SURVEY ON AGGRESSION DETECTION IN ONLINE USER CONTENT

Dr.A Ajina[*] , Sahana B S[**] , Sandhya G[**] , Sushma Ellur[**], Tanuja R S[**]

[*]Asst. Prof, Dept. of CSE, Sir MVIT- Bangalore
[**]Dept. of CSE, Sir MVIT- Bangalore

***Abstract*- In today's world, social media has become ubiquitous, used by many to share content, express opinions, comment, among others. While the existence of such a free space has many advantages, it is inevitably accompanied by drawbacks. Some of them being cyber-bullying, harassment, propagation of hate, radicalization, usage of profane language targeting an individual, community, ethnicity or an organization. Though many social media sites take relevant measures once an incident is reported, it is often after harm and damage is done. Instead of working on issues in the aftermath of disturbing incidents, it is best to prevent aggressive or abusive content from being put online. In view of this, a lot of research is being done in the Natural Language Processing domain to classify content between aggressive and non-aggressive, hate and profanity, and some work towards multi-class classification. In this paper, we review some of the work that has been done in the field of Natural Language Processing with the detection of hate or aggressiveness in social media content as the use case.**

I. INTRODUCTION

Social media has always been a debatable subject. Some consider it as a boon while others have a difference in opinion regarding the same. As of June 2017, the world's population stands at 7.5 billion with 3.8 billion people using the internet while the active users were estimated to be 2.89 billion which accounts for a penetration of 29%.

An average kid gets their first cell phone at the tender age of eleven and starts going online, thus opening lots of doors for cyberbullying. Anonymously people can text or email or post a nasty comment or vicious remark or unappealing photos which can get shared with others thus creating a nasty, disgusting, ineluctable spiral of abuse. These rates only get higher when it comes to a tween or teenager. Bullying is a major public health problem. Studies suggests that there is an association between adolescent bullying and poor mental health. Over half of adolescents and teens have been bullied while the same ratio of them have engaged themselves in bullying. One-thirds have experienced cyber threats. Over half of young people don't confront this to their parents.

Now that problem has been laid out, we must also come up with solution for the same. Although a quick solution could be as simple as shutting down of the computer or mobile devices or even deactivating the accounts, these cannot be considered as long-term solutions.

Here we shall focus on harsh comments that have been posted on social media which can be a potential form of cyberbullying. In this paper there is classification of the comments into nonaggressive, hate and profane statements using NLP and Deep learning. Hate speech is an online common form for expressing prejudice and aggression. They may convey racist, xenophobic and many forms of verbal aggression. Hate speech is typically defined as the act that disparages a person or people on the basis of a number of characteristics that may include and not limited to race, ethnicity, sexual orientation, gender, religion and nationality[15] Although deep learning methods show hopeful future in text mining tasks however they are not always better when compared to traditional supervised approaches considering the fact that the performance of the former is subject correct choice of algorithm and number of hidden layers combined with feature representation techniques and tuning of hyperparameters. Some of the bog-standard traditional approaches include CNN and RNN.

Some challenges include presence of words or sentences of multiple languages, convoluted sentences, sentences expressing meaning indirectly. While the former is true for unsupervised learning, in the case of supervised learning, the mindset, opinions, and beliefs of the annotators become an issue [7]. Others include the lack of grammar correctness and syntactic structure of social media posts limited context provided by each individual post irony and sarcasm racial and minority insults, which might be unacceptable to one group, but acceptable to another one.

II. LITERATURE REVIEW

One of the earliest works in the detection of hate speech was carried out by (Warner and Hirschberg, 2012) wherein feature templates were generated from the corpus and fed to SVM classifier. Brown clusters were used for feature representation. They recognized seven categories (anti-semitic, anti-black, antiasian, anti-woman, anti-muslim, anti-immigrant or otherhate) to which the input can be identified with and built six classifiers (ignoring otherhate). The study mainly dealt with identifying anti-semitic data. A very similar approach of using SVM to perform a multi-class classification was adopted by Malmasi et al (2018). In this research, the authors have studied the efficacy of single classifier trained for each feature, ensemble classifier using the individual classifiers, meta-classifier trained with a linear SVM and Radical Basis Function (RBF). Both the metaclassifiers outperform the rest of the classifiers, with RBF kernel performing the best.

Burnap and Williams (2014) studied hate in twitter data. The authors suggest a surge in the speed of hate occurs in the aftermath of incidents like publicized murder, riots, terrorism, considering them as the "trigger" events. The data that was

collected for their study was from twitter and immediately after the infamous murder of Lee Rigby, a British Army Soldier. They implemented a Bayesian Logistic Regression, SVM, RFDT classifiers to classify the tweets as hateful/antagonistic or not. The results suggest that overall the most efficient features for classifying hate speech are n-gram typed dependencies combined with ngram hateful and antagonistic terms.

S. Madisetty et al. (2018) approach uses three deep learning methods, namely, CNN with a static and non-static channel, LSTM, BiLSTM. A majority voting-based ensemble method is used to combine these classifiers. The data was classified into three classes: Overtly Aggressive (OAG), Covertly Aggressive (CAG) and non-aggressive (NAG) posts. It was observed Aggression is hidden in CAG posts and it is very difficult to identify that type of posts. There was also a

lot of confusion between OAG and CAG. The authors however did not incorporate any feature-based models in addition to the deep learning models.

Marzieh Mozafari *et al* [16] proposed a pretrained (Bidirectional Encoder Representations from Transformers) BERT language model to detect hate speech on twitter dataset. The contextual information is then extracted from
BERT's pre-trained layers and then fine-tuned using annotated datasets.This outperformed all other techniques (including ensemble, binary classifiers, multi classifiers) in terms of precision, recall, and F1-score. It also helped to detect biases occurred during the annotation of datasets.

Laura P. Del Bosque *et al* [10] proposed aggression detection as a linear regression problem where documents were assigned an aggressiveness score between 1 and 10, 1 being the least aggressive and 10 being most aggressive. They also explored other techniques such as lexicon-based systems (various lexicon includes SentiWordNet, ANEW, swearword.com), fuzzy systems and multilayer perceptron neural network. Among all these techniques linear regression yielded better results. However, the score predicted was heavily dependent on the use of profane language in the document the structure and semantics of the sentences and the correlation between them were not considered.

Ying Chen *et al* 2012, proposed a Lexical Syntactic Feature (LSF) architecture to detect offensive language in social media. They obtained dataset from the text comments of top 18 YouTube videos (as of 2012) and performed automatic spelling and grammar correction. Ngrams, Bag of Words(BoW), lexical and syntactic features were extracted at sentence level and other user specific data like their language style and their sentiment of their historic comments were considered to train using various Machine Learning techniques like NaiveBayes and SVM.This model outperformed the traditional model for incorporating additional user centric data.However, this model fails to differentiate between hate speech and profanity, as they have assumed profanities are always undoubtedly offensive when directed at users or object, which is not quite true. Thus, resulting in more True Negatives.

In this paper [11] hate speech is detected by converting the paragrapg2vec for joint modeling of the comments and words. By using the embeddings to train a binary classifier they identify between the hateful and clean comments.

Jun-Ming et al. identified a wide range of emotions in bullying traces and proposed a fast learning procedure to train a model to recognizing them automatically. They classified the emotions in bullying traces as seven categories like anger, embarrassment, empathy, fear, pride, relief and sadness. Their learning procedure included collecting seed words, collecting online documents, creating feature extractors and building text classifiers.

A detailed study [8] of relationships between cyberbullying, cyber aggression, social graph feature, temporal commenting behavior, linguistic content and image content is done. By considering the majority voting criterion to a label they classify it accordingly. They applied Linguistic Inquiry and Word Count (LIWC), a text analysis program to find which categories of words have been used for cyberbullying/cyber aggression in labelled media session.

In this paper, the study used Twitter dataset to classify tweets as aggressive bullying/spam. The study was based on user attributes, network attributes and text attributes. After the crowd flower task of annotating the data and applying various techniques such as Naive Bayes, random forest. The conclusion included that using network attributes increased the performance of the model [5].

In this paper, the comments from YouTube were considered. Two experiments were conducted. 1)Training binary classifiers to classify the given instance into sensitive topic or not. 2) Multiclass classifier to classify an instance from a set of sensitive instances. Their findings suggested that former performed better than the latter. The dataset was classified into three classes namely, sexuality, race or culture. Experiments were performed on each of them classes and the results were pruned. As a second experiment, the study combined the classes. The observation was that the labelspecific classifiers performed better than multiclass classifier. [3]

Cynthia Van Hee et al (2015) research was to gain insight into the linguistic characteristics of cyberbullying by collecting and annotating Dutch social media corpus. The study classified the dataset of comments into a scale ranging from 0-2. Sexual talk received a harmfulness score of zero, which conveyed that those instances contained harmless sexual talk. Utterances considered sexual harassment were assigned a score of 1 or 2. The study sed Support VectorMachines (SVM) as the classification algorithm with BOW and sentiment lexicons for representing features. The authors suggest that the results were better in classifying a sub category which was highly lexicalized.

III. TABLES AND FIGURES

REFERENCES	DATASET	FEATURE REPRESENTATION	TECHNIQUES
S. Madisetty et al [2]	Facebook	Glove word embeddings	CNN, LSTM, and BiLSTM.
Malmasi et al [7]	Twitter	n-grams, skip-grams and clustering (brown cluster)	SVM, RBF
Jun-Ming Xu et al [6]	Twitter	Unigram, bigram	SVM
Segun et al [9]	Facebook	word2vec, Glove, SSWE, and fastText.	CNN, LSTM
Burnap et al[14]	Twitter	Bag of words, type dependencies, hateful and derogatory n-grams	Bayesian Logistic Regression, Random Forest DecisionTree (RFDT), SVM
Nemanja Djuric et al [11]	Yahoo	Word embeddings, Bag of words paragraph2vec	SVM
Despoina Chatzakou et al [5]	Twitter	User,text,network level features	Naive Bayes,Random Forest Classifier
Karthik Dinakar et al [3]	Youtube	TF-IDF weighted unigrams,label specific unigrams	Naive Bayes,SVM
Cynthia Van Hee et al [4]	Ask.fm	Word unigram and bigram bagsof-words	SVM
Ying Chen *et al* [12]	Youtube	N-grams, Bag of Words (BoW), lexical and syntactic	Naive Bayes, SVM
Laura P. Del Bosque *et al* [10]	Twitter	Bag of Words.	Linear Regression, Fuzzy logic, Multilayer Perceptron Neural Network, Lexicon based Training

Marzieh Mozafari *et al* [16]	Twitter	word2vec, Glove	BERT (Bidirectional Encoder Representations from Transformers)

Table 1: summary of the techniques used

IV. CONCLUSION

In this paper, we presented a survey of the previous work done in the automatic detection of hate speech. Various Machine learning and Deep Learning methods have been employed for this purpose. Several ways of representing the features have been used like the bag of words, word embeddings, n-grams, POS tags being very adopted very often. Social media still remains a hotbed to spread hate speech and other offensive content. Aggression, cyberbullying, usage of abusive language is largely under the umbrella of hate speech. Some papers also work for the detection of these specifics. The future work includes incorporating various robust deep learning architectures to overcome some of the challenges in the
detection of hate speech and to achieve better accuracy

V. ACKNOWLEDGEMENT

We would like to thank Visvesvaraya Technological University for supporting this study. In addition, we are indebted for the support and encouragement of our guide Dr. A Ajina.

VI. REFERENCES

[1] Gabriel Emile Hine, Jeremiah Onaolapo, Emiliano De Cristofaro, Nicolas Kourtellis Ilias Leontiadis, Riginos Samaras, Gianluca Stringhini, Jeremy Blackburn, "Key, Cucks, and God Emperor Trump:A Measurement Study of 4chan's Politically Incorrect Forum and Its Effects on the Web ", 2017

[2] Sreekanth Madisetty, Maunendra Sankar Desarkar, "Aggression Detection in Social Media using Deep Neural Networks", 2018 [3] Karthik Dinakar, Roi Reichart, Henry Lieberman, "Modeling the Detection of Textual Cyberbullying", 2011

[4] Cynthia Van Hee, Els Lefever, Ben Verhoeven, Julie Mennes, Bart Desmet, Guy De Pauw, Walter Daelemans and Veronique, "Automatic Detection and Prevention of Cyberbullying",2015

[5] Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Emiliano De Cristofaro], Gianluca Stringhini], Athena Vakali, "Mean Birds: Detecting Aggression and Bullying on Twitter".

[6] Jun-Ming Xu, Xiaojin Zhu, Amy Bellmore, "Fast Learning for Sentiment Analysis on Bullying".

[7] Shervin Malmasi, Marcos Zampieri, "Challenges in Discriminating Profanity from Hate Speech", 2018

[8] Homa Hosseinmardi1(B), Sabrina Arredondo Mattson2, Rahat Ibn Rafiq, Richard Han1, Qin Lv1, and Shivakant Mishra1, "Analyzing Labeled Cyberbullying Incidents on the Instagram Social Network", 2015

[9] Segun Taofeek Aroyehun, Alexander Gelbukh, "Aggression Detection in Social Media: Using Deep Neural Networks, Data Augmentation, and Pseudo Labeling", 2018

[10] Laura P. Del Bosque and Sara Elena Garza, "Aggressive Text Detection for Cyberbullying", 2014

[11] Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, Narayan Bhamidipati, "Hate Speech Detection with Comment Embeddings".

[12] Ying Chin, Yilu Zhou, Sencun Zhu1, Heng Xu, "Detecting Offensive Language in Social Media to Protect Adolescent Online Safety", 2012.

[13] W. Warner and J. Hirschberg, "Detecting Hate Speech on the World Wide Web," no. Lsm, pp. 19–26, 2012.

[14] P. Burnap and M. L. Williams, "Hate Speech, Machine Classification and Statistical Modelling of Information Flows on Twitter: Interpretation and Communication for Policy Decision Making," in Proceedings of the Conference on the Internet, Policy & Politics, 2014, pp. 1–18.

[15] Noopur Tarwani and Prof. Uday Chorasia, "Survey of Cyberbulling Detection on Social Media Big-Data"MayJune 2017.

[16] A BERT-Based Transfer Learning Approach for Hate Speech Detection in Online Social Media Marzieh Mozafari, Reza Farahbakhsh and No¨el Crespi.