Detecting Vandals on Wikipedia Based on User Interaction Logging

Kristof Komlossy

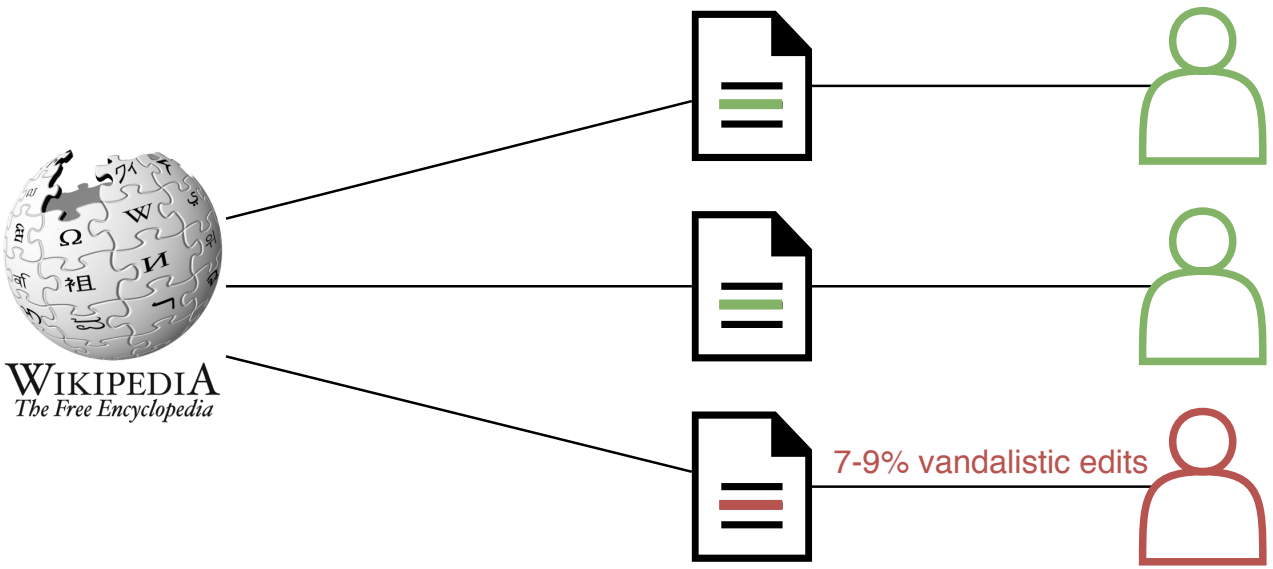
Bauhaus-Universität Weimar

27.08.2018

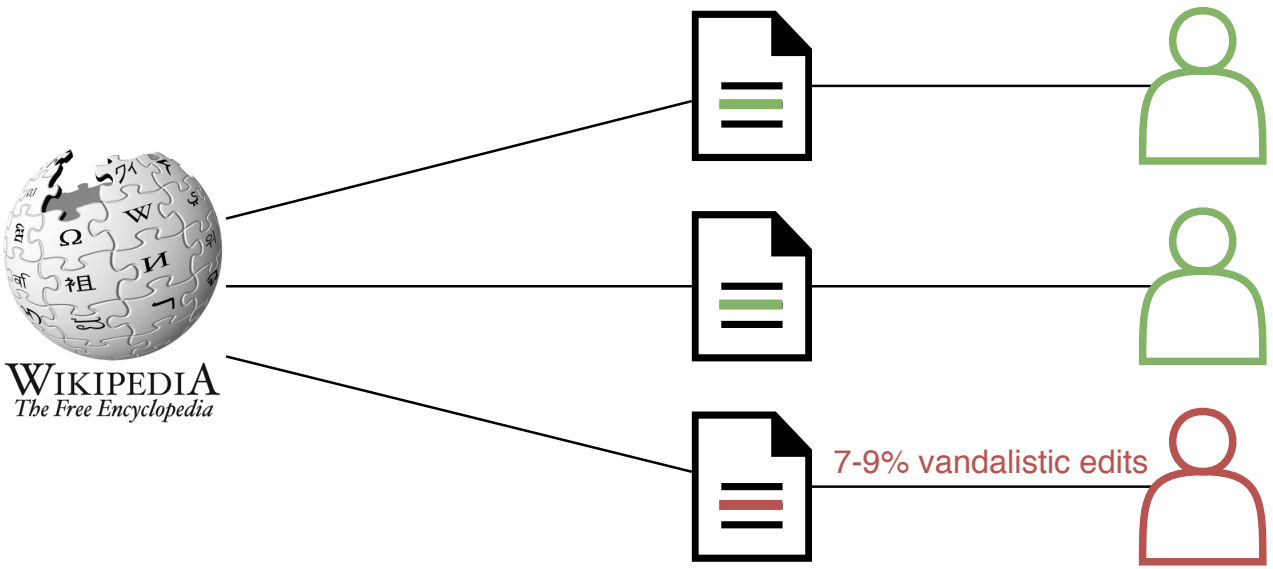
Overview

- Introduction
- Dataset Construction
- Machine Learning Model
- Experiments and Evaluation

Wikipedia Vandalism



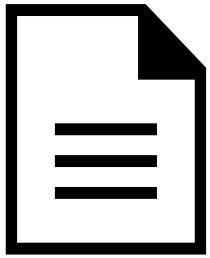
Wikipedia Vandalism



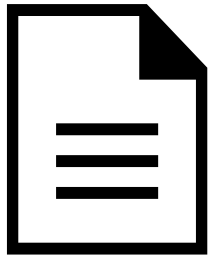
Vandalistic Edits:

- Insert wrong statements: "Weimar has over 1 million citizens"
- Insert gibberish: "Weimar has over 60 thousand hmtjbmgkkjrpstw"
- Delete text: "Weim izens"
- ...

Vandalism Detection

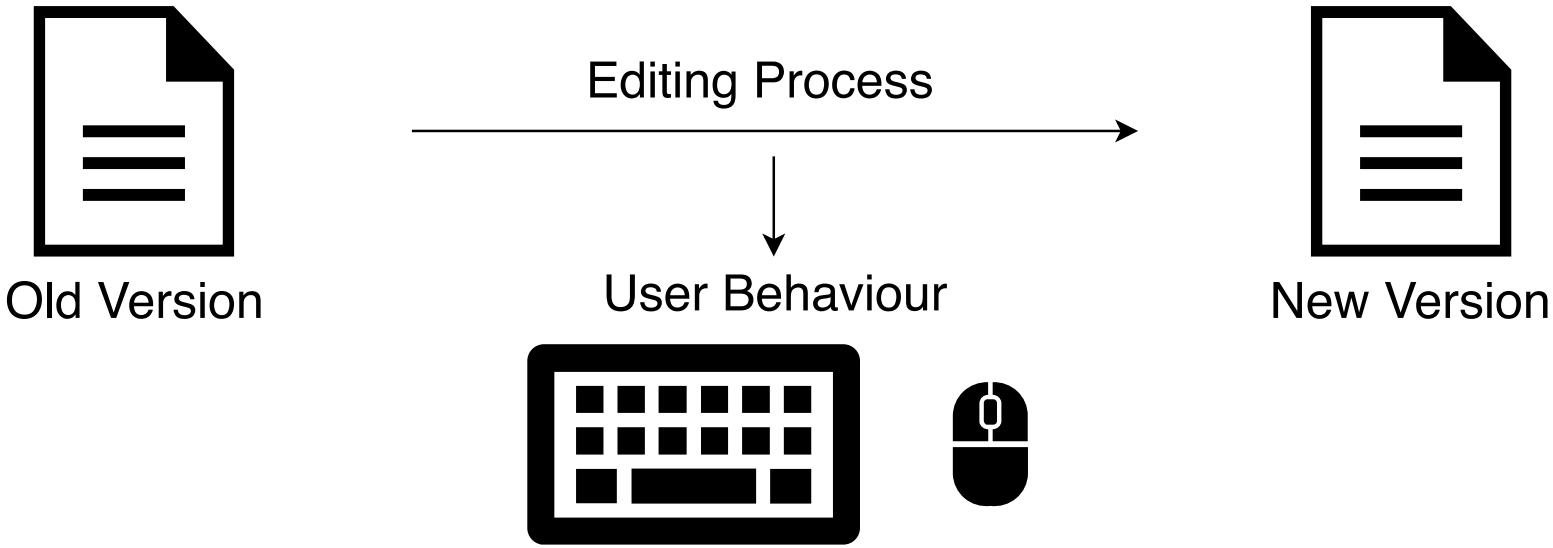


Old Version

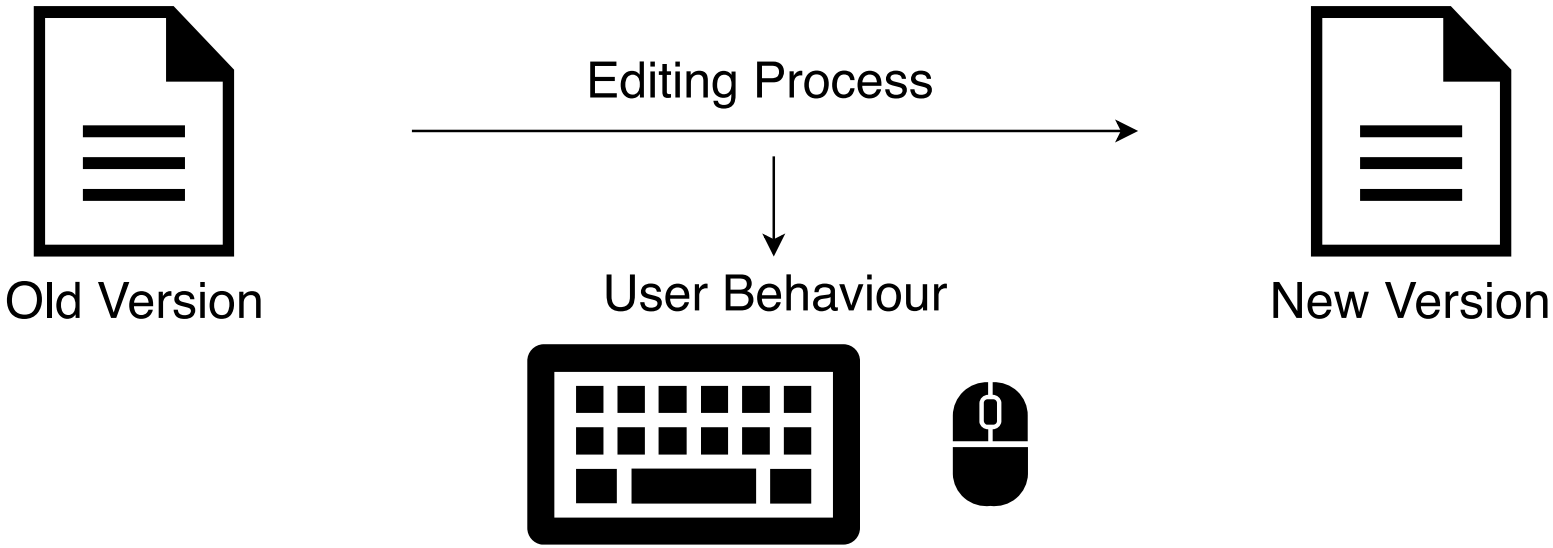


New Version

Vandalism Detection



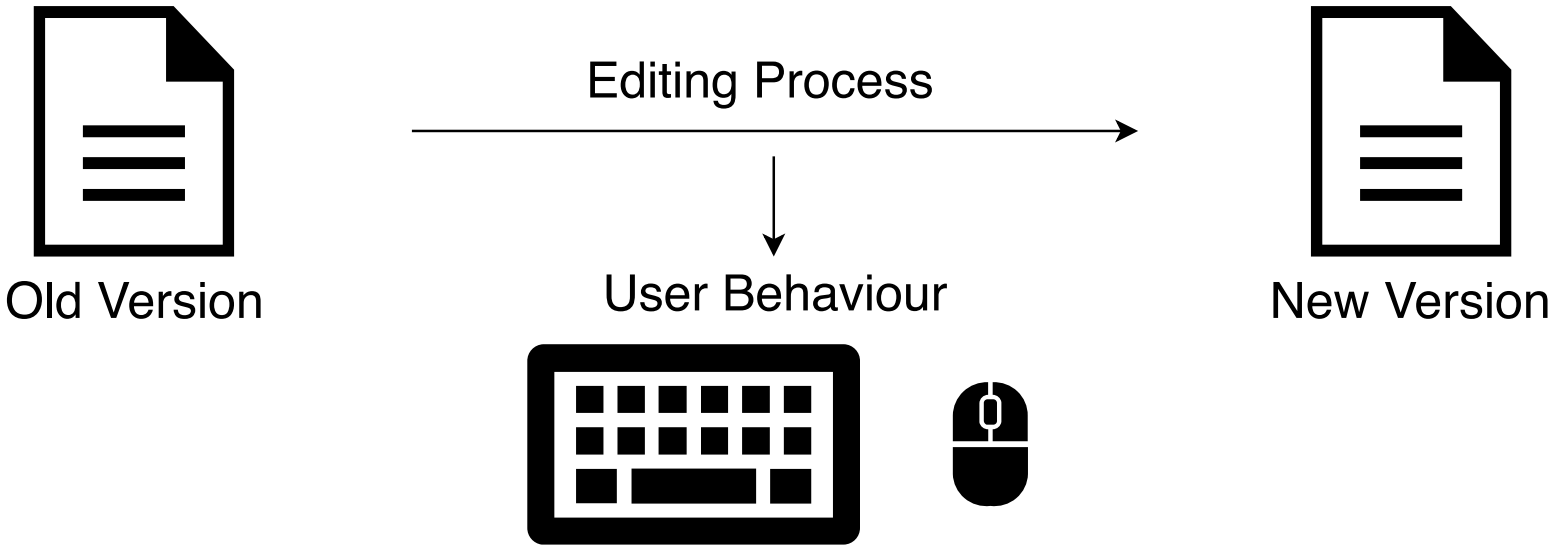
Vandalism Detection



User Interactions:

- Re-authentication (Pusara et al. 2004)
- Identify current mood (Khan et al. 2008)
- Authorship attribution (Plank et al. 2016)

Vandalism Detection



User Interactions:

- Re-authentication (Pusara et al. 2004)
- Identify current mood (Khan et al. 2008)
- Authorship attribution (Plank et al. 2016)
- **Vandalism detection?**
  - Early detection of vandalism
  - Combination with existing techniques

Contributions

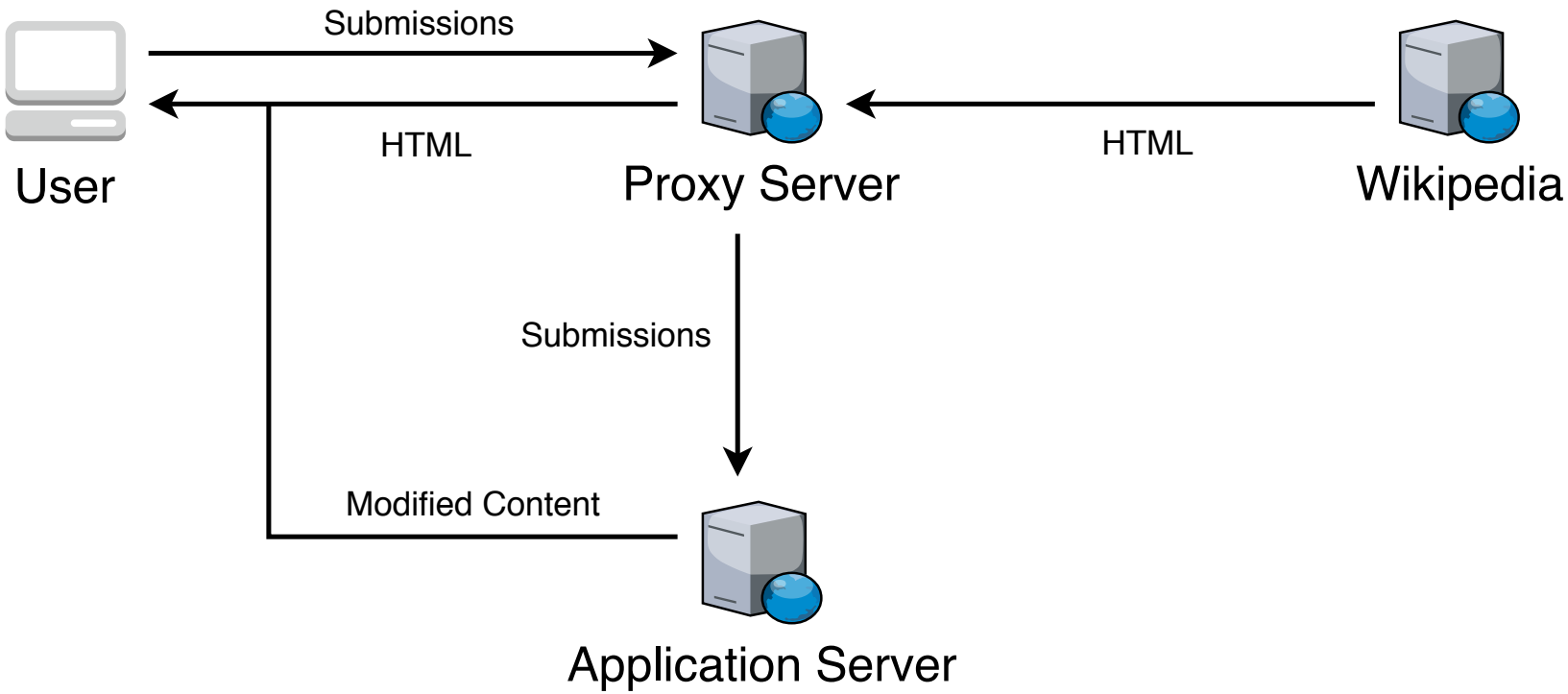
- Creation of system to edit Wikipedia in a non-invasive way
- Development of tool to track and record user interactions
- Design of crowdsourcing tasks to collect vandalism and non-vandalism edits
- Creation of dataset containing more than 3,800 user interactions editing Wikipedia
- First-time evaluation of vandalism detection based on user behaviour

Dataset Construction

- Appropriate dataset does not exists
- Collection of user interactions from real Wikipedia not possible
- Dataset requirements:
  - Authentic
  - Large

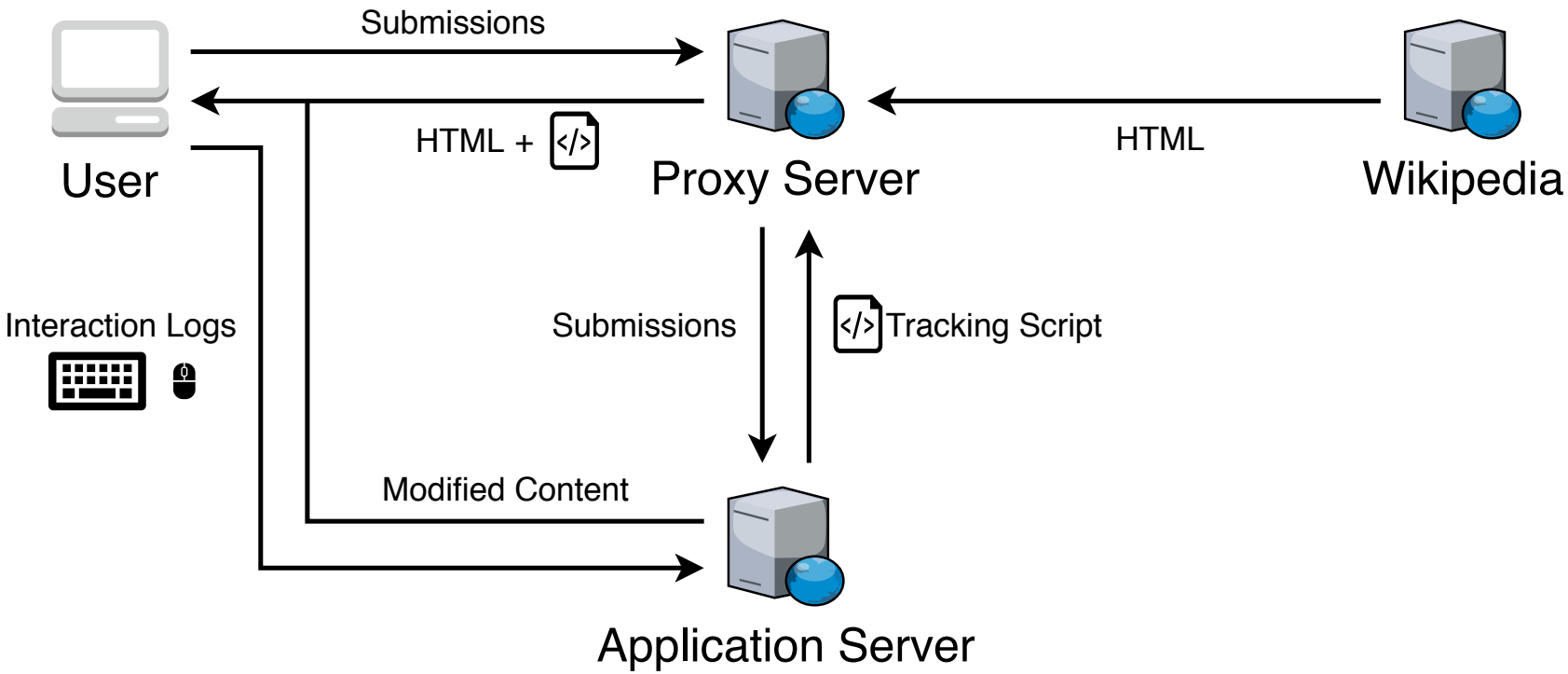
Requirement: Authentic

Server Setup



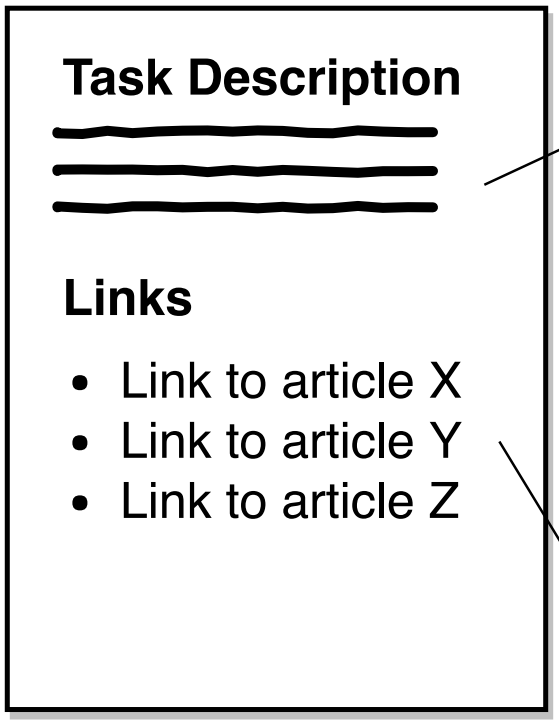
Requirement: Authentic

Server Setup



Requirement: Large

Crowdsourcing



Different Tasks:

1. Perform Vandalism

   e.g. destroy article, insert gibberish, ...

2. Correct Spelling Errors

   e.g. "... print the valeu of ..."

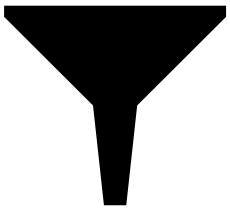
3. Insert New Fact

   e.g. "Elizabeth Weiffenbach was an art teacher at Lafayette High School in Buffalo, New York, from the school's opening in 1903 until her retirement in 1952."

60 different articles
30 workers per article

Dataset

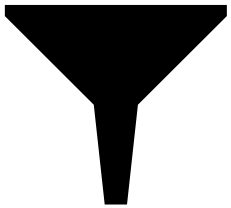
5,400 edits



Rejections: Low-Quality Work:

- No processing of article
- Copied and pasted the new fact
- ...

4,015 edits

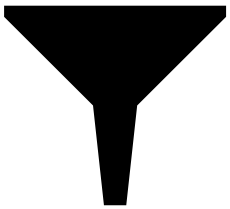


Post-processing: Technical Issues (corrupt log files)

3,808 edits

Dataset

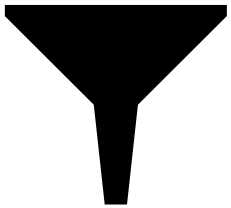
5,400 edits



Rejections: Low-Quality Work:

- No processing of article
- Copied and pasted the new fact
- ...

4,015 edits



Post-processing: Technical Issues (corrupt log files)

3,808 edits

Statistics:

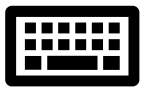
- Number of edits: 3,808
- Number of workers: 335
- Total costs: approx. 500$

- Total vandalism: 53%
- Simple vandalism: 50%
- Complicated vandalism: 45%
- Mixed vandalism: 5%

Model

My Model

Total features: 3,736

	Features
General	Edit Duration ...
	#Left Clicks Cursor Speed Cursor Curvature ...
	#Pressed Letters #Pressed Numbers Dwell Time Flight Time ...

Model

My Model

Total features: 3,736

Features
General: Edit Duration ...
#Left Clicks Cursor Speed Cursor Curvature ...
#Pressed Letters #Pressed Numbers Dwell Time Flight Time ...

Wikimedia Model

Total features: 904

Features
Cursor Speed Cursor Curvature ...
Dwell Time Flight Time ...

Experiments

- 20% test set, 80% training set
- Experiments are executed a hundred times and averaged
- Usage of random forest algorithm
- Evaluation with F1 measure ($2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$)

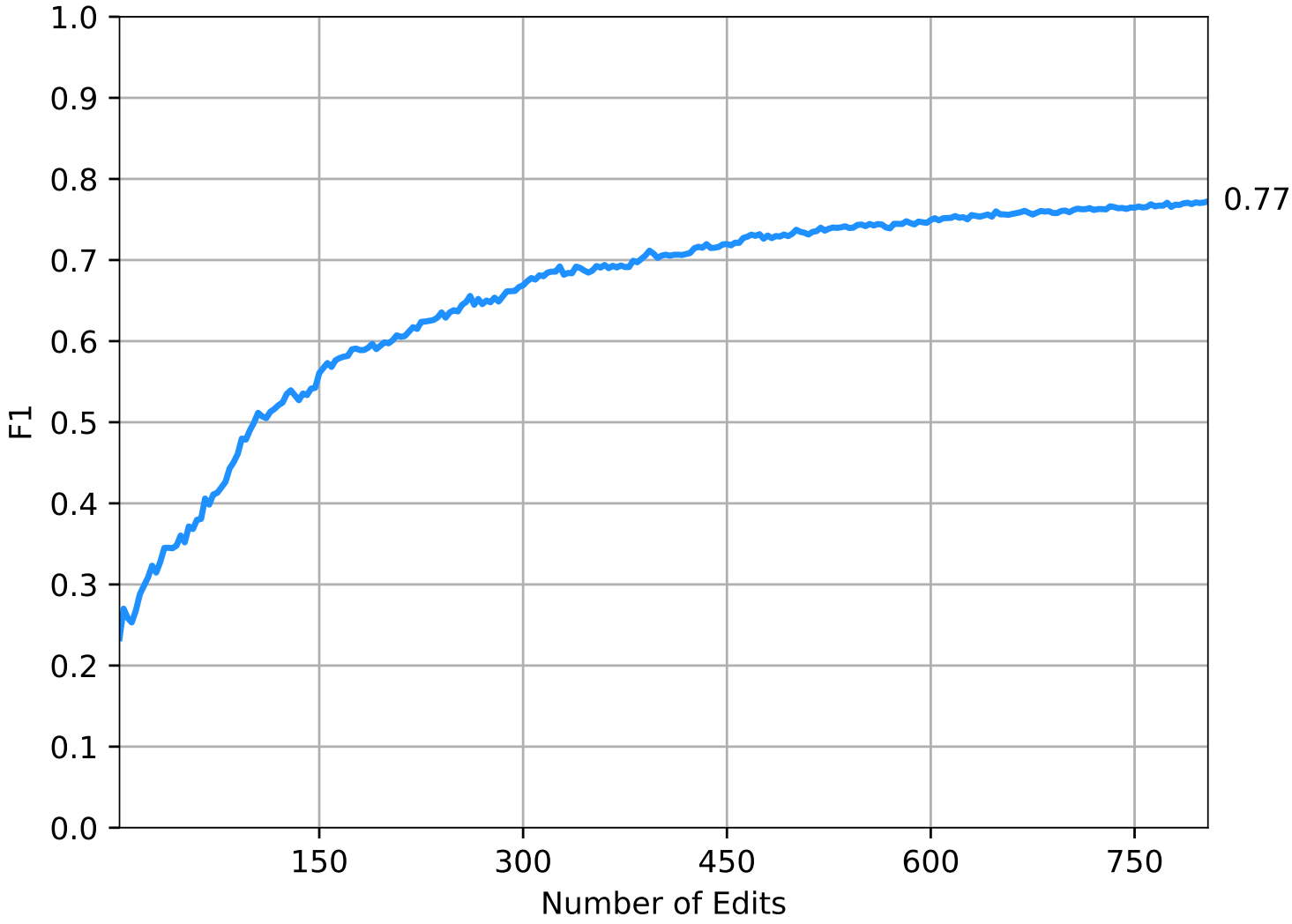
Experiments

- 20% test set, 80% training set
- Experiments are executed a hundred times and averaged
- Usage of random forest algorithm
- Evaluation with F1 measure $(2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}})$

- Research Questions:
  1. Is it possible to detect vandalism using only the behavior?
  2. Can the user's privacy be protected?
  3. How early can vandalism be detected?

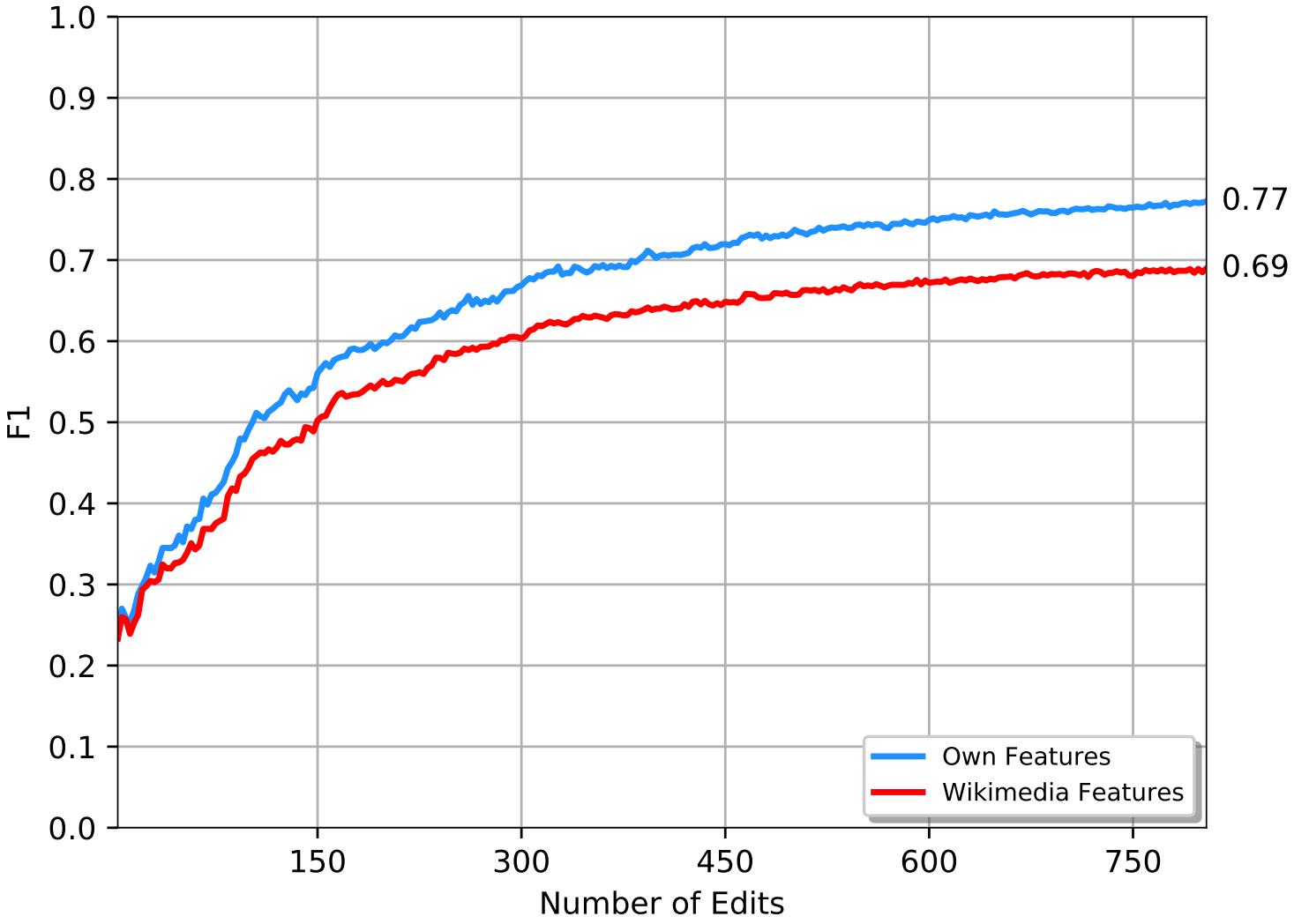
Vandalism Detection Possible?

Experiment 1: Increasing Training Set Over Workers



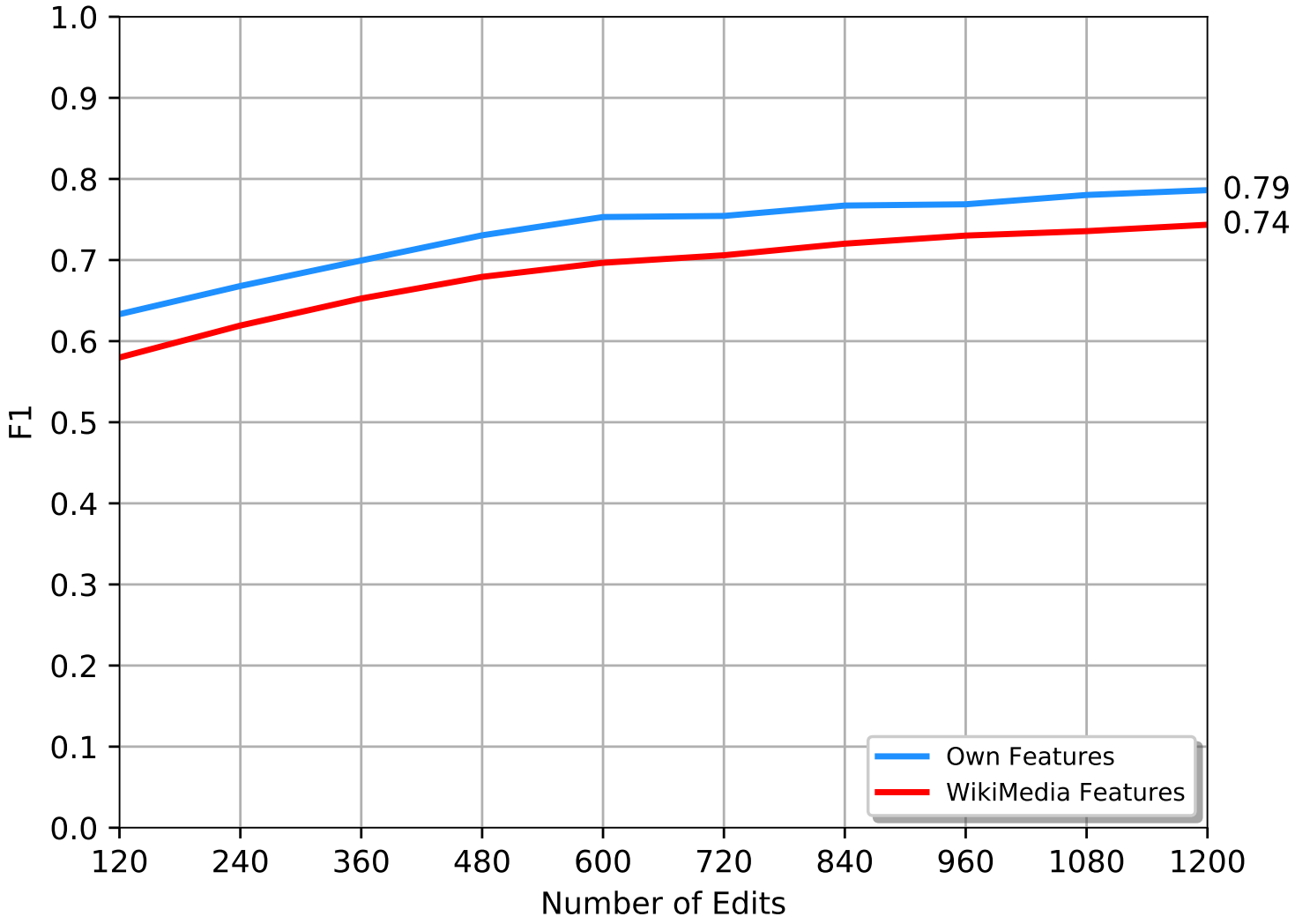
Vandalism Detection Possible?

Experiment 1: Increasing Training Set Over Workers

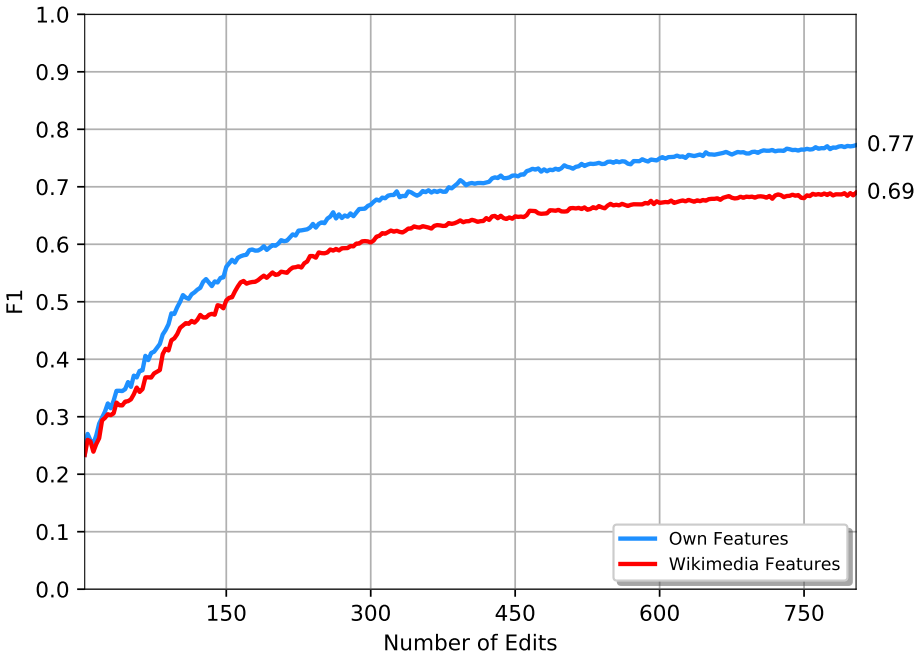


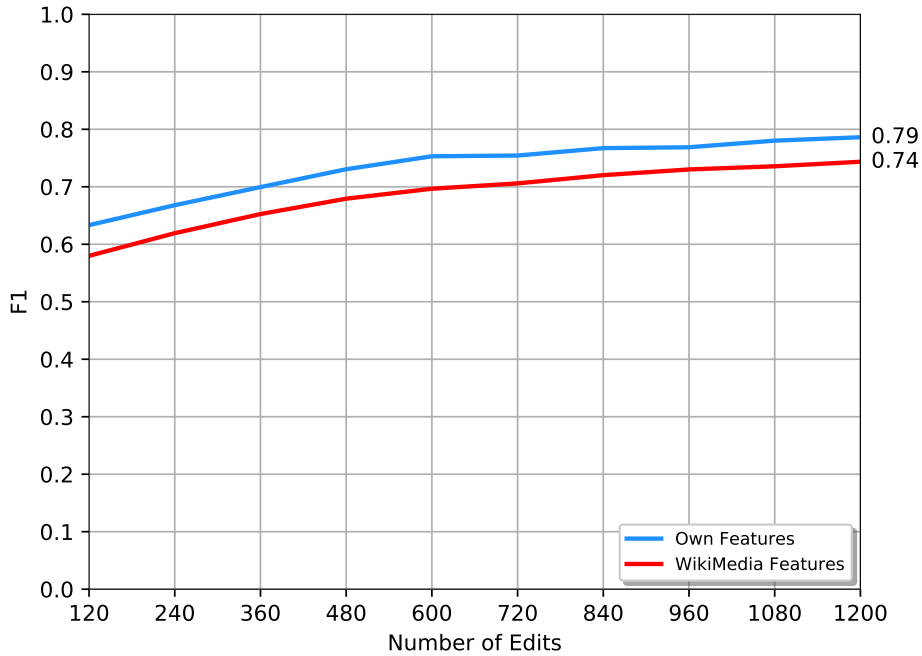
Vandalism Detection Possible?

Experiment 2: Increasing Training Set Over Edits Per Worker



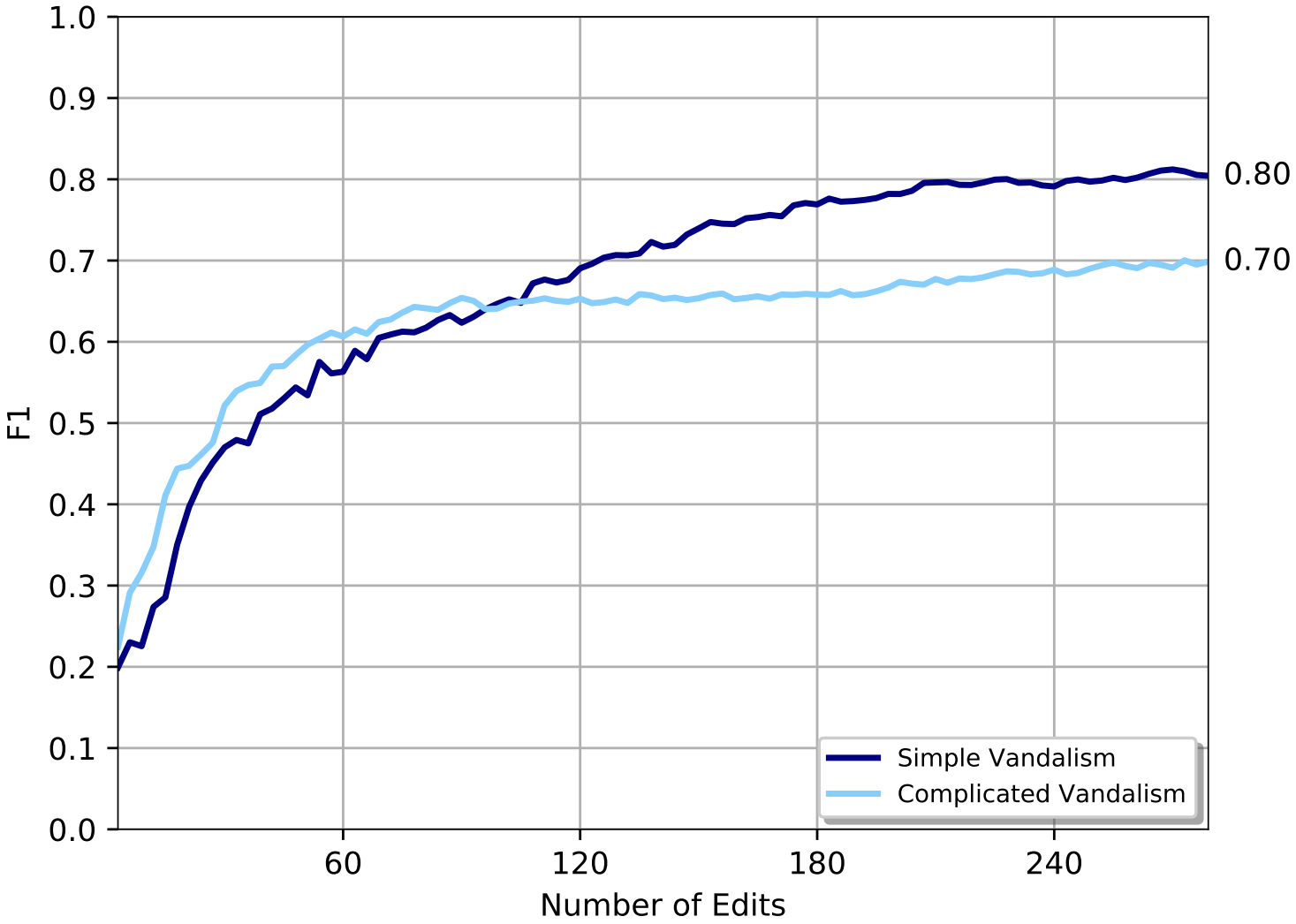
Vandalism Detection Possible?





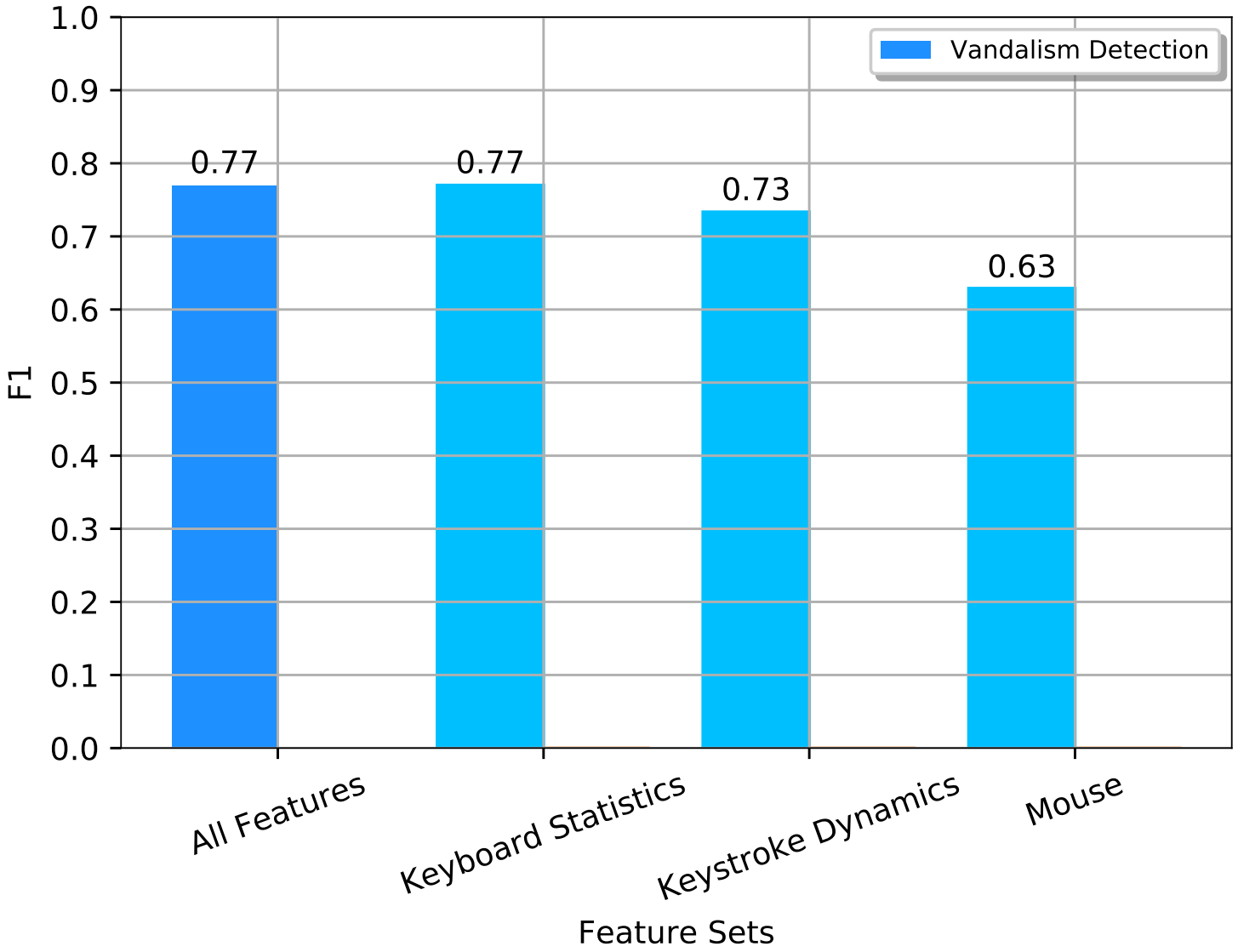
Vandalism Detection Possible?

Experiment 3: Simple vs. Complicated Vandalism



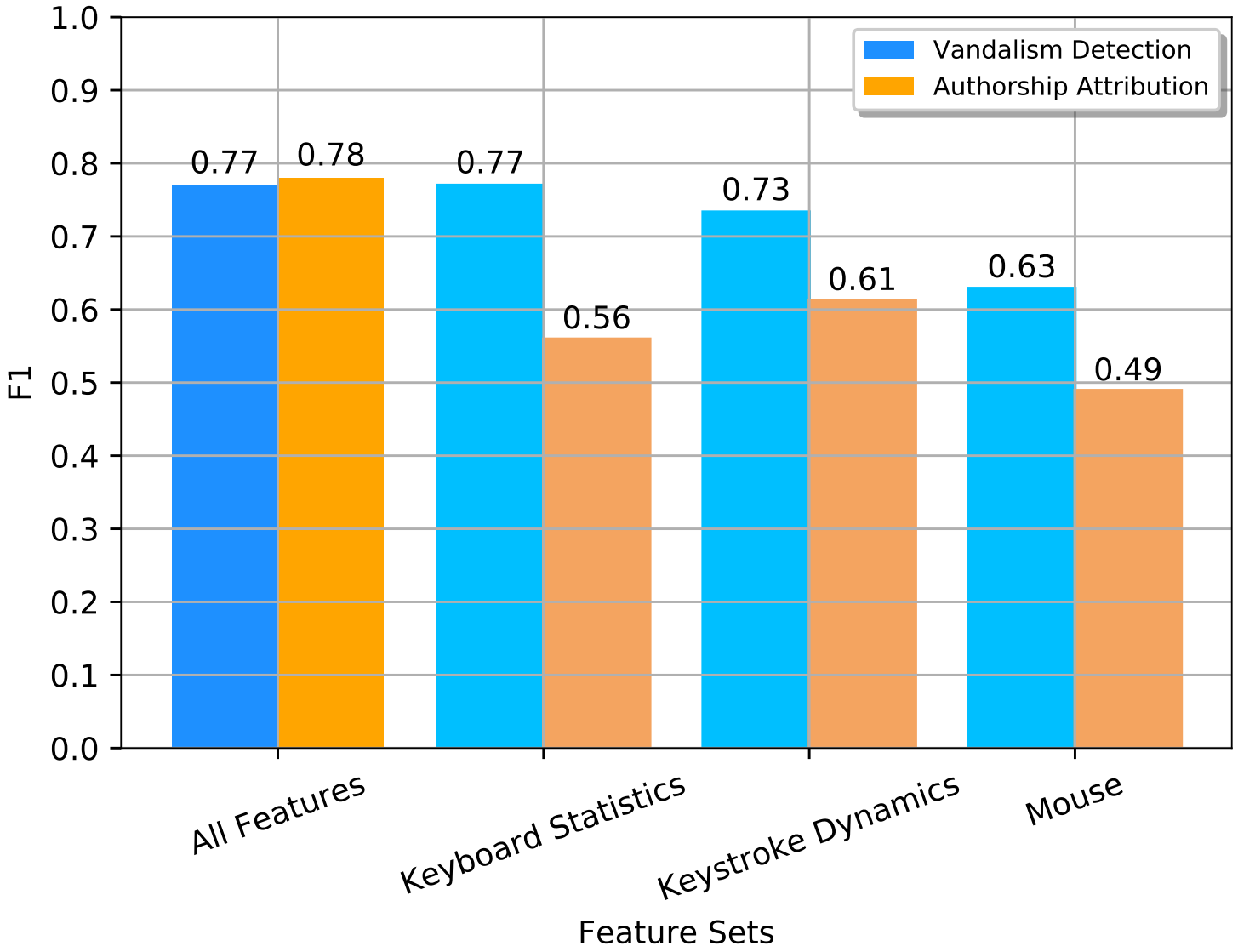
Can the user's privacy be protected?

Experiment 4: Feature Selection



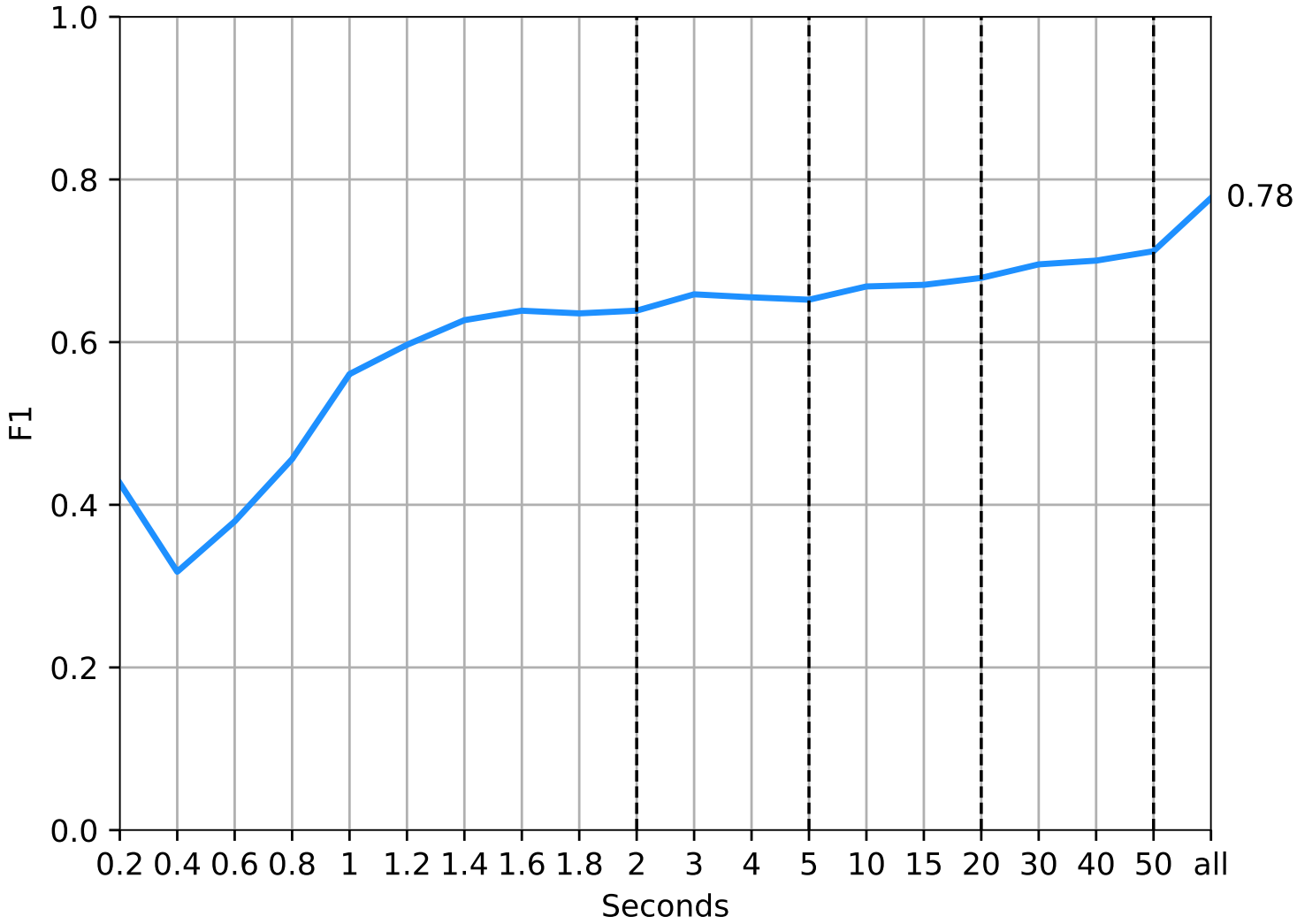
Can the user's privacy be protected?

Experiment 4: Feature Selection



How early can vandalism be detected?

Experiment 5: Vandalism Detection Over Time



Conclusion

- Behavioral vandalism detection is possible
  - F1 score of 0.77
- Data privacy can be protected
  - Reducing amount of recorded data
  - Detecting vandalism without being able to identify users
- Fast detection is possible
  - First tendency after approx 1.4 seconds
  - 15 seconds for complicated vandalism

Conclusion

- Behavioral vandalism detection is possible
  - F1 score of 0.77
- Data privacy can be protected
  - Reducing amount of recorded data
  - Detecting vandalism without being able to identify users
- Fast detection is possible
  - First tendency after approx 1.4 seconds
  - 15 seconds for complicated vandalism

Future Work

- Larger dataset
- Recorded behavior from the real Wikipedia
- Combine behavioral approach with existing techniques
- Wikipedia-dependent features