HOAX DETECTION ON WIKIPEDIA

1. Dataset

We plan to take data from the following sources with appropriate pre-processing and augmentation procedures.

- Kaggle Wikipedia dataset
- Wikimedia dumps
- https://osf.io/rce8m/ Wikipedia REST API
- scrape from links to archived hoax articles on Wikipedia's list of hoaxes
- hoaxpedia dataset
- Stanford dataset

2. Stakeholder

- Wikimedia foundation
- Other Internet platforms for factual information
- Fact checkers relying on Wikipedia summaries
- Ordinary people who seek general information online

3. KPI

- Mean hamming distance of classification as bitstrings (equivalent to MAE and MSE in our case);
- More specifically, the distances caused by Type I error and Type I error, respectively;
- Proportion of correct classifications, Type I error, and Type II error.

4. Proposed Modeling Approach

A classification model to determine if a given Wikipedia page is a hoax or not, based on the following features:

- Sentence (and/or word) length
- Editor history
- Number of citations and citation variety
- ...

We will refrain from massive usage of NLP or LLM models, but might slightly apply them or look for previous studies using these methods to evaluate our model.

5. Backup ideas in case we need to pivot

- Keep Wikipedia, change target of classification to predict pageviews?