - With structural information (syntax and predicate argument structure) and shallow semantics (word sense linked to an ontology and coreference)
    - Sourced from various genres of news, conversational telephone speech, weblogs, usenet newsgroups, broadcast, talk shows.

Metrics / Benchmarks

- Metrics:
  - Recall, Precision, F1 score, Accuracy, MCC
  - Efficiency and Resource Utilization
- Can be benchmarked against standard NLP and NLU Datasets [Task Specific]
  - CoNLL-2003
  - OntoNotes 5.0
- Can be benchmarked or tested on test data from Kaggle [Noisy ; Resource Constrained Environment]

Key Challenges

- Data Collection and Generation
- Extracting and categorizing sensitive v/s non-sensitive names.
  - Multilingual names, cultural names
  - Rare / unseen names
- Developing Novel Methods over existing models
- High Penalty for False Negatives
  - Precision-Recall Trade-off
  - Potentially leaking sensitive information has higher impact than giving false positive results.
- Data Quality

Vision and Future Scope

- Open Source API / Model
- Web GUI
- Contextual Understanding / Semantic Similarities
- Zero / Few Shot Learning
- More Applications:
  - Refactoring / Releasing code for open source projects
  - Medical Domain
  - Compliance [HIPAA, GDPR]
- Improving the system over-time (with rewards-based incremental self-training)

References

- Automated Concatenation of Embeddings for Structured Prediction
  - Xinyu Wang, Yong Jiang, Nguyen Bach, Tao Wang
- GoLLIE: Annotation Guidelines improve Zero-Shot Information-Extraction
  - Oscar Sainz, Iker García-Ferrero, Rodrigo Agerri, Oier Lopez de Lacalle, German Rigau, Eneko Agirre
- BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding
  - *NAACL 2019:* Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova
- Deep contextualized word representations
  - *NAACL 2018* : Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, Luke Zettlemoyer

RUTGERS

THANK YOU