*Vaibhav Suri*

*Student, GITAM University,*

*psourabh92@gmail.com*

*Student, GITAM University,*

*Visakhapatnam, India*

*Sourabh Parime*

*Cyberbullying Detection and Prevention: Data*

*Mining and Psychological Perspective*

We implemented sentiment analysis using the above dataset with two classes of data: one class was positively identified as an occurrence of cyberbullying while the other was identified as not being an occurrence of cyberbullying. Since textual data is unstructured, it doesn't find a place in the usual databases and is stored in the form of text files. In our process, we first extracted the text from these text files while ignoring their

XML or HTML contents

Dataset

In our study we obtained datasets from the Myspace social networking website, which were manually marked as cyberbullying. This data was then extracted into multiple XML files and stored in a folder. Using the RapidMiner data mining tool, the data from multiple documents was then read into local repositories

Preprocesses

Tokenize: In this operation, the extracted text is broken into a set of words or tokens, where a token is the most basic unit of a text

2. Filter Tokens: The tokens are filtered depending upon their length. In this step the max and min parameters chosen depend upon the context of the application. Since the usual "chat lingo" consists of relatively smaller word lengths, the parameters were chosen appropriately.

Stem: The stemming operation involves reducing words to their root-words, [11] for example stemming the word "computation" to its root "comput". This process facilitates the identification of similar tokens.

4. Filter Stop words: This operation eliminates the common words in the English language since they form the noise in the data.

Classifier

used consisted of a linear SVM (Support Vector Machine) with

Improving existing algorithms we propose a hybrid process in which the first stage involves computing and recording the frequencies of words in a dataset over a time period for the high-activity hours of teenagers on the internet. In the second stage, sentiment analysis is applied to the dataset to identify offensive texts with certain degree of certainty in order to avoid false positives. We use the stemming process to facilitate the identification of similar tokens and remove the stop words to eliminate the noise in the data

Accuracy :NONE

Drawbacks

They have no features in their method and the dataset is so weak due to that is collected manually