Data Science, 2024

Mining massive datasets: Wikipedia Bots' Detector

APPLIED SCIENCES FACULTY

Denys Botuk

Yurii Roziaiev

Tetiana Dulina

Project Objectives

- Data:
  - Obtain 40k Wikipedia edits
  - Use 20% sampling for the stream
  - Show distributions of edits per humans and bots
- Classifier:
  - Train a model to classify user as a human or bot
  - Do not use username field
  - Evaluate the model
- Bloom Filter:
  - Train on model’s predictions
  - Find optimal parameters to around 10% error rate
- Make the system work with PySpark Streaming

Data Fetching & Sampling

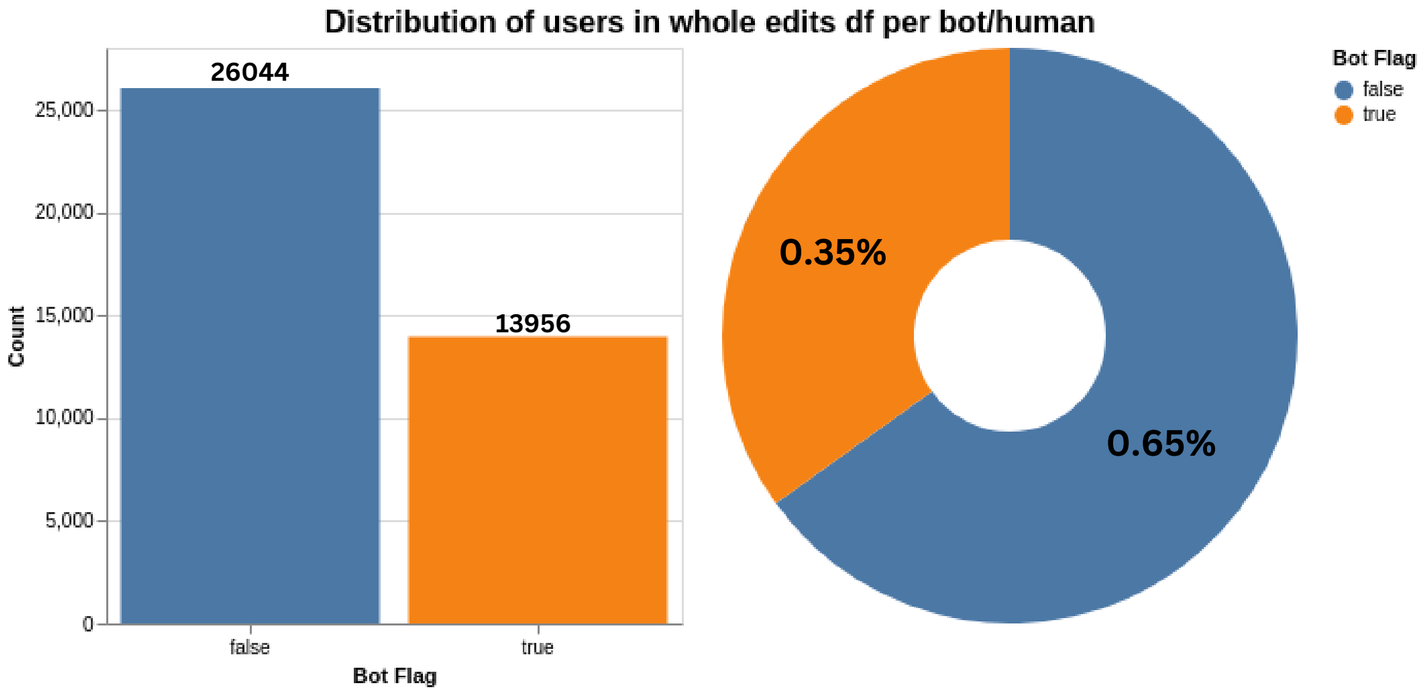
- Edits endpoint
  - https://stream.wikimedia.org/v2/stream/recentchange
- Sampling 20% of the stream
- 40000 edits


```
event: message
id: [{"topic":"eqiad.mediawiki.recentchange","partition":0,"offset":-1},
{"topic":"codfw.mediawiki.recentchange","partition":0,"timestamp":1732125085001}]
data: {"$schema":"/mediawiki/recentchange/1.0.0","meta":
{"uri":"https://www.wikidata.org/wiki/Q6988441","request_id":"946c0e16-6d94-4001-a166-
7e8d88a404f9","id":"e2eb68df-d685-466f-8366-7a495d4234fe","dt":"2024-11-
20T17:51:25Z","domain":"www.wikidata.org","stream":"mediawiki.recentchange","topic":"codfw.mediawiki.rec
entchange","partition":0,"offset":1262420065},"id":2345703731,"type":"edit","namespace":0,"title":"Q6988
441","title_url":"https://www.wikidata.org/wiki/Q6988441","comment":"/* wbeditentity-update-languages-
short:0||mul */ Update default label to name in native
language","timestamp":1732125085,"user":"Iamcarbon","bot":false,"notify_url":"https://www.wikidata.org/w
/index.php?diff=2277570159&oldid=1716872425&rcid=2345703731","minor":false,"patrolled":true,"length":
{"old":6986,"new":7058},"revision":
{"old":1716872425,"new":2277570159},"server_url":"https://www.wikidata.org","server_name":"www.wikidata.
org","server_script_path":"/w","wiki":"wikidatawiki","parsedcomment":"<span dir=\"auto\"><span
class=\"autocomment\">Changed label, description and/or aliases in mul: </span></span> Update default
label to name in native language"}
```

Raw Data Distribution

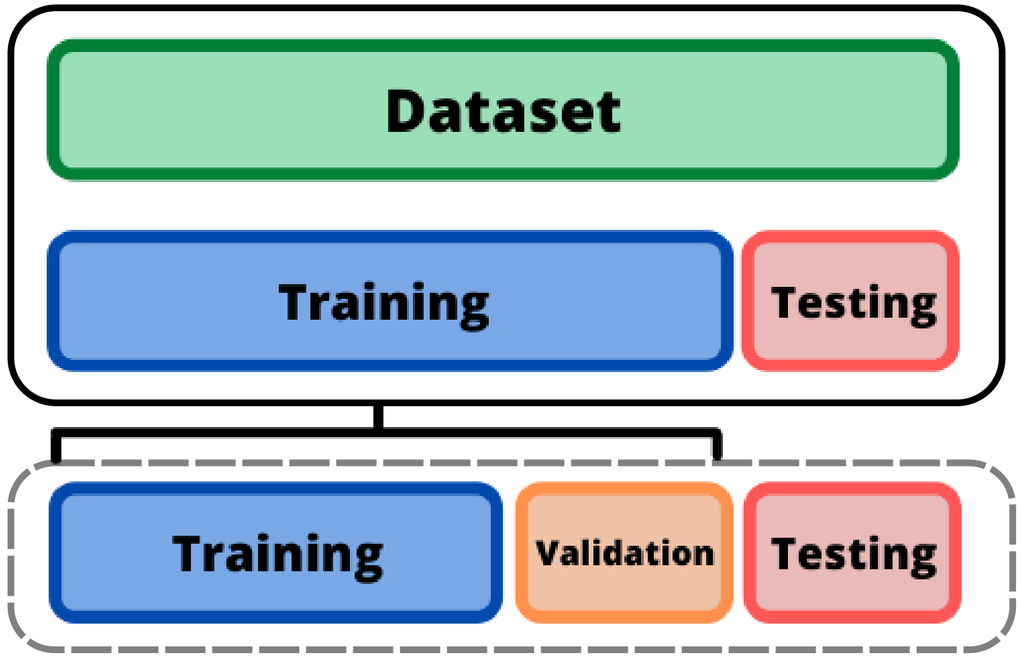
- Unique humans: 4586
- Unique bots: 159
- Hybrid: **41**

user	bot
Russian Rocky	0
Russian Rocky	1

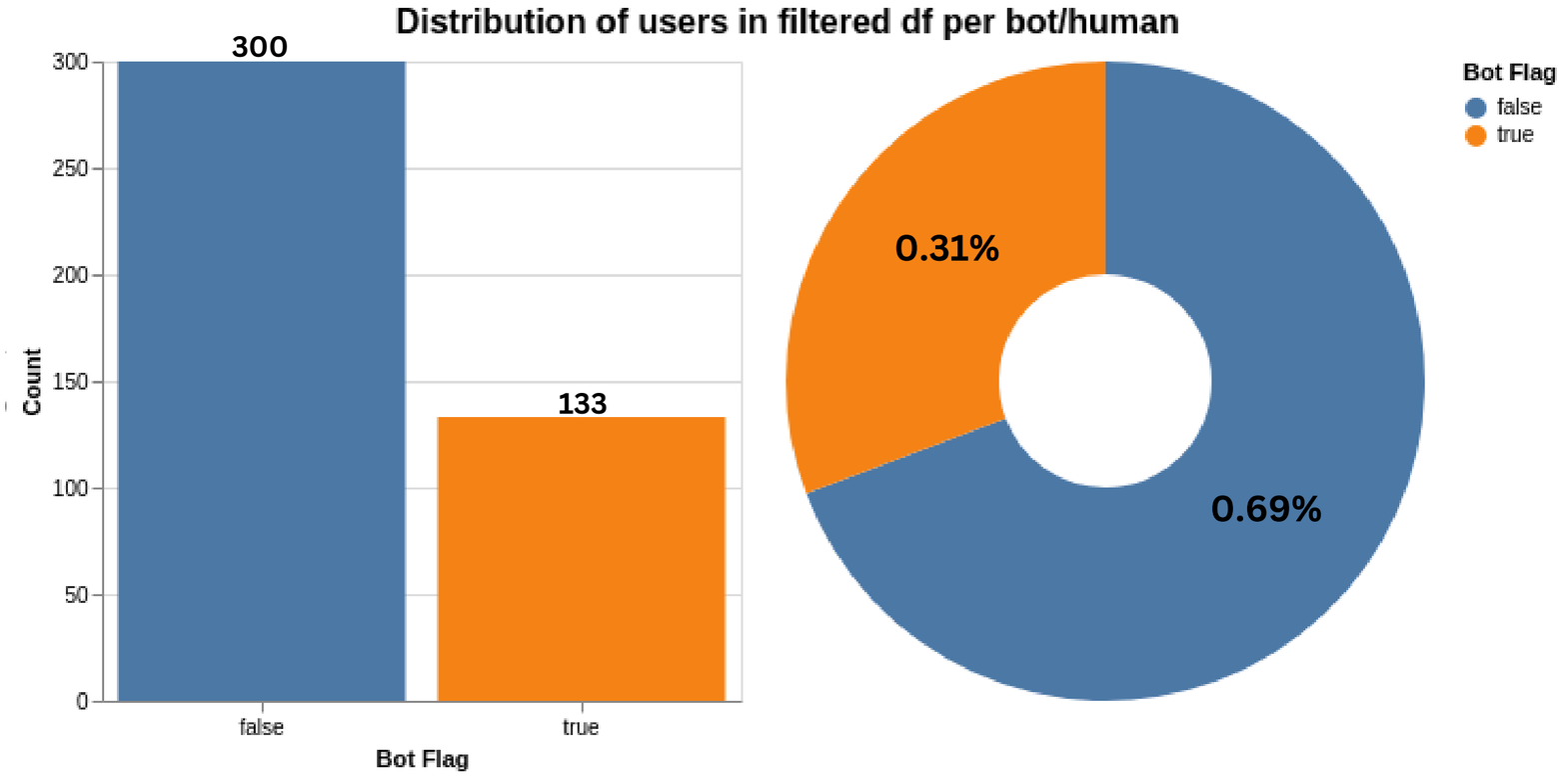


Data Preparation

- **Random split**
- **Grouped split**
  - 200/215 test edits by 1 bot
- **Deduplication**:
  - 26044 -> 4560 human
  - 13956 -> 133 bot
- **Class balancing**:
  - 4560 -> 300 human



Filtered Data Distribution



Classification Model

Random Forest Classifier:

- 4 features (type, namespace, comment, length)
- custom features:
  - length_diff
  - comment_length
  - stopwords filter + tf-idf on comment
  - one-hot encoding on type and namespace
- 1 target feature (bot)

```
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   $schema             40000 non-null  object
 1   meta                40000 non-null  object
 2   id                  39248 non-null  float64
 3   type                40000 non-null  object
 4   namespace           40000 non-null  int64
 5   title               40000 non-null  object
 6   title_url           40000 non-null  object
 7   comment             40000 non-null  object
 8   timestamp           40000 non-null  datetime64[ns]
 9   user                39999 non-null  object
 10  bot                 40000 non-null  int64
 11  notify_url          38143 non-null  object
 12  server_url          40000 non-null  object
 13  server_name         40000 non-null  object
 14  server_script_path  40000 non-null  object
 15  wiki                40000 non-null  object
 16  parsedcomment       40000 non-null  object
 17  minor               22990 non-null  float64
 18  patrolled           16641 non-null  float64
 19  length              22990 non-null  object
 20  revision            22990 non-null  object
 21  log_id              1857 non-null   float64
 22  log_type            1857 non-null   object
 23  log_action          1857 non-null   object
 24  log_params          1857 non-null   object
 25  log_action_comment  1857 non-null   object
```

Hyperparameters Tuning

Hyperparameters: numFeatures, numTrees, maxDepth

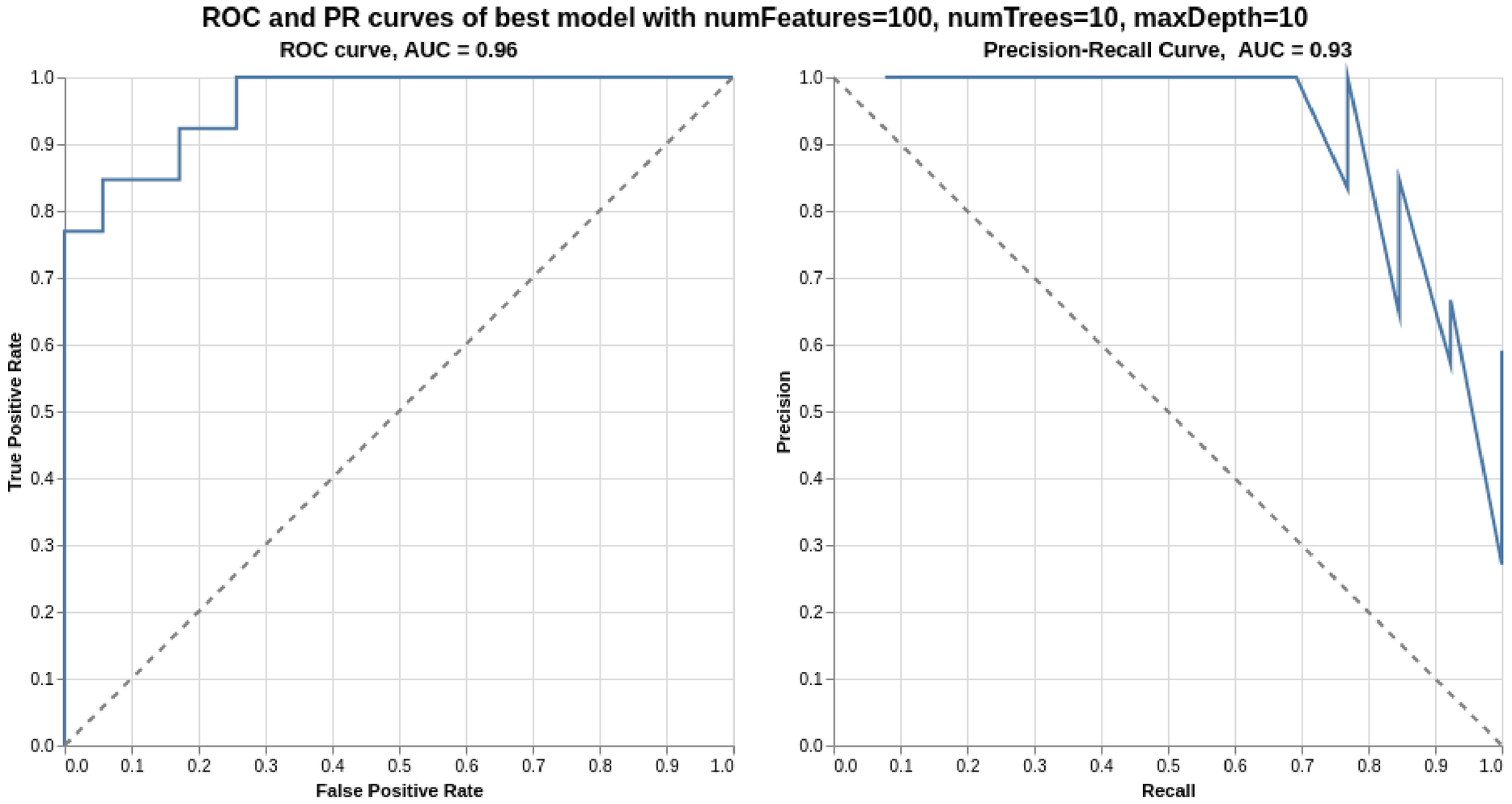
Grid Search:

- numFeatures: [100, 500, 1000]
- numTrees: [10, 50, 100]
- maxDepth: [5, 10, 20]

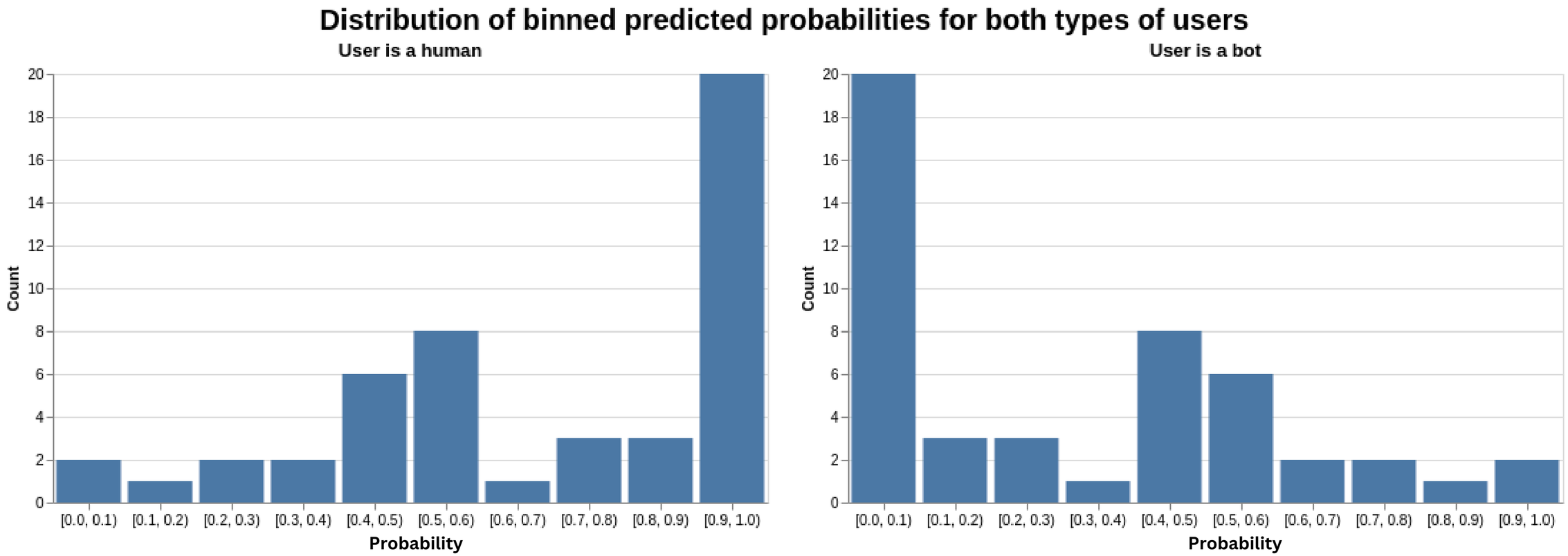
Evaluation: f1, accuracy, precision, recall, ROC AUC, PR AUC

Optimal parameters: numFeatures=100, numTrees=10, maxDepth=10

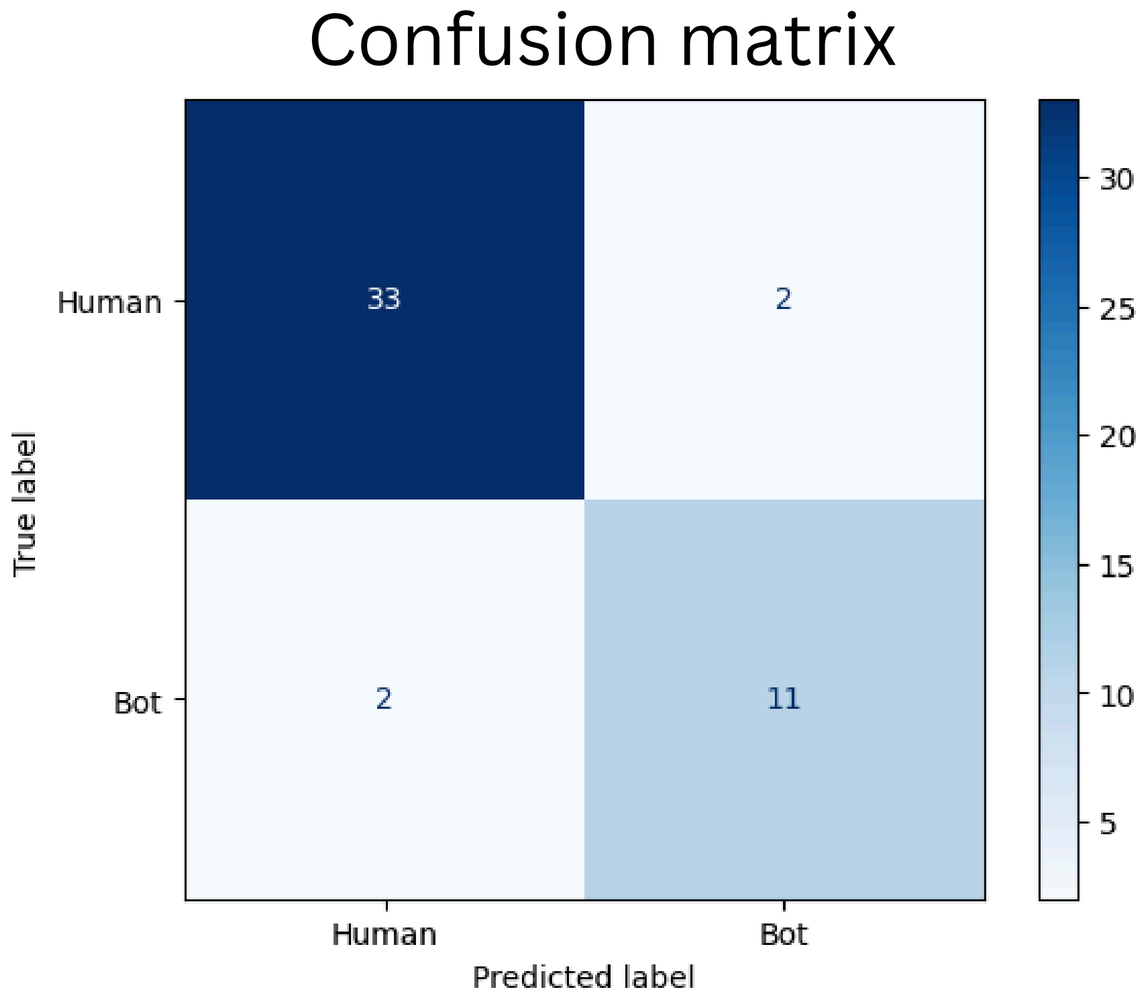
Model Evaluation



Model Evaluation



Model Evaluation



Bloom Filter

Input

Blacklist of bots (usernames that classifier predicted as bots)

Hash function

mmh3.hash(x, seed=42) % (a + b) % filtersize

where
a and **b** are random integers between 1 and filtersize - 1
filtersize - number of bits in the filter
x - bot’s username

Optimal parameters

Optimal filter size: $$m = -\frac{n \ln(p)}{(\ln(2))^2} = -\frac{13 \times \ln(0.1)}{(\ln(2))^2} \approx 62$$

Optimal number of hash functions: $$k = ln(2) \times \frac{m}{n} = ln(2) \times \frac{62}{13} \approx 3$$

p = 0.1 - desired FPR (error rate)

n = 13 - number of stored elements

Deployment & Demo

- Docker image
- PySpark Streaming
- Socket input
- Live blacklisting
- Output to console

```
-------------------------------------------
+--------------------+--------------------+--------+----+-----------+
|               title|           title_url|    user| bot|blacklisted|
+--------------------+--------------------+--------+----+-----------+
|Категорія:Сторінк...|https://uk.wikipe...|BunykBot|true|       true|
+--------------------+--------------------+--------+----+-----------+

100 1090k    0 1090k    0     0  52111      0 --:--:--  0:00:21 --:--:--
Batch: 19
-------------------------------------------
100 1127k    0 1127k    0     0  51794      0 --:--:--  0:00:22 --:--:--
|               title|           title_url|     user|  bot|blacklisted|
+--------------------+--------------------+---------+-----+-----------+
|Template:Casertan...|https://en.wikipe...|     Rodw|false|       true|
|   Dimitrana Ivanova|https://fr.wikipe...|RobokoBot| true|       true|
+--------------------+--------------------+---------+-----+-----------+
```

```
+--------------------+--------------------+------------------+-----+---------+
|               title|           title_url|              user|  bot|blacklist|
+--------------------+--------------------+------------------+-----+---------+
|           Q51561119|https://www.wikid...|             KrBot| true|    false|
|Transformers : Le...|https://fr.wikipe...|           Pautard|false|    false|
|User talk:Changem...|https://en.wikipe...|         Wiiformii|false|    false|
|Category:WikiProj...|https://en.wikipe...|        Psquintero|false|    false|
|Category:Draft-Cl...|https://en.wikipe...|        Psquintero|false|    false|
|Petit dictionnair...|https://fr.wikiso...|        Poslovitch|false|    false|
|Category:Media co...|https://commons.w...|          DPLA bot| true|    false|
|Категорија:Шаблон...|https://sr.wikipe...|     Ранко Николић|false|    false|
|Категорија:Шаблон...|https://sr.wikipe...|     Ранко Николић|false|    false|
|Британские СМИ: а...|https://ru.wikine...|InternetArchiveBot| true|    false|
|Category:Digital ...|https://commons.w...|          DPLA bot| true|    false|
|Category:Draft-Cl...|https://en.wikipe...|        Psquintero|false|    false|
|Категория:Страниц...|https://ru.wikipe...|            Valmin|false|    false|
|Category:NA-impor...|https://en.wikipe...|        Psquintero|false|    false|
|File:2017-06-21 L...|https://commons.w...|     SchlurcherBot| true|    false|
|Category:NA-impor...|https://en.wikipe...|        Psquintero|false|    false|
|      Sabine Monauni|https://de.wikipe...|               Aka|false|     true|
|Category:Media co...|https://commons.w...|          DPLA bot| true|    false|
|Category:Digital ...|https://commons.w...|          DPLA bot| true|    false|
|Категорија:Шаблон...|https://sr.wikipe...|     Ранко Николић|false|    false|
+--------------------+--------------------+------------------+-----+---------+
only showing top 20 rows
```

Q&A