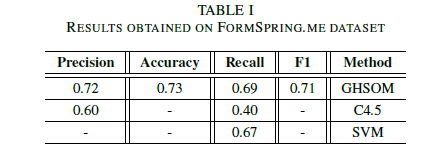
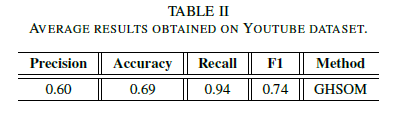
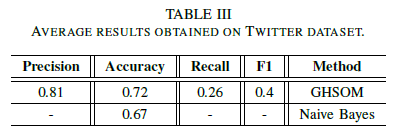
**Unsupervised Cyber Bullying Detection in Social Networks**

* 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





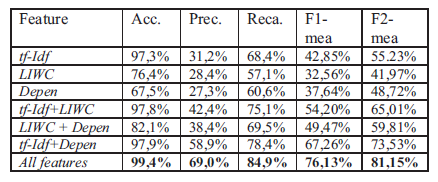
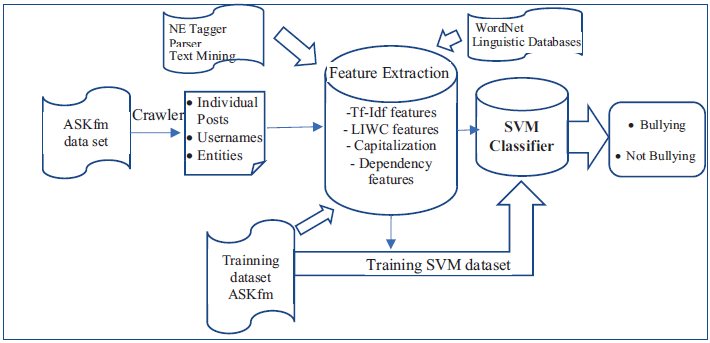


* We now know multiple sources that we can setup as our data sets ( YouTube, twitter, FormSpring)
* While cyber bullying is a well-studied problem from a social point of view, only recently it has attracted the attention of computer scientists, especially towards automatic detection tasks. For this reason, 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
    - 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**

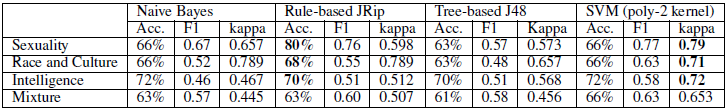
* Most of the technical studies have focused on the detection of cyberbullying through identifying harassing comments rather than preventing the incidents by detecting the bullies.
* Proposed methods: we introduce the three types of models used for calculating and assigning the bulliness score to the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

**Cyberbullying System Detection and Analysis**

* Cyber-bullying has recently been reported as one that causes tremendous damage to society and economy
* The system relies on the detection of three basic natural language components corresponding to Insults, Swears and Second Person.
* Preprocessing: Web links and unknown characters were removed. For each sentence, the incorrect wording is corrected in the following way. The word is first mapped to WordNet lexical database. If an entry is not found, we seek whether it has an entry in the list of saved usernames, Named-entities (using Illinois Named-entity tagger), SMS dictionary / abbreviations (using SMS dictionary Netlingo (www.netlingo.com/acronyms.php). If no entry is found at any of the linguistic dictionaries, we check for the presence of character duplication that will be removed. If neither the original nor the transformed word is recognized, the word is inputted to Norvig spell-correcting algorithm (http://norvig.com/spellcorrect. html), the unknown word is therefore substituted by the suggested correct wording only if its Edit distance with respect to the original is one.
  + we also want to diminish the impact of false negative by avoiding deleting deliberate user’s incorrect wording
* 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.

|  |  |  |
| --- | --- | --- |
|  | **Previous System** | **Our System** |
| Accuracy | ranging between 34% and 66% | Definitely higher |
| Methodology | 1. SVM 2. Bag of words 3. TD-IDF 4. Profane words 5. Second person pronouns 6. Other personal pronouns 7. The weight of the words in each sentence. | 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. Dependency features 11. Lexicons and stemming 12. Machine learning     1. SVM     2. Naive Bayes     3. Decision Tree 13. Hybrid classifiers 14. Deep learning |
| Application | No Application | A graphical user interface will be used for furthermore illustration |
| Dataset | Small scale of dataset | Large scale of dataset |