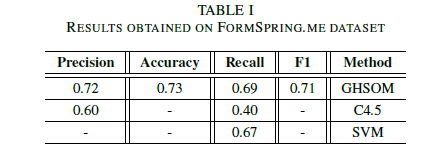
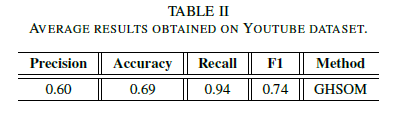
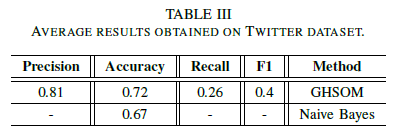
**Unsupervised Cyber Bullying Detection in Social Networks**

* While cyber bullying is a well-studied problem from a social point of view, only recently it has attracted the attention of computer scientists, especially towards automatic detection tasks. For this reason, only relatively few articles on the subject and very few datasets are available.
* We proposed to adopt an unsupervised approach to detect cyber bully traces over social networks







* We now know multiple sources that we can setup as our data sets ( YouTube, twitter, FormSpring)
* While cyber bullying is a well-studied problem from a social point of view, only recently it has attracted the attention of computer scientists, especially towards automatic detection tasks. For this reason, only relatively few articles on the subject and very few datasets are available.
* We proposed to adopt an unsupervised approach to detect cyber bully traces over social networks.
* Supervised learning technique for detecting harassment, using a bag of words model based on content, sentiment and contextual features of documents to train an SVM classifier. Recall level 61.9%.
* Gender approach: They used SVM and the results obtained improved the baseline by 39% in precision, 6% in recall, and 15% in F-measure.
* Applying different binary and multiclass classifiers on a manually labeled corpus of YouTube comments. This approach reached 66.7% of accuracy. Also, in this case authors used an SVM learner.
* Using Amazon’s Mechanical Turk. Authors used rule based learning method and a bag-of-words approach based on a C4.5 decision tree learner and an instance-based learner. They identify true positives cyber bullying posts with an overall accuracy of 47.7%.
* Proposed Model: Our model will avoid a bag-of-words (BoW) (bully traces) can be pre-filtered using syntactic and semantic analysis, using NLP algorithms also sentiment analysis features.
  + Syntactic features
    - Bad words
    - Bad words density: number of bad words that appear in a sentence, for each severity level, divided by the words in the same sentence.
    - Badness of a sentence: weighted average of the “bad” words
    - Density of upper case letters: This feature is given by the ratio between the number of upper case letter and the length (number of chars) of the whole sentence.
    - Exclamations and questions marks:
  + Semantic features
    - Bigrams: Using Part Of Speech analysis, it’s possible to detect, as a feature, the presence of commonly occurring bigram pairs in a bullying sentence such as “you are”, “yourself ”, and so on.
    - Trigrams: The adoption of N-Gram windows inside text can help at least to mitigate some controversial sentences that contain negations
  + Sentiment features
    - Sentiment polarity of a sentence: polarity score is defined as the mean of polarity scores of all the terms. The polarity function is calculated by using the SentiWordNet1 lexicon.
    - Emoticons:
  + Social features
    - Direct User Tagging: On Twitter this can be easily detected by looking at the presence of the special @USER tag.
    - Author profiling: This feature measures the politeness of the author of posts. Our model tries to reflect this behavior to avoid misleading posts.
    - Messages exchanged with a user: This feature tries to gain information about an eventually pre-existent discussion to which the current post analyzed belongs.

**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 (tfidf).
  + 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.
* 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.