Semi-supervised Learning for Cyberbullying Detection in Social Networks

Abstract: Current approaches on cyberbullying detection are mostly static: they are unable to handle noisy, imbalanced or streaming data efficiently. Existing studies on cyberbullying detection are mainly supervised learning approaches, assuming data is sufficiently pre-labelled. However this is impractical in the real-world situation where only a small number of labels are available in streaming data. In this paper, we propose a semi-supervised leaning approach that will augment training data samples and apply a fuzzy SVM algorithm. The augmented training technique automatically extracts and enlarges training set from the unlabeled streaming text, while learning is conducted by utilizing a very small training set provided as an initial input. The experimental results indicate that the proposed augmented approach outperformed all other methods, and is suitable in the real-world situations, where sufficiently labelled instances are not available for training. For the proposed fuzzy SVM approach we handle complex and multidimensional data generated by streaming text, where the importance of features is discriminated for the decision function. The evaluation conducted on different experimental scenarios indicates the superiority of the proposed fuzzy SVM against all other methods.

Dataset:   
we utilized data provided by Fundacion Barcelona Media2 for the workshop on content analysis from the Web 2.0. The given data was collected from the three different SNs including Myspace, Kongregate, and Slashdot.

Methodology:

1. Linguistic Features
   1. Keywords based features, which involve binary representation of the keywords to see if the keywords are presented or not.
   2. The number of swearwords in posts, divided by the total number of the words in messages.
   3. Presence of pronouns
   4. To capture a degree of users’ emotions
   5. The normalized value of capital letters within messages
   6. Special characters
   7. Users’ age and gender
2. Cyberbullying Detection
   1. Algorithm 1: Model building using Augmented training set
      1. Naive Bayes
      2. Logistic Regression
      3. Random Forest
      4. Stochastic Gradient Descent classifier: Designed for the text data. It generates feature space using an STWV filter to transform text strings into term-weight vectors based on Vector Space Model
   2. Algorithm 2: Kernel Fuzzy C-Mean Clustering Algorithm
      1. Fuzzy Approach:
      2. Clustering Process: enables us to evaluate all features and calculate their degree of relevance to a specific group. Clustering is used to find high intracluster and low inter-cluster similarities. The idea is to find natural groupings among similar objects.
   3. Algorithm 3: Fuzzy SVM Classification algorithm
      1. Fuzzy Classifier: fuzzy classifier is used to handle unbalanced and unstable text streams generated from social networks. The dataset is fed into the KFCM model to extract membership values.

Drawbacks:

1. Insufficient training instances
2. Uncertain and imbalance feature distribution
3. The classification of cyberbullying messages is a critical issue because of different impacts made by the false positives(identify non-cyberbullying instances as cyberbullying itself) and the false negatives (system should not miss out the cyberbullying posts)