**Experts and Machines against Bullies: A Hybrid Approach to Detect Cyberbullies**

* Most of the technical studies have focused on the detection of cyberbullying through identifying harassing comments rather than preventing the incidents by detecting the bullies.
* Proposed methods: we introduce the three types of models used for calculating and assigning the bulliness score to the social network users: a multi-criteria evaluation system, a set of machine learning models and two hybrid models that combine the two.  
  + Multi-Criteria Evaluation Systems (MCES): By assigning weights and importance levels to features or criteria, MCES can combine different sources of knowledge to make decisions.  
    - The likelihood, that a bully user belongs to a certain category, was indicated on four-point scale ‘Unlikely’, ‘Less likely’, ‘Likely’ and ‘Very likely’ corresponding to values 0.125, 0.375, 0.625 and 0.875 respectively. The 'I don't know' option was also available.
    - The importance was indicated on a four-point scale of 1: not informative, 2: partially informative, 3: informative and 4: very informative.
  + Machine Learning Approaches: We used three well-known machine learning methods, which use pre-labelled training data for automatic learning: a Naive Bayes classifier, a classifier based on decision trees and Support Vector Machines (SVM) with a linear kernel  
    - The ratio of capital letters in a comment.
    - The number of emoticons.
    - The occurrence of a second person pronoun followed by a profane word in profanity.
    - The term frequency–inverse document frequency (Tf-Idf).
  + Hybrid Approach 1: Using the outcome of the expert system as an extra feature for training the machine learning models. The hybrid system is formed by adding the following features to the machine learning classifier: 1) the results of the MCES, 2) the features’ categories that were used in the expert system as new set of features, and 3) the combined features (C1 and C2).
  + Hybrid Approach 2: Using the results of the machine learning model as a new criterion for the expert system. As previously done in the MCES, we assigned equal weights to all the criteria used in the system, including the machine learner criterion.
* Dataset: Youtube comment section
* Results:
  + The discrimination capacity of the MCES was 0.72.
  + Among the machine learning classifiers the decision tree classifier performed the worst, followed by the SVM classifier. Naive Bayes with discrimination capacity of 0.66 outperformed the other two algorithms.
* Drawbacks: Although the classifiers scored relatively low, but the addition of temporal features such as time of activity could provide a pattern of bullying incidents and when to expect an abnormal behavior.
* Authors: Maral Dadvar, Dolf Trieschnigg, and Franciska de Jong

**A Pattern-Based Approach for Sarcasm Detection on Twitter**

* Feature Extraction:
  + SENTIMENT-RELATED FEATURES
* 59.2% Accuracy
* 65.1% Precision
* 39.7% Recall
  + PUNCTUATION-RELATED FEATURES
* 60.4% Accuracy
* 61.3% Precision
* 56.6% Recall
  + - Number of exclamation marks
    - Number of question marks
    - Number of dots
    - Number of all-capital words
    - Number of quotes
  + SYNTACTIC AND SEMANTIC FEATURES
* 53.7 % Accuracy
* 65.1 % Precision
* 16.1 % Recall
  + - Use of uncommon words
    - Number of uncommon words
    - Existence of common sarcastic expressions
    - Number of interjections
    - Number of laughing expressions
  + PATTERN-RELATED FEATURES
* 90 % Accuracy
* 90.6 % Precision
* 89.3 % Recall
  + - classified words into two categories: high-frequency words and content words

Methodology:

* + Natural Language Processing (NLP)
  + Toolkit weka
  + Support Vector Machine (SVM)
* Dataset: Twitter tweets
* Our proposed approach reaches an accuracy of 83.1% with a precision equal to 91.1%.
* Drawbacks: syntax-related features present a very low accuracy and recall. PoS-tagger performances are lower than when applied to a formal text.
* Authors: MONDHER BOUAZIZI AND TOMOAKI OTSUKI