OPTIMAL ONLINE CYBERBULLYING DETECTION

Daphnē–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!

AvOID: A Novel Algorithm for Optimal Online Cyberbullying Detection

Framework

- Optimal classification strategy $H_B$ or $H_N$
- Optimal stopping strategy (Continue or Stop?)
- Posterior probability is set to prior probability of bullying message $P(y_0) = p$
- Optimal stopping theory problem for Markov processes

Features

- Type: Features
  - Number of sentiment words
  - Number of word tags
  - Mean value of valence, arousal and dominance respectively

Classification Strategy

- Optimal classification strategy:
  $D^\text{optimal}_R = \arg \min_{R,C} \left[ C_R \pi_R + C_N (1- \pi_R) \right]$
- Results to the smallest average cost:
  $J(R) = J(R, D^\text{optimal}_R) = \mathbb{E} \left[ \sum_{n=1}^{N} c_n + g(\pi_n) \right]$

Optimal Stopping Strategy

- **Goal**: Use least number of features for detecting a cyberbullying message without loss of accuracy
- Optimal solution via dynamic programming (DP):
  - Cost of stopping $J_n(\pi_n) = \min_{y_{n+1}} \left[ g(\pi_n, y_{n+1}) + \sum_{i=1}^{N} A_i(y_{n+1}) \times \right.$
  - Cost of continuing $J_{n+1}(\pi_n) = \frac{p(y_{n+1}|H_B)\pi_n}{A_n(y_{n+1})}$

Conclusions

- *Related Work*
  - Prior work:
    - Focuses on only 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

- *Numerical Results*
  - Real-world labeled Twitter dataset consisting of 10,600 tweets
  - Performance: Error? Number of features?

- *Optimal Algorithm*
  - Proposed novel algorithm for cyberbullying detection
    - Optimal classification strategy (optimize classification performance)
    - Optimal stopping strategy (minimize time to raise an alert)

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