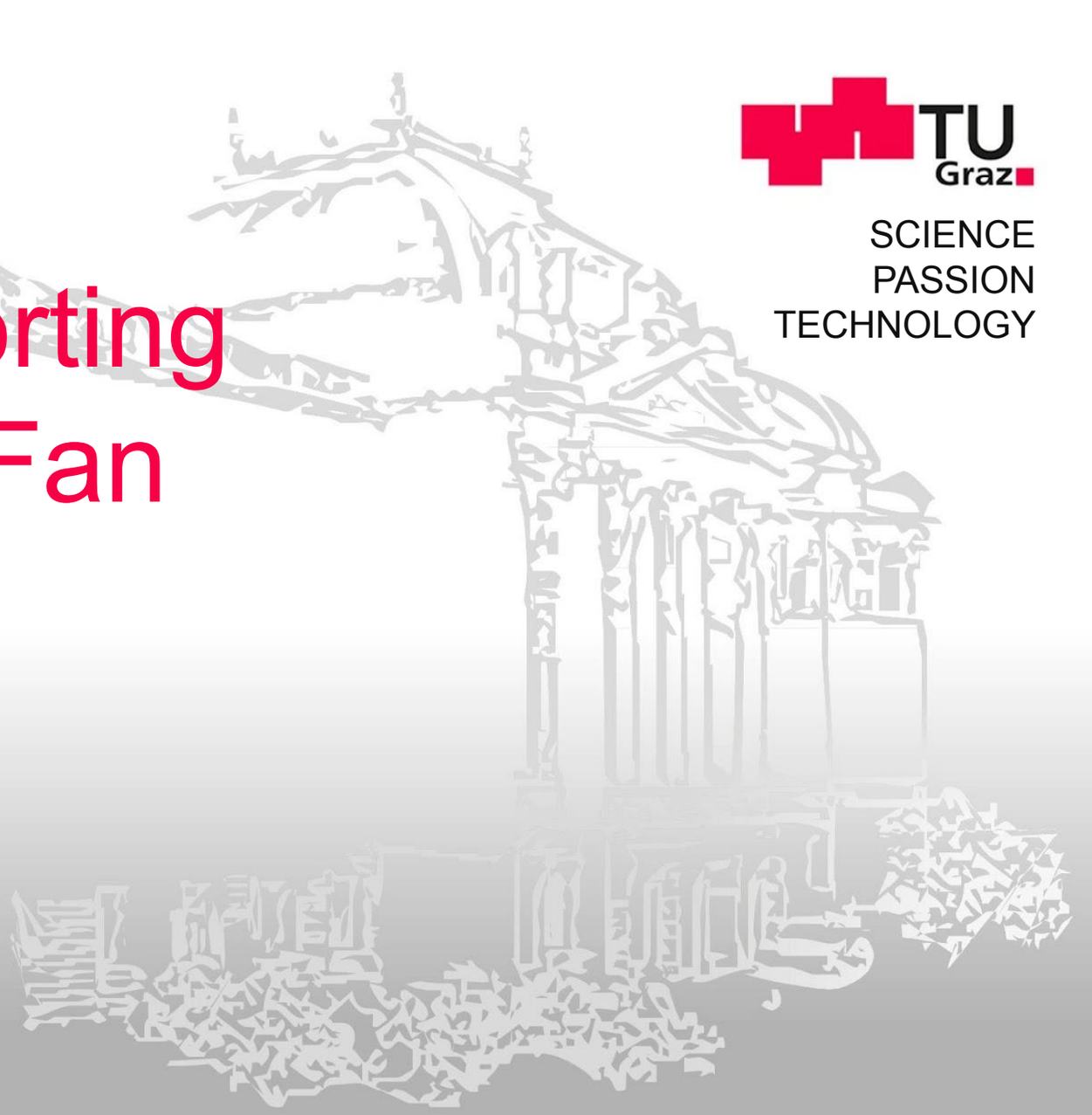


TU Graz

SCIENCE
PASSION
TECHNOLOGY

Identifying Discomforting Content in Lengthy Fan Fiction

Seminar/Project Data Science

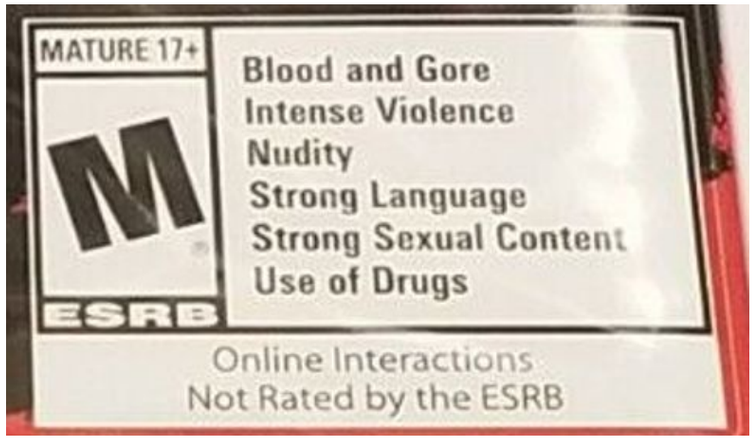
Daniel Hebenstreit

ISDS
INSTITUTE OF INTERACTIVE SYSTEMS AND DATA SCIENCE

Introduction

- Discomforting content is common in literature
  - abduction of children in "Rumpelstiltskin"
  - burning witches in "Hansel and Gretel"

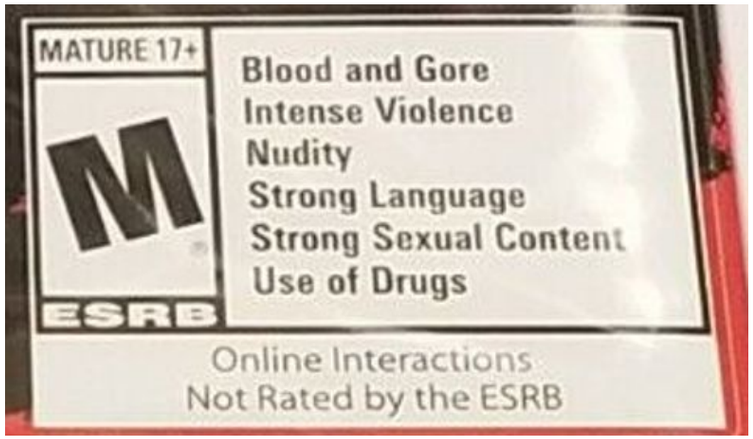
Introduction



Read Dead Redemption Back Cover

- Discomforting content is common literature
  - abduction of children in "Rumpelstiltskin"
  - burning witches in "Hansel and Gretel"
- Perception of what is discomforting is different for every individual
  - often referred to as triggering content
  - traditional media protects us via trigger warnings
    - e.g. use of violence, FSK18+...

Introduction



Read Dead Redemption Back Cover

- Discomforting content is common literature
  - abduction of children in "Rumpelstiltskin"
  - burning witches in "Hansel and Gretel"
- Perception of what is discomforting is different for every individual
  - often referred to as triggering content
  - traditional media protects us via trigger warnings
    - e.g. use of violence, FSK18+...
- Fan fiction is mostly unregulated
  - subgroups vulnerable to potentially triggering content

Dataset

- Data of the ‘PAN 2023’ trigger detection challenge
  - 307.102 Fan fiction works from AO3 + respective trigger labels
  - one or multiple trigger labels per document
  - top challenge solutions already published
    - predictions on the validation set of 17.104 works
    - measured with F1-micro and F1-macro score

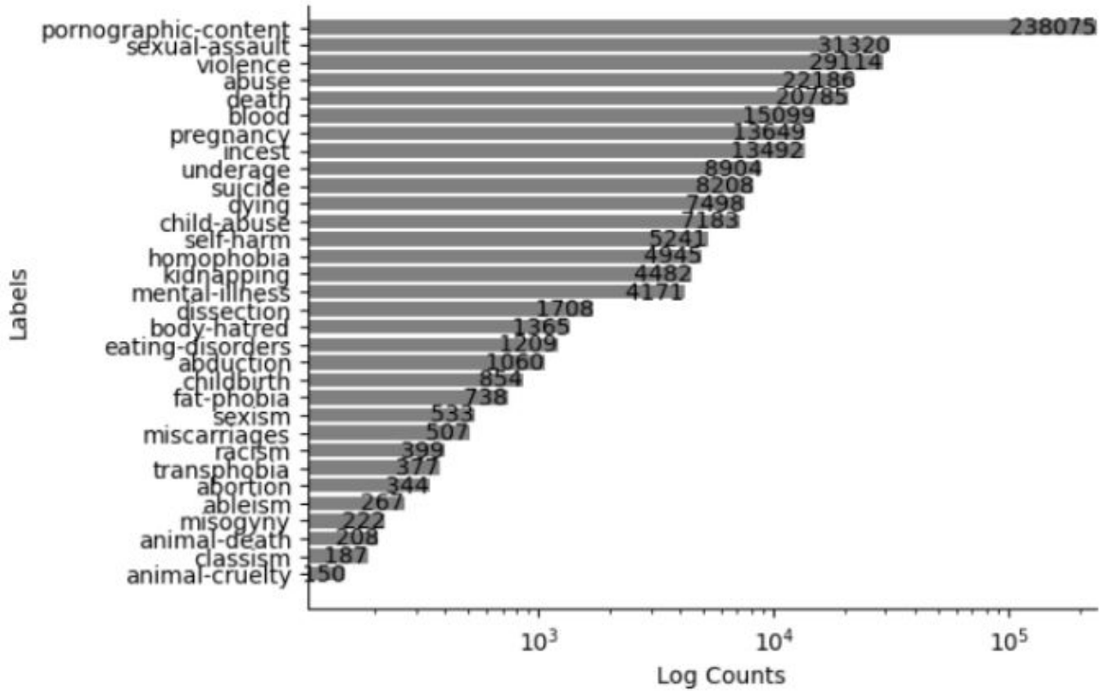
Dataset

- Data of the ‘PAN 2023’ trigger detection challenge
  - 307.102 Fan fiction works from AO3 + respective trigger labels
  - one or multiple trigger labels per document
  - top challenge solutions already published
    - predictions on the validation set of 17.104 works
    - measured with F1-micro and F1-macro score
- Goal is to advance the research in this area
  - can we improve the state of the art?
    - w.r. to the metric
    - w.r. to the computational complexity

Dataset Properties

Why is this dataset intriguing to analyse?

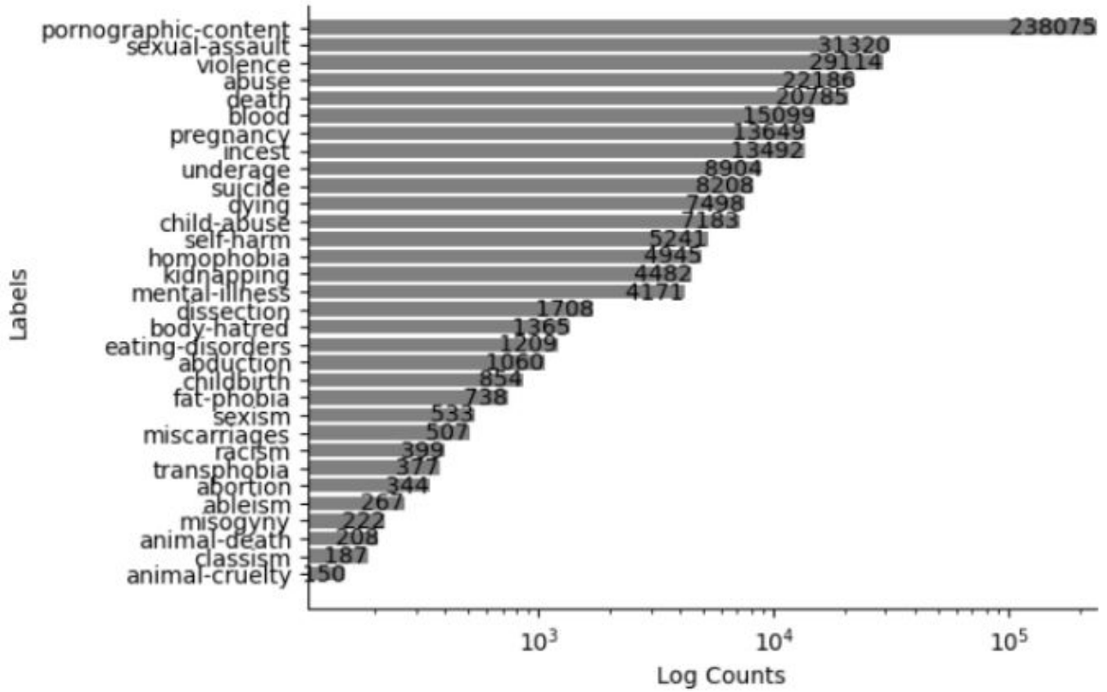
- Label Imbalance
  - some labels more frequent
    - pornographic-content (77%)
    - animal-cruelty (0.0004%)



Dataset Properties

Why is this dataset intriguing to analyse?

- Label Imbalance
  - some labels more frequent
    - pornographic-content (77%)
    - animal-cruelty (0.0004%)
- Length of Documents
  - surpass SOTA model input capacities
    - Bert (512 tokens) BigBird (4096 tokens)
  - median 2.124 words
  - maximum 13.253 words



Preprocessing

- Removing HTML code
- Lowercasing
- Removing special characters, numbers, multiple spaces
- *Transformers*: tokenization
- *BoW*: lemmatization and n-grams

Bag-of-Words

- Tf-idf encodings of uni-grams
- Linear Support Vector Machine
  - low computation time
  - typically excellent results in text classification tasks
  - one-versus-all strategy
- Implicitly solves length of text problem
- Class imbalance?
  - inverse label frequencies as weights

Transformers

[Tong Wu et al.
Distribution-Balanced Loss for Multi-Label Classification in Long-Tailed Datasets.
2021. arXiv: 2007.09654]

- Class Imbalance
  - distribution balanced loss
  - takes information of label co-occurence into account
  - assign lower weight on "easy-to-classify" negative instances

Transformers

[Tong Wu et al. Distribution-Balanced Loss for Multi-Label Classification in Long-Tailed Datasets. 2021. arXiv: 2007.09654]

- Class Imbalance
  - distribution balanced loss
  - takes information of label co-occurence into account
  - assign lower weight on "easy-to-classify" negative instances
- Length of text problem
  - training longformers too expensive
  - Bert (512 tokens) - information loss?
    - truncation
    - summarization
    - combination

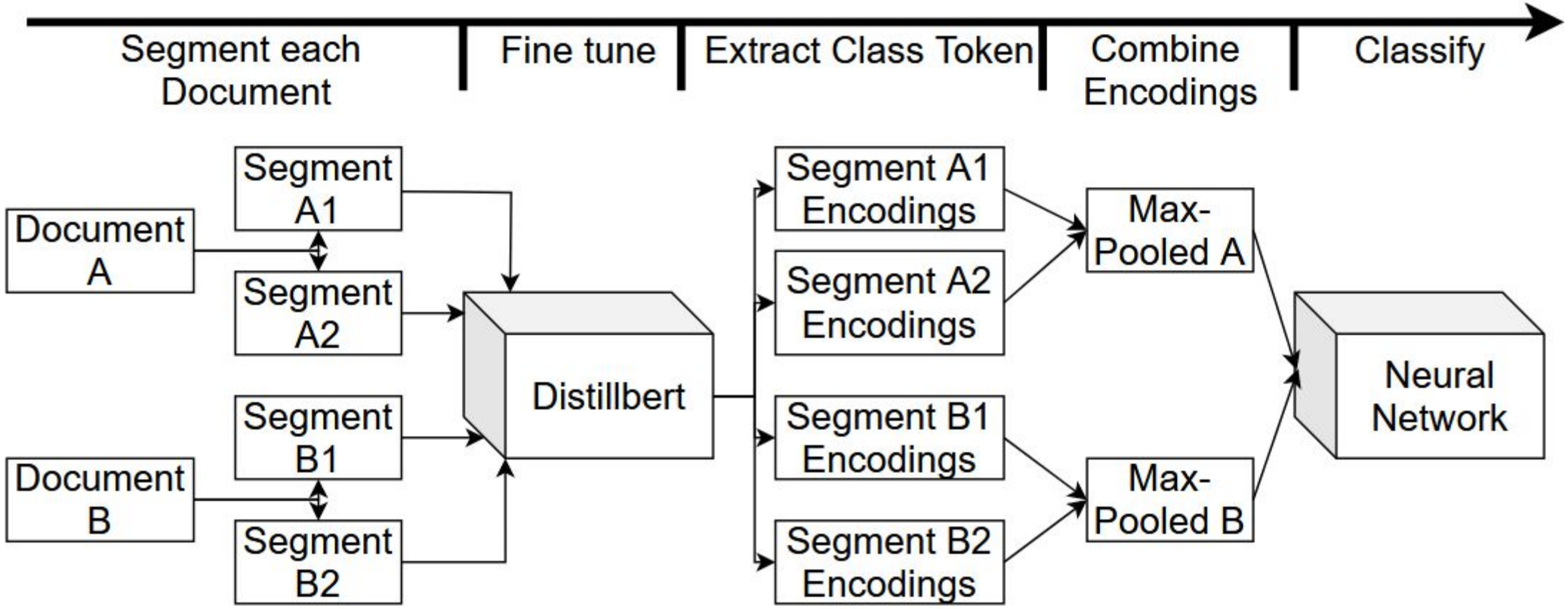
Transformers: TextGuide

Krzysztof Fiok et al. “Text Guide: Improving the Quality of Long Text Classification by a Text Selection Method Based on Feature Importance”. IEEE Access 9 (2021)

- Summarize Text based on important Keywords
- Extract features associated to highest coefficents of our BoW model
  - resembles important words
    - e.g. top 5 words "homophobia": gay, faggot, homophobic, fag, queer
- Iterate each text and search for important words
  - If we match a keyword, extract it and its k-Neighbours
  - Add them to the Summary, repeat until Max Input Size is reached
  - e.g. k=1, I_Word=”killed”, Sentence: “Harry killed Voldemort last year”

    -> Harry killed Voldemort

Transformers: Max-Pooled Bert Encodings



Results

- Max Pooled Encodings **superior**
- TextGuide as computationally **cheap alternative** (~7 times faster)
- SVM/Bert comparably poor performance

Table 1: *Validation Scores of the Applied Approaches.*

Approach	F1-Macro	F1-Micro
PAN 2023 Baseline	0.2575	0.7274
SVM	0.2256	0.7150
SVM Balanced	0.2768	0.5644
Hierarchical SVM	0.2657	0.7201
Bert	0.1586	0.6754
TextGuide	0.3366	0.7465
PAN 2023 Top Solution	0.3720	0.7360
Max-Pooled Encodings	**0.4198**	**0.7613**