Automated Cyberbullying Detection using Clustering Appearance Patterns

[Walisa Romsaiyud](https://ieeexplore.ieee.org/search/searchresult.jsp?searchWithin=%22First%20Name%22:%22Walisa%22&searchWithin=%22Last%20Name%22:%22Romsaiyud%22&newsearch=true)

[Kodchakorn na Nakornphanom](https://ieeexplore.ieee.org/search/searchresult.jsp?searchWithin=%22First%20Name%22:%22Kodchakorn%22&searchWithin=%22Last%20Name%22:%22na%20Nakornphanom%22&newsearch=true)

[Pimpaka Prasertsilp](https://ieeexplore.ieee.org/search/searchresult.jsp?searchWithin=%22First%20Name%22:%22Pimpaka%22&searchWithin=%22Last%20Name%22:%22Prasertsilp%22&newsearch=true)

[Piyaporn Nurarak](https://ieeexplore.ieee.org/search/searchresult.jsp?searchWithin=%22First%20Name%22:%22Piyaporn%22&searchWithin=%22Last%20Name%22:%22Nurarak%22&newsearch=true)s

[Pirom Konglerd](https://ieeexplore.ieee.org/search/searchresult.jsp?searchWithin=%22First%20Name%22:%22Pirom%22&searchWithin=%22Last%20Name%22:%22Konglerd%22&newsearch=true)

[Walisa Romsaiyud[1]](https://ieeexplore.ieee.org/search/searchresult.jsp?searchWithin=%22First%20Name%22:%22Walisa%22&searchWithin=%22Last%20Name%22:%22Romsaiyud%22&newsearch=true)in this paper wants to enhance the Naïve Bayes classifier for extracting the words and examining loaded pattern clustering This algorithm consists of two main methods creating partitions by iteratively relocating from entire datasets into clusters using k-mean clustering and capturing any specific partition with the frequency of words with multinomial model feature vector and drawing the probability of words occurring in a document for predicting the eight classes. This preprocesses the data by removing the special and nonprintable characters also they reduce the duplicate words then they generate clusters, data sources are clustered the features two categories of the messages as polite messages and abusive messages, which the contents of the messages are identified, based on a crime pattern and the normalized documents using K-means clustering technique and they used naïve Bayes as a classifier to classify the abusive messages into 8 categories. Their accuracy equal 95.79% and their drawback is that clustering process is not parallelized also the use of K-means is poor on categorical data

Cyberbullying Classification using Text Mining

[Noviantho](https://ieeexplore.ieee.org/search/searchresult.jsp?searchWithin=%22Authors%22:%22Noviantho%22&newsearch=true); [Sani Muhamad Isa](https://ieeexplore.ieee.org/search/searchresult.jsp?searchWithin=%22First%20Name%22:%22Sani%20Muhamad%22&searchWithin=%22Last%20Name%22:%22Isa%22&newsearch=true); [Livia Ashianti](https://ieeexplore.ieee.org/search/searchresult.jsp?searchWithin=%22First%20Name%22:%22Livia%22&searchWithin=%22Last%20Name%22:%22Ashianti%22&newsearch=true)

[Noviantho](https://ieeexplore.ieee.org/search/searchresult.jsp?searchWithin=%22Authors%22:%22Noviantho%22&newsearch=true) wants to increase the accuracy using Naive Bayes method and Support Vector Machine (SVM) then applying n-gram 1 to 5 for the number of class 2, 4, and 11 for each method. They preprocesses the data by removing conversations that have total characters under 15 letters and deleting meaningless words then they tokenize the sentence, remove stop words, transform letter to lower case ,return the words to its basic, keep only the words that are between 3 and 25 then they generate n-grams. They use tf idf as feature extraction They make classifier that classify the data into 2 classes cyberbully, non-cyberbully then another classifier that classify the data into 4 Classes non-cyberbully, cyberbully level severity low, cyberbully level severity middle, cyberbully level severity high and classifier that classify the data into 11 classes (non-cyberbully, cyberbully level severity 1 – 10).they used naïve Bayes and many kinds of svm classifier with experimenting each classifier with n-gram from 1 to 5 the best accuracy was 97.11% with svm poly kernel because the data used in this study are non-linear separable. The drawback in this study is that they using tf idf only as feature extraction and it does not capture position in text, semantics, co-occurrences in different documents For this reason, TF-IDF is only useful as a lexical level feature.

**Automatic Sarcasm Detection using feature selection**

**Paras Dharwal ; Tanupriya Choudhury ; Rajat Mittal ; Praveen Kumar**

[Paras Dharwal](https://ieeexplore.ieee.org/search/searchresult.jsp?searchWithin=%22First%20Name%22:%22Paras%22&searchWithin=%22Last%20Name%22:%22Dharwal%22&newsearch=true) in his paper his primary goal is was understand different methods for Automatic sarcasm detection .they preprocesses the data by tokenize the sentence, then they generate n-grams and they analyze that the sentence is negative or positive by sentiment then they recognize the pattern. They used tfi df and chisquare as feature extraction, and they used naïve bayes and svm as classifiers with f-score with n-gram 0.56 sentiments 0.41 and topics 0.35.the drawback is that Naive Bayes method also does not lead to higher precision, and the chances of poor results are quite high in this. Methods like, SVM need to be more precise in their analysis because these can provide more accuracy.