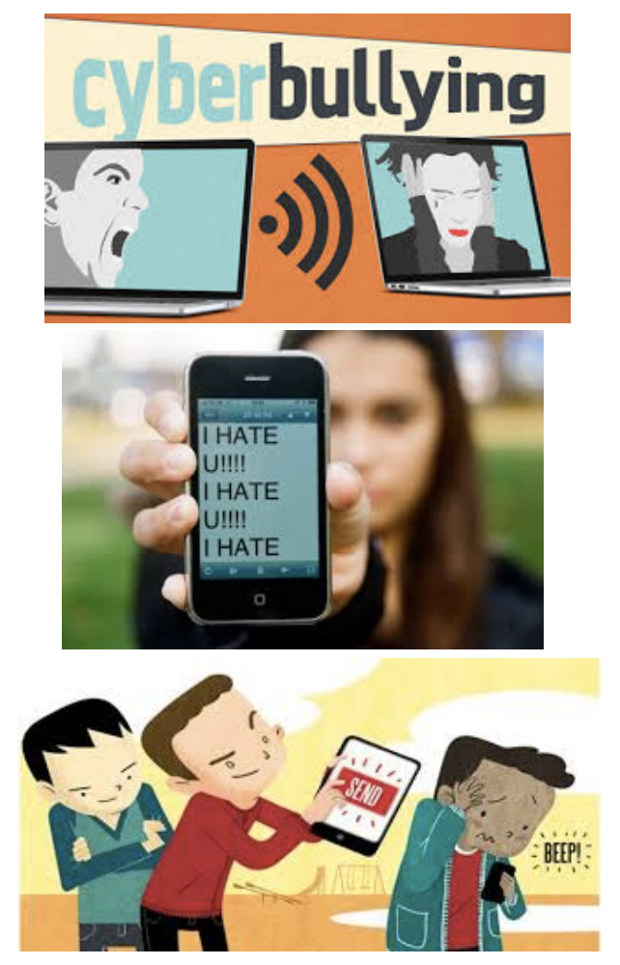
http://goo.gl/KzUNR8 http://goo.gl/6Nrc1r

OPTIMAL ONLINE CYBERBULLYING DETECTION

Daphney–Stavroula Zois*, Angeliki Kapodistria*, Mengfan Yao# and Charalampos Chelmis#
*Electrical and Computer Engineering Department, #Computer Science Department
University at Albany, SUNY, Albany, NY 12222, USA
{dzois, akapodistria, myao, cchelmis}@albany.edu

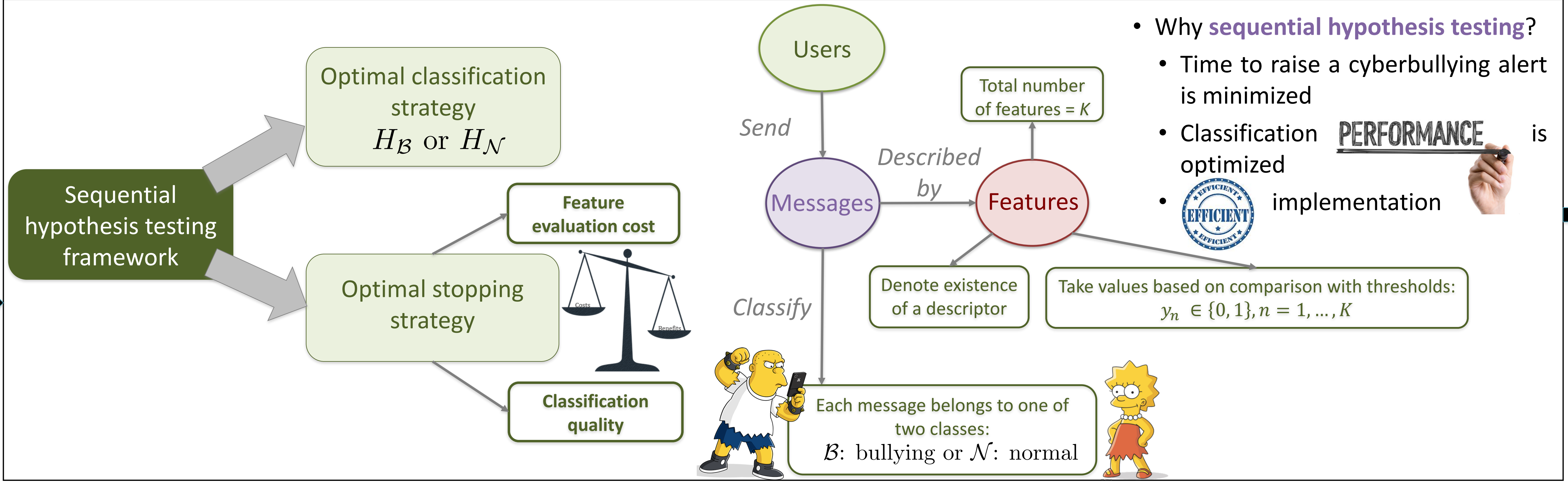
Motivation

- Bullying can occur **anytime** and **anywhere**
- Consequences are **devastating**: learning difficulties, psychological suffering, suicide
- Two key practical issues in cyberbullying detection thus far remain unaddressed:
  - Scalability
  - Timeliness



Accurately detect cyberbullying messages using **text–based features** in a **scalable** and **timely manner**!

Framework

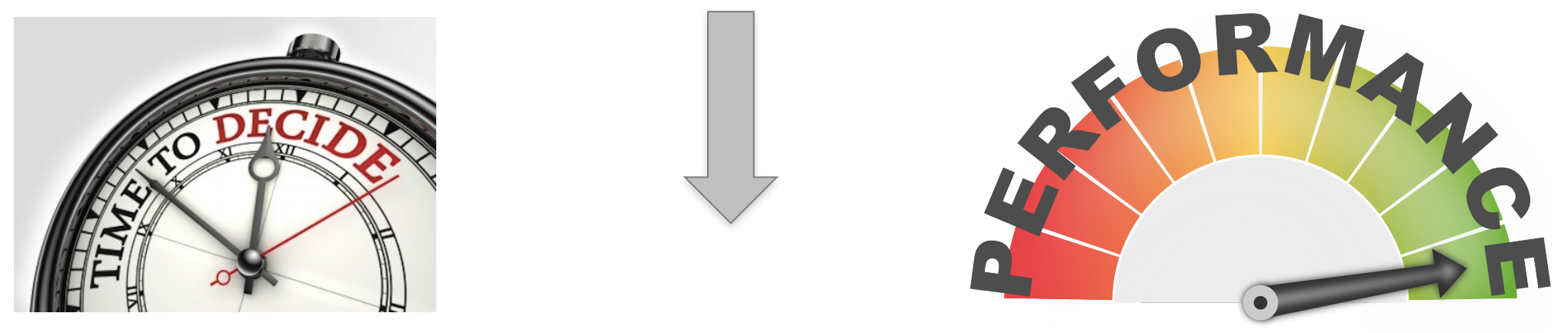


Sequential hypothesis testing framework
Optimal classification strategy $H_{\mathcal{B}}$ or $H_{\mathcal{N}}$
Optimal stopping strategy
Feature evaluation cost
Classification quality
Users
Send
Messages
Described by
Features
Total number of features = K
Classify
Denote existence of a descriptor
Take values based on comparison with thresholds: $y_n \in \{0,1\}, n = 1,\dots,K$
Each message belongs to one of two classes: $\mathcal{B}$: bullying or $\mathcal{N}$: normal

- Why **sequential hypothesis testing**?
  - Time to raise a cyberbullying alert is minimized
  - Classification PERFORMANCE is optimized
  - EFFICIENT implementation

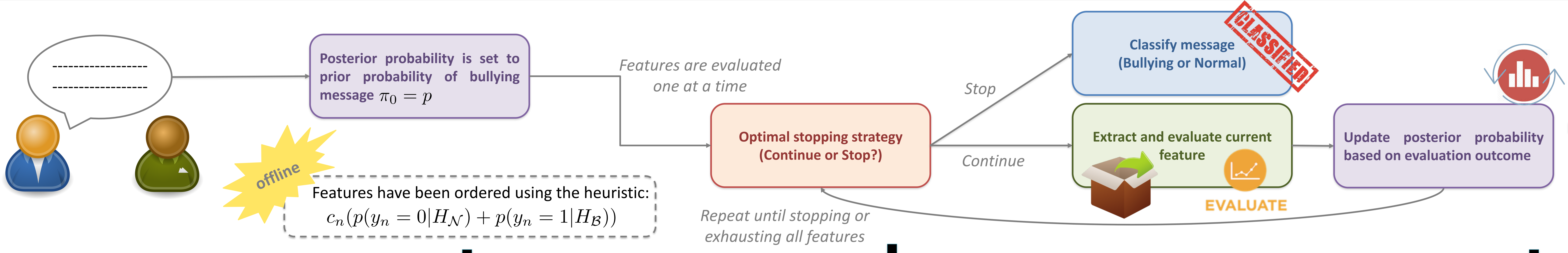
Related Work

- **Prior work**:
  - Focuses **only** on **classification performance**
  - Decision is made using **all features**
- In contrast, **our framework**:
  - Focuses on both **classification performance** and **timeliness**
  - Decision is made using **optimal subset of features**



- **Reduced time** to reach a decision **without sacrificing classification performance**

AvOID: A Novel Algorithm for Optimal Online CyberbullyIng Detection



Posterior probability is set to prior probability of bullying message $\pi_0 = p$
Features are evaluated one at a time
Optimal stopping strategy (Continue or Stop?)
Stop
Continue
Classify message (Bullying or Normal)
CLASSIFIED
Extract and evaluate current feature
EVALUATE
Update posterior probability based on evaluation outcome
Repeat until stopping or exhausting all features
offline

Features have been ordered using the heuristic:

$$c_n(p(y_n = 0|H_{\mathcal{N}}) + p(y_n = 1|H_{\mathcal{B}}))$$

Classification Strategy

- Optimal classification strategy:

$$\mathcal{D}_R^{optimal} = \arg\min_{1\leqslant j\leqslant L}\left[C_{\mathcal{B}j}\pi_R + C_{\mathcal{N}j}(1-\pi_R)\right]$$

- Results to the smallest average cost:

$$\tilde{J}(R) = J(R, \mathcal{D}_R^{optimal}) = \mathbb{E}\left[\sum_{n=1}^{R} c_n + g(\pi_R)\right]$$

Features

Type	Features
	# of exclamation marks, # of uppercase letters, # of emoticons, # of acronyms, # of second person pronouns, # of curse hashtags, # of curse words, density of curse words
	mean value of valence, arousal and dominance respectively

Optimal Stopping Strategy

Optimization Problem

- **Goal**: use **least number of features** for detecting a cyberbullying message without loss of accuracy

Minimize cost function

$$\boxed{\min_{R\geqslant 0}\tilde{J}(R) = \min_{R\geqslant 0}\mathbb{E}\left[\sum_{n=1}^{R} c_n + g(\pi_R)\right]}$$

- Optimal stopping theory problem for Markov processes

Optimal Solution

- Optimal solution via dynamic programming (DP):

$$\underbrace{\bar{J}_n(\pi_n)}_{\text{Optimal cost–to–go}} = \min\Big[\overbrace{g(\pi_n)}^{\text{Cost of stopping}}, \underbrace{c_{n+1} + \sum_{y_{n+1}} A_n(y_{n+1}) \times \bar{J}_{n+1}\left(\frac{p(y_{n+1}|H_{\mathcal{B}})\pi_n}{A_n(y_{n+1})}\right)}_{\text{Cost of continuing}}\Big]$$

Posterior Probability Rule

- Update **posterior probability** via:

$$\pi_n = \frac{p(y_n|H_{\mathcal{B}})\pi_{n-1}}{\pi_{n-1}p(y_n|H_{\mathcal{B}}) + (1-\pi_{n-1})p(y_n|H_{\mathcal{N}})}$$

Conclusions

- Proposed **novel algorithm** for cyberbullying detection

Optimal classification strategy (*optimize classification performance*)

Optimal stopping strategy (*minimize time to raise an alert*)

- Validated performance using **real–world** Twitter dataset comprising more than 10K messages

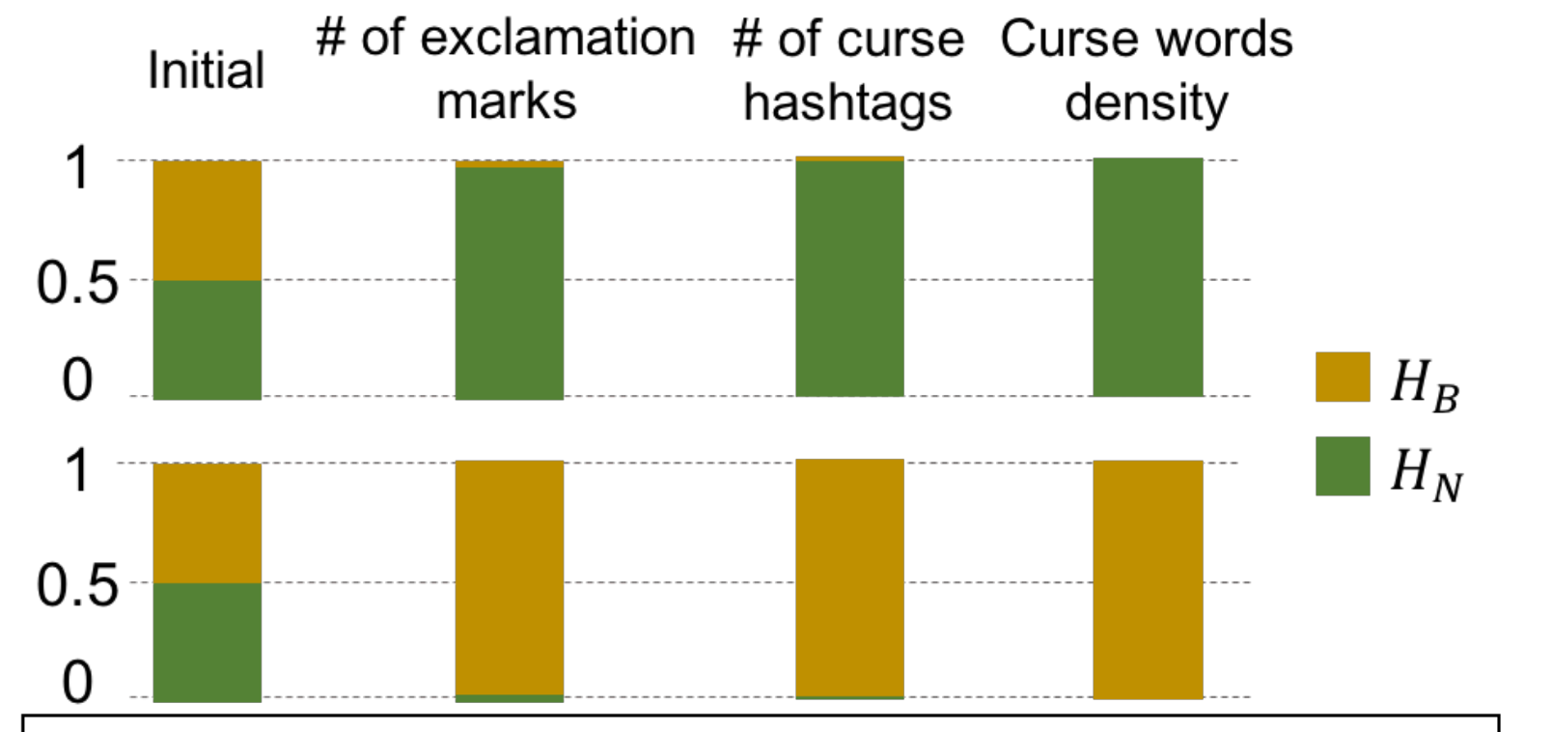
64% reduction in number of features

Numerical Results

Real–word labeled Twitter dataset consisting of **10,600 tweets**

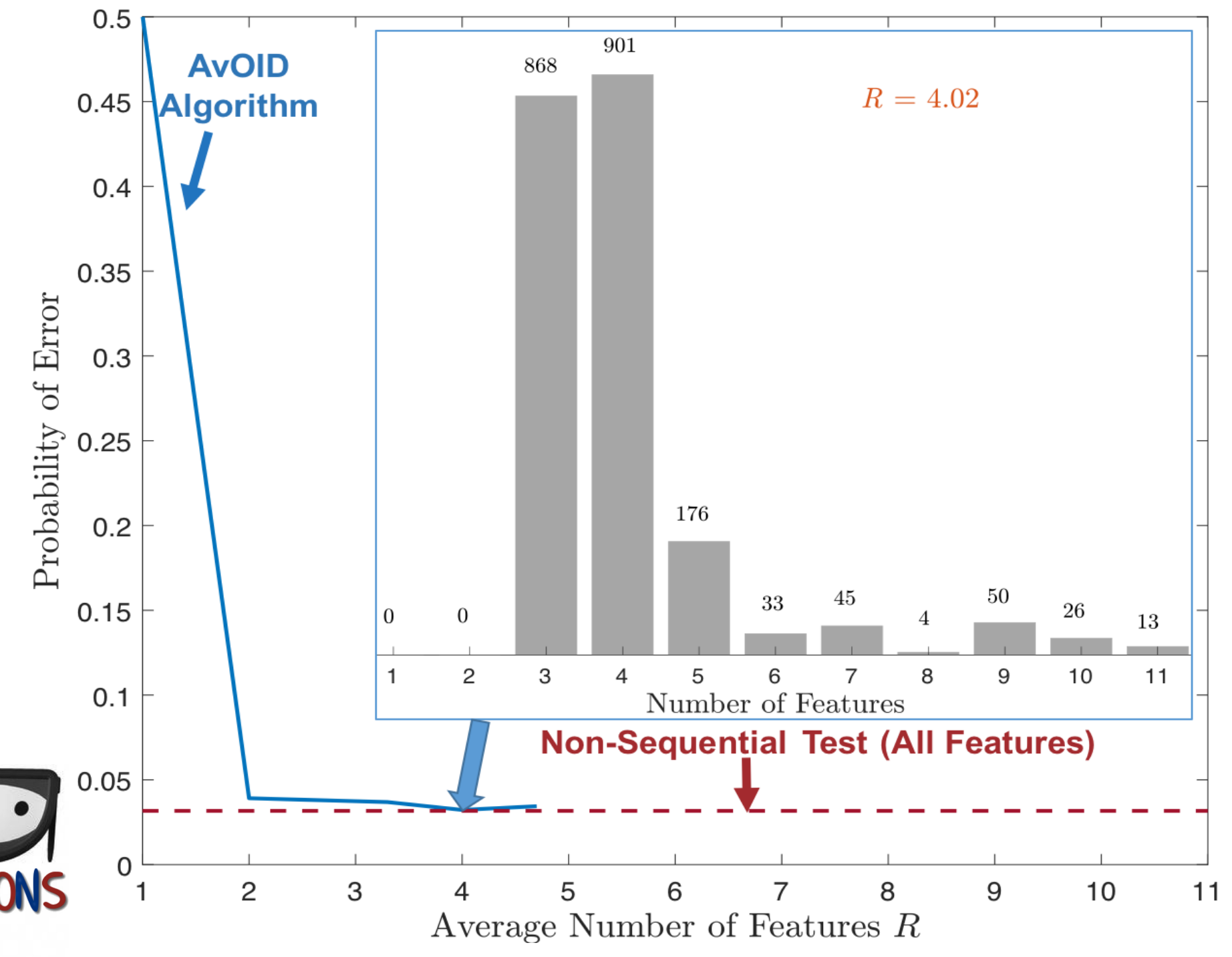
Cyberbullying Detection: *Is it a cyberbullying message?*

yeah, i am realisingthat.....as i recently shiftedtoTweetDeck ...features andfunctionality seemssimilar.Only this is desktopapp



RT @Sanity_3: LiberalMuslim girl fromBelgium is exasperatedwhy didn't Hitlerkill all the Jews. Justwow! https: / /t.co/2aq8DH6JMA

Performance: *Error? Number of features?*



OBSERVATIONS

- Achieves same error probability by using approximately **4 out of 11 features on average**
- In most cases, 3 – 4 features needed for classification