RuATD - Artificial Text Detection

—

NN Methods final project

Anna Aksenova & Ekaterina Taktasheva

Plan

1. Why?
2. Who?
3. Data/Baseline/Metrics
4. How?
5. References

Why?

Business and Scientific Value

Business

- fake news detection
- fake product reviews detection
- spamming/phishing detection
- copyright

Science

- generative models benchmark
- detecting features for the future improvement of ATG

Who?

Our team

Anna Aksenova

– baselines

– NN-based methods

Ekaterina Taktasheva

– data analysis

– feature-based methods

Data/Baselines/Metrics

Competition

RuATD dataset

- train / val / test
- 129065 / 21511 / 64533 texts
- binary classification (H vs. M)

Baseline:

- tf-idf + LogReg
- fine-tuning of ruBERT

Metrics: accuracy

How?

Set of Experiments

1. Statistical methods
   a. Feature-based methods:

      Using common lexical/linguistic text features (readability, diversity, etc.) as indicators of text complexity and coherence

   b. Stylometry-based analysis:

      Analysis of the stylistic features as indicators of author style

2. NN-based methods
   a. Fine-tuned ruRoBERTa (Roberta as SOTA on TweepFake dataset)
   b. GPT models (/other generative models like T5)

3. Combination of statistical and neural features

References

References

1. Automatic Detection of Machine Generated Text: A Critical Survey (Jawahar 2020)
    a. An overview of NN-based methods for ATD
    b. RoBERTa (TweepFake) mistakes analysis
2. Computer-Generated Text Detection Using Machine Learning: A Systematic Review (Beresneva 2016)
    a. Use of lexicographical and statistical features for ATD
3. Defending Against Neural Fake News (Zellers et al. 2019)
    a. GROVER – model for text generation and ATD
    b. GPT2 model as discriminator used for text classification
4. Giant Language model Test Room (Gehrmann 2019)
    a. Use GPT to detect text generated by GPT
    b. Frequency analysis is important

Links

Github

Trello