Proposal: Cyber-Bullying Detection and Classification System

Marc Mailloux
marcmailloux@knights.ucf.edu

Rolando Nieves
rolando.j.nieves@knights.ucf.edu

Maxim Shelopugin
maxim.shelopugin@knights.ucf.edu

1. MOTIVATION & PROBLEM STATEMENT

Among all the forms of bullying and harassment recognized today, bullying via platforms hosted on the Internet has proliferated at an alarming rate. The rate of proliferation has been concerning enough to motivate the United States Centers for Disease Control and Prevention (CSC) to acquire and report data regarding "electronic" bullying (or, as it is more commonly known, "cyber-bullying") via its biennial Youth Risk Behavior Surveillance System (YRBSS) [1].

Collection of data from cyber-bullying victims has helped immensely as it pertains to awareness and prevention. It should be possible to further improve in both areas if data collection overcomes the limitation of the "a posteriori" nature of current methods (i.e., prevention methods derived from data voluntarily disclosed by victims after harassment has occurred). An automated system capable of examining electronic communication traffic, identifying and classifying any traffic that could be considered as bullying, could lead to solutions that can either filter out any offensive content automatically, or significantly shorten the response time to an incident, possibly preventing the more dire secondary effects of cyber-bullying.

2. RELATED WORK

There has been much research done around this topic, as evidenced by the body of work on display on the Cyber-Bullying Research Center research summary page [2]. Approaches employed to date include relatively simple Bayes type classifiers, as well as more complex deep learning systems (Sweta Agrawal et. al. [3]). One existing implementation that serves as an excellent exemplar to this proof-of-concept proposal is the *Anti Bully* project by Michelle Li [4]. The work in this proposal differs from *Anti Bully* in two important ways:

- The system proposed in this paper will not rely solely on Naive Bayes classification algorithms the way *Anti Bully* does.
- The proposed system will be able to further refine the binary classification done in *Anti Bully* (which only identifies traffic as **bullying** or **not bullying**) by categorizing bullying traffic among one of the classes listed in Section 3.

The *Anti Bully* system code base does include a good data set which can be used in the proof-of-concept proposed in this paper, albeit with some modifications (see Section 4.1).

3. CYBER-BULLYING DETECTION SYSTEM

The proposal presented in this paper advocates for the implementation of a proof-of-concept meeting the desired features as outlined in Section 1. Leveraging technological advances in Natural Language Processing, along with algorithms borrowed from the areas of Artificial Intelligence and Machine Learning, the system will be trained to recognize and classify cyber-bullying traffic. The classification will go beyond simply identifying bullying traffic, but will also categorize the bullying into three (3) distinct classes:

Cultural Harassment Offensive content primarily focusing on the victim's race or religious beliefs.

Sexual Harassment Amplification of stereotypical behavior for a given gender, or gender identification.

Personal Attacks Ridicule based on a victim's outward-visible attributes, such as appearance, mannerisms, or intelligence.

The proposed system, at its core, will be very similar to classification systems common in the areas of Artificial Intelligence and Machine Learning. The system will combine a classifying apparatus based on machine learning algorithms, such as Artificial Neural Networks (ANN), k-Nearest Neighbor (kNN), or Support Vector Machines (SVM), with a natural language vector representation mechanism, such as *Word2Vec*.

4. EVALUATION

The primary focus while evaluating the system's performance will be to maximize detection of bullying traffic (i.e., maximize true positives while minimizing false negatives), with the minimization of otherwise innocuous traffic (i.e., maximizing true negatives while minimizing false positives).

Focusing on the *Recall* classification metric (i.e., the proportion of true positives classified properly) first and foremost will provide a clear assessment regarding the aforementioned behavior we seek from the system. A system

ACM ISBN 978-1-4503-2138-9.
DOI: 10.1145/1235

that perfectly identifies all cyber bullying instances while not misclassifying any innocuous traffic would exhibit a *Recall* score of 1.0. Although such perfect performance is likely not attainable, tuning of the system will focus on maximizing the *Recall* score.

Upon quick inspection, no single supervised learning classifier seems like an out-right favorite to perform optimally in the problem domain addressed by the proof-of-concept system. Thus, the team will implement various classifiers that can easily plug into the proof-of-concept system. Only one classifier will be used at any one time, and a 95% confidence interval statistical analysis will determine which one of the classifiers, if any, is significantly better than the others.

4.1 Data Set

The data set we propose to use in this proof-of-concept, as alluded to in Section 2, will be a modified version of the data set provided by the *Anti Bully* system [4]. The data set as provided is labeled, but the labels only contain two (2) classes: **bullying** and **not bullying**. In order to meet the goals for the system as detailed in Section 1, the data set labeling will have to be enhanced such that the samples labeled as **bullying** are distributed among the categories identified in Section 3.

5. EXPECTED OUTCOMES AND RISK MANAGEMENT

Similar classification exercises, such as that documented by Sweta Agrawal [3], achieved an average Recall score of 0.87. Thus, we expect this proof-of-concept to at least meet this performance baseline.

One risk that may lead to a revision of the proof-of-concept design lays with the data set. As it stands now, we have not been able to ascertain how many of the samples in the *Anti Bully* data set fit into the categories as listed in Section 3. Some of the categories may be either severely under-represented or not present at all. Should that be the case, the system will either need to employ classification weighting (to account for data set class imbalance) or classification re-design (for the case when a category is not found in the data set).

6. PLAN AND ROLES OF COLLABORATORS

The proof-of-concept implementation exercise will be divided into five (5) primary task areas. The order in which the tasks are presented in this paper does not necessarily represent the chronological order in which the tasks will be carried out.

6.1 Data Set Labeling

As stated in Section 4.1, the source data set for this proof-of-concept will need to be refined in order to accommodate the classification taxonomy as described in Section 3. Although all team members will contribute to this task, **Mr. Marc Mailloux** will serve as the task's point of contact.

6.2 Text Representation

Transforming the training and test sets from plain english into a form that the proof-of-concept system can ingest is fundamental. For this task, **Mr. Marc Mailloux** will serve as both the primary developer and the point of contact.

6.3 Classifiers

The classifier task, given its wide breadth, will be divided among all team members. A list of the classifiers that will be used, along with a short explanation justifying their use, as well as the classifier's implementation point of contact, follows:

- Artificial Neuron Network (ANN) - ANNs possibly represent the most flexible of classifiers, both in the form of input they accept as well as the output they produce. The point of contact for this classifier will be **Mr. Rolando Nieves**.

- Support Vector Machines (SVM) - SVMs have the possibility of requiring the least amount of computing power during both training (as compared with ANNs) and classification. The point of contact for this classifier will be **Mr. Marc Mailloux**.

- Naive Bayes - This was the classifier used in the *Anti-Bully* system this proof-of-concept will be based on. It will be interesting to assess the impact, if any, of implementing the same classifier on a multi-class environment. The point of contact for this classifier will be **Mr. Maxim Shelopugin**.

The statistical significance testing, as described in Section 4, will be implemented by **Mr. Rolando Nieves**. Mr. Nieves will also serve as the point of contact for the classifier task.

6.4 Presentation

Mr. Maxim Shelopugin will be responsible for producing any materials required to present the results of this proof-of-concept to interested parties, including a summary poster.

6.5 Final Report

Although all team members will contribute content to it, **Mr. Rolando Nieves** will be responsible for producing the final report that will document the results observed while exercising the proof-of-concept system documented in this proposal.

7. REFERENCES

[1] "Cyberbullying Research Center: Facts Page," 2018.

[2] "Cyberbullying Research Center: Research Summaries Page," 2018.

[3] S. Agrawal and A. Awekar, "Deep learning for detecting cyberbullying across multiple social media platforms," in *European Conference on Information Retrieval*, pp. 141–153, Springer, 2018.

[4] M. Li, "Anti Bully," 2016.