Detect Cyberbullying on Social Media

Introduction:

Federal government website ‘stopbullying.gov’ defines cyberbullying as bullying which takes place over digital devices like computers, cell phones, tablets and can occur through SMS, text, apps, or online where people can view, participate in, or share content. It also includes sending, posting, or sharing negative, harmful, false, mean, or personal information that can bring about embarrassment or humiliation to the recipient'.

Online interaction data contributes to a significant percentage of textual data such as posts, messages, and comments on social networks. With the proliferation of the communication, there is a corresponding increase in concern surrounding the nature of this content. Textual interactions that signify disturbing and negative phenomena such as online harassment, cyberbullying, cyber threats, stalking, and hatred are on the rise. This detrimental online behavior can have significant traumatic effects on the individual and lead to severe psychological problems.

Problem Statement:

In this project, we train our model on Formspring data and predict the messages from Myspace data if it is bullying content, or otherwise, using Logistic Regression, Multinomial Naive Bayes and Perceptron network algorithms.

Datasets:

1) Formspring

It is a question-and-answer-based social network. The site allows its users to follow others privately. People could also ask questions of their followers from the homepage. Formspring garnered controversies, as many teenagers committed suicide on account of harassment and cyberbullying due to the anonymity of the entries.

The data represents 50 ids from Formspring.me. For each id, the profile information and each post (question and answer) was extracted. Each post was loaded into Amazon's mechanical turk and labeled by three workers for cyberbullying content.

The data contains the following profile fields :-
BIO - profile biography created by owner of the id
DATE - the date the id was crawled
LOCATION - location provided by the owner of the id
USER ID - The actual id itself

The data contains the following information on each post :-
TEXT - the question and answer
ASKER - the id of the person asking the question (blank if anonymous)

Occurrences of label data -

ANSWER - YES or NO as to whether the post contains cyberbullying
SEVERITY - cyberbullying severity from 0 (no bullying) to 10

Formspring Dataset Link: https://www.kaggle.com/swetaagrawal/formspring-data-for-cyberbullying-detection

This dataset has approximately 18500 posts out of which 2500 posts contain bullying content.
There is 15% bullying in training data.

2) MySpace

It is a social networking website offering an interactive, user-submitted network of friends, personal profiles, blogs, groups, photos, music, and videos. The data has been manually labeled for bullying content by three independent coders. The data is in the form of xml files.

The data contains following fields:
1) User id: who created post
2) Username
3) Sex
4) Age
5) Place (City, Province, Country)
6) Date
7) Post text

This dataset has total 11000 posts. We use 400 posts for prediction.

Dataset Link: http://www.chatcoder.com/DataDownload

Feature Engineering:

We perform various preprocessing and feature extraction techniques to build our model. Our model involves the following steps to be performed:

1. Data cleaning and formatting – noise removal.
2. Tokenizing
3. Stop words removal
4. Spell checking
5. Stemming
6. Building vocabulary from training data and it's synonyms fetched from wordnet
7. Term frequency computation
8. Building the bullying dictionary.

The features we use, to train our models, based on the above processes are:

1. Term frequency
2. Bullying word dictionary and it's synonyms (fetched from wordnet).

Model:

Based on the above results, we use these 2 primary features – term frequency and bullying dictionary - for training, validation and prediction of bullying. We train the following algorithms using our feature vectors and use them for validation and predcition:

1. Logistic Regression
2. Multinomial Naive Bayes
3. Perceptron

The parameters/values used for fine-tuning our algorithms are:

1. Logistic Regression:

   penalty: l2
   dual: False
   tol: 0.0000001
   solver: saga/liblinear
   max_iter: 10000
   n_jobs: -1

2. Multinomial Naive Bayes:

   alpha=1.0,
   fit_prior=True,
   class_prior=None

3. Perceptron Network

   penalty: l2
   alpha: 0.00001
   max_iter:1000
   shuffle:True
   n_jobs: -1
   random_state: 3

Results:

	Logistic Regression	**Multinomial Naive Bayes**	**Perceptron Network**
Training Accuracy	87.31%	85.3%	82.94%
Testing Accuracy	86.93%	84.80%	82.86%
Number of texts predicted as Bullying in MySpace data (out of 400)	66	151	96

Learning:
This project had the following learning outcomes:

1. We learnt the different feature vectors which are needed to get the best throughput.
2. We understood the feature extraction algorithms and implemented them.
3. We had to research and learn the various algorithms that can be used for this nature of data.
4. We understood, by fine tuning the parameters, the various ways through which we can improve efficiency.

5. The efficiency throughput of various algorithms on our dataset.
6. We understood, on a broad view, what are the algorithms and how they would fit into the real world examples, as in which algorithm would be the best suited for the nature of data.

Project outcomes:

From the aforementioned results, we observe the following:

1. MySpace has very less (~12.5%) bulying.
2. Logistic Regression has the best prediction, compared to the other 2 algorithms – although Multinomial Naive Bayes has a higher accuracy.
3. FormSpring (training data) has bullying of about 15%.

References:

1. Prediction of Cyber bullying Incidents on the Instagram Social Network
2. Cyberbullying Identification Using Participant-Vocabulary Consistency
3. Modeling the Detection of Textual Cyberbullying
4. Reynolds, K, A. Kontostathis and L. Edwards. 2011. Using Machine Learning to Detect Cyberbullying. In Proceedings of the 2011 10th International Conference on Machine Learning and Applications Workshops (ICMLA 2011). December 2011. Honolulu, HI.
5. Formspring and MySpace Dataset
6. MySpace – Wikipedia
7. StopBullying.gov