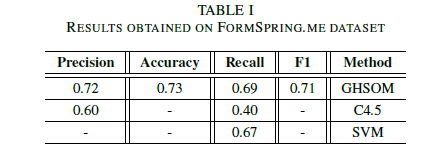
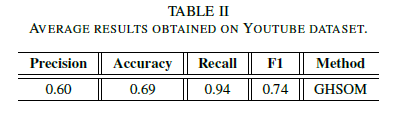
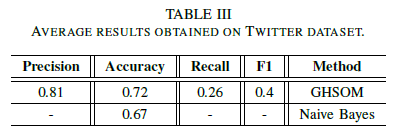
**Unsupervised Cyber Bullying Detection in Social Networks**

* 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





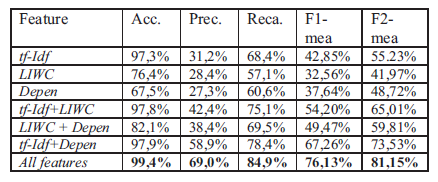
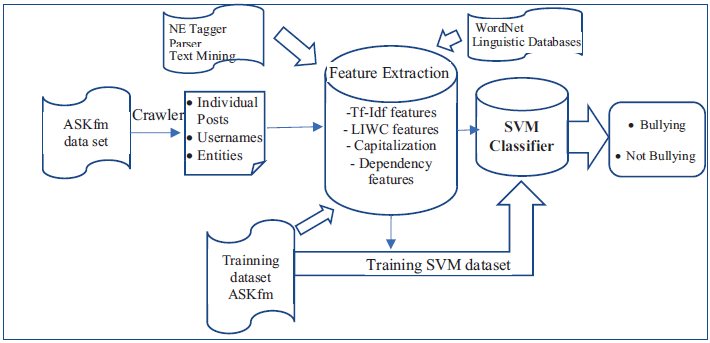


* While cyber bullying is a well-studied problem from a social point of view, only recently it has attracted the attention of computer scientists, especially towards automatic detection tasks. For this reason, only relatively few articles on the subject and very few datasets are available.
* Proposed 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
    - 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**

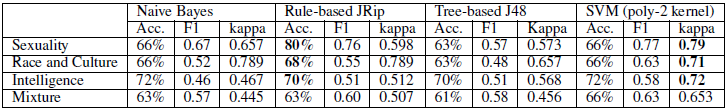
* Most of the technical studies have focused on the detection of cyberbullying through identifying harassing comments rather than preventing the incidents by detecting the bullies.
* Proposed methods: we introduce the three types of models used for calculating and assigning the bulliness score to the 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 (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.
* 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.

**Cyberbullying System Detection and Analysis**

* The system relies on the detection of three basic natural language components corresponding to Insults, Swears and Second Person.
* Lexicons found in the text such as smiley faces, brushing faces, among others, are replaced by their textual equivalent expressions. This will ensure that such symbols are also taken into account in the feature space that will be explained later on.
* Proposed Methods: ***the whole is greater than the sum of its parts***. A combination ofmodestly accurate features coming from heterogeneous data modalities can outperform methods that employ a single modality.
  + Tf-Idf: Our implementation introduces two key novelties. First, WordNet lexical database [17] as well as some SMS repositories
  + Linguistic Inquiry and Word Count(LIWC): concerns the categories; Second person, Total number of pronouns, Swear words, Negative emotion, Anxiety, Anger, Sadness and Sexual. This yields a total of 8 features.
  + Unusual capitalization
  + Dependency features: occurrence of Insult/Swear word is found
* We used support vector machines (SVM) classifier. 
* This work opens up new direction for future research through using advanced parser, dimension reduction and taking into account user’s profile in order to strengthen the detection capabilities.

**Common Sense Reasoning for Detection, Prevention, and Mitigation of Cyberbullying**

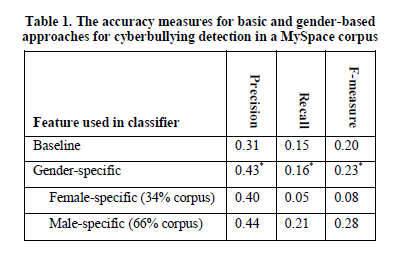
Cyberbullying or harassment on social networks is as much a threat to the viability of online social networks for youth today as spam once was to email in the early days of the internet.



Proposed models: To detect explicit bullying language pertaining to (1) sexuality, (2) race & culture and (3) intelligence. Binary classifiers outperform their multiclass counterparts: JRip and Support Vector Machines were the best performing in terms of accuracy and kappa values.

* Tf-Idf.
* Ortony lexicon for negative affect.
* List of slurs & profanity.
* POS bigrams.
* Topic-specific unigrams & bigrams.
* Future work: We are currently embarking on the use of a family of latent variable models to model, understand and predict self-harm in adolescents, a phenomenon that is not very well understood in the field of abnormal psychology.

**Improved Cyberbullying Detection Using Gender Information**

* We used a supervised learning approach to detect cyberbullying. We constructed a Support Vector Machine classifier using WEKA.
* Four types of features: Profane words, second person pronouns, other personal pronouns, and the weight of the words in each sentence.

Future work: Considering contextual features of the text as well as the word level features. The ground truth annotation can be done through crowdsourcing, investigate other features which may differentiate the writing styles of the users such as age, profession, and educational level.

|  |  |  |
| --- | --- | --- |
|  | **Previous System** | **Our System** |
| Accuracy |  | Definitely higher |
| Methodology | 1. N-gram 2. Counting 3. TF-IDF 4. Pronoun occurrence 5. Skip-grams 6. Classification    1. SVM    2. Logistic regression | 1. Sentiment and contextual features analysis 2. Bag of words 3. Syntactic features 4. Semantic features 5. Sentiment features 6. Social features 7. Linguistic Inquiry and Word Count 8. TF-IDF 9. Unusual capitalization 10. Lexicons and stemming 11. Classification     1. SVM     2. Logistic regression 12. Hybrid classifiers 13. Deep learning |
| Application | No Application | A graphical user interface will be used for furthermore illustration |
| Dataset | Small scale of dataset | Large scale of dataset |

**Cyberbullying Detection using Time Series Modeling**

* Unlike previous approaches that consider a fixed window of a cyber-predator’s questions within a dialogue, we exploit the whole question set and model it as a signal, whose magnitude depends on the degree of bullying content
* Pre-Processing: Stop-words removal, tokenization, stemming and Part-Of-Speech tagging.
* 1st model: simple bag of-words approach, in which all tokens are considered as input feature.
* 2nd model: n-gram character representation, with n equals to 3.
* All four variations were examined, namely Tf-Idf, term frequency, term occurrences and binary term occurrences.
* They used SVM, SVD (Singular Value Decomposition) and   
  DTW (Dynamic Time Warping).
* Their highest result was using the binary term occurrence and it scored 49.8%.

**Machine Learning Approach for Detection of Cyber-Aggressive Comments by Peers on Social Media Network**

* In this paper we have devised methods to detect cyberbullying using supervised learning techniques. We present two new hypotheses for feature extraction to detect offensive comments directed towards peers which are perceived more negatively and result in cyberbullying.
* Methodology
  + Normalization
  + Correcting words
* Standard Feature Extraction
  + N-gram model
  + Counting: Count the number of times each of these tokens occurs in each of the text strings
  + TF-IDF Score
  + The standard features scored 82.69% on accuracy.
* Additional Features
  + Capturing pronouns
    - Scored 86.58% accuracy
  + Skip- grams: For example consider “You are an idiot” as a comment, if we use 2-skip-gram, count of 'You are' as one feature and 'an idiot' as other is added in our feature-matrix.
    - Scored 86.87% accuracy.
* Feature Selection
  + Chi-Square Method
* Classification
  + Support vector machine (SVM)
    - scored 0.7029 on precision and 0.5829 on recall
  + Logistic Regression
    - scored 0.644 on precision and 0.6147 on recall
* Future work should be directed towards detecting sarcastic comments.

**A Multilingual System for Cyberbullying Detection: Arabic Content Detection using Machine Learning**

* With the abundance of Internet and electronic devices bullying has moved its place from schools and backyards into cyberspace; to be now known as Cyberbullying
* None till now covered cyberbullying in Arabic language.
* Methodology
* Features used
  + Unigrams
  + TF-IDF
  + Lexicon
  + Bigrams
* Machine Learning
  + Naive Bayes
    - Scored 90.8514 %
  + Support Vector Machine (SVM)
    - Scored 94.1%
* Further plans incur training the system using deep learning methods instead of machine learning and then comparing the differences of the outcomes from the two schemes.