JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

DETECTION OF CYBERBULLYING ON SOCIAL MEDIA USING MACHINE LEARNING

[1]Savitha K, [2]Nandini K, [3]Dr.Chandra Sekar S

[1] PG Student, [2]Assistant Professor, [3]Professor
[1] Department of Computer science Engineering,
[1] P.S.V College of Engineering and Technology, krishnagiri, Tamil Nadu, India

Abstract: Cyber bullying detection is solved in this project as a binary classification problem where we are detecting two major form of Cyber bullying: hate speech on twitter and personal attacks on Wikipedia and classifying them as Cyber bullying or not. The system uses SVM (Support Vector Machine) for hate speech and Random Forest Classifier for personal attacks. It instead of simply looking for patterns, it goes beyond what's happening in the past to predict future outcomes based on the prediction existing data. The system produces the output set labeled as either offensive or non offensive. SVM aims to minimize an error by generating optimal hyper plane in an iterative manner. The real step towards of a Machine learning model is collecting data, Data preparation, model selection, Feature extraction and Analyze and prediction. Finally saving the trained model module is implemented. This system shows us more accuracy for detection Cyber bullying content. It also helps people from the attacks of social media bullies.

Keywords: Cyber bullying, Natural Language Processing, Machine Learning, SVM, Random Forest Classifier.

I. INTRODUCTION

Now more than ever technology has become an integral part of our life.With the evolution of the internet. Social media is trending these days. But as all the other things misusers will pop out sometimes late sometime early but there will be for sure. Now Cyber bullying is common these days. Sites for social networking are excellent tools for communication within individuals. Use of social networking has become widespread over the years, though, in general people find immoral and unethical ways of negative stuff. We see this happening between teens or sometimes between young adults. One of the negative stuffs they do is bullying each other over the internet. In online environment we cannot easily said that whether someone is saying something just for fun or there may be other intention of him. Often, with just a joke, "or don't take it so seriously," they'll laugh it off. Cyberbullying is the use of technology to harass, threaten, embarrass, or target another person. Often this internet fight results into real life threats for some individual. Some people have turned to suicide. It is necessary to stop such activities at the beginning. Any actions could be taken to avoid this for example if an individual's tweet/post is found offensive then maybe his/her account can be terminated or suspended for a particular period.

So, what is cyber bullying?? Cyber bullying is harassment, threatening, embarrassing or targeting someone for the purpose of having fun or even by well-planned means.

Researches on Cyber bullying Incidents show that 11.4% of 720 young people surveyed in the NCT DELHI were victims of cyber bullying in a 2018 survey by Child Right and You, an NGO in India, and almost half of them did not even mention it to their teachers, parents or guardians. 22.8% aged 13-18 who used the internet for around 3 hours a day were vulnerable to Cyber bullying while 28% of people who use internet more than 4 hours a day were victims. There are so many other reports suggested us that the impact of Cyber bullying is affecting badly the peoples and children between age of 13 to 20 face so many difficulties in terms of health, mental fitness and their decision making capability in any work. Researchers suggest that every country should have to take this matter seriously and try to find solution. In 2016 an incident called Blue Whale Challenge led to lots of child suicides in Russia and other countries. It was a game that spread over different social networks and it was a relationship between an administrator and a participant. For fifty days certain tasks are given to participants. Initially they are easy like waking up at 4:30 AM or watching a horror movie. But later they escalated to self harm which let to suicides. The administrators were found later to be children between ages 12-14.

II. LITERATURE SURVEY

H. Ting, et al. [1] proposes an approach based on social networks analysis and data mining for cyber bullying detection. In the approach, there are three main techniques for cyber bullying discovery will be studied, including keyword matching technique opinion mining and social network analysis. In addition to the approach, we will also discuss the experimental design for the evaluation of the performance.

P. Galán-García, et al. [2] propose an alter-ego with no relation to the actual user, creates a situation in which no one can certify the match between a profile and a real person. This problem generates situations, repeated daily, in which users with fake accounts, or at least not related to their real identity, publish news, reviews or multimedia material trying to discredit or attack other people who may or may not be aware of the attack. These acts can have great impact on the affected victims' environment generating situations in which virtual attacks escalate into fatal consequences in real life. In this paper, we present a methodology to detect and associate fake profiles on Twitter social network which are employed for defamatory activities to a real profile within the same network by analysing the content of comments generated by both profiles. Accompanying this approach we also present a successful real life use case in which this methodology was applied to detect and stop a cyberbullying situation.

A. Mangaonkar, et al.[3] aims at analyzing Cyberbullying content based on English tweets on one such social network that is Twitter. We analyzed tweets based on textual analysis and performed classification also. Through this they concluded our findings and future scope of work for detection of Cyberbullying on more complex data.

R. Zhao, et al[4] propose a representation learning framework specific to cyberbullying detection. Based on word embeddings, we expand a list of pre-defined insulting words and assign different weights to obtain bullying features, which are then concatenated with Bag-of-Words and latent semantic features to form the final representation before feeding them into a linear SVM classifier. Experimental study on a twitter dataset is conducted, and our method is compared with several baseline text representation learning models and cyberbullying detection methods. The superior performance achieved by our method has been observed in this study.

V. Banerjee, et al, [5] proposed a novel cyberbullying detection method dependent on deep neural network. Convolution Neural Network is utilized for the better outcomes when contrasted with the current systems.

K. Reynolds et al, [6] Through machine learning, we can detect language patterns used by bullies and their victims, and develop rules to automatically detect cyber bullying content. The data we used for our project was collected from the website Formspring.me, a question-and-answer formatted website that contains a high percentage of bullying content. The data was labeled using a web service, Amazon's Mechanical Turk. We used the labeled data, in conjunction with machine learning techniques provided by the Weka tool kit, to train a computer to recognize bullying content. Both a C4.5 decision tree learner and an instance-based learner were able to identify the true positives with 78.5% accuracy.

J. Yadav et al, [7] proposed a new approach is proposed to cyberbullying detection in social media platforms by using the novel pre-trained BERT model with a single linear neural network layer on top as a classifier, which improves over the existing results. The model is trained and evaluated on two social media datasets of which one dataset is small size and the second dataset is relatively larger size.

M. Dadvar et al, [8] investigate the findings of a recent literature in this regard. We successfully reproduced the findings of this literature and validated their findings using the same datasets, namely Wikipedia, Twitter, and Formspring, used by the authors. Then we expanded our work by applying the developed methods on a new YouTube dataset (~54k posts by ~4k users) and investigated the performance of the models in new social media platforms. We also transferred and evaluated the performance of the models trained on one platform to another platform. Our findings show that the deep learning based models outperform the machine learning models previously applied to the same YouTube dataset. We believe that the deep learning based models can also benefit from integrating other sources of information and looking into the impact of profile information of the users in social networks.

S. Agrawal et al,[9] showed that deep learning based models can overcome all three bottlenecks. Knowledge learned by these models on one dataset can be transferred to other datasets. We performed extensive experiments using three real-world datasets: Formspring (12k posts), Twitter (16k posts), and Wikipedia(100k posts). Our experiments provide several useful insights about cyberbullying detection. To the best of our knowledge, this is the first work that systematically analyzes cyberbullying detection on various topics across multiple SMPs using deep learning based models and transfer learning.

Y. N. Silva et al,[10] describe the challenges associated with building a computer model for cyberbullying identification, presents key results from psychology research that can be used to inform such a model, introduces a holistic model and mobile app design for cyberbullying identification, presents a novel evaluation framework for assessing the effectiveness of the identification model, and highlights crucial areas of future work. Importantly, the proposed model—which can be applied to other social networking sites—is the first that we know of to bridge computer science and psychology to address this timely problem.

III. PROPOSED WORK

In this Paper, we divided the modules into the following.

- Data Collection
- Dataset
- Data Preparation
- Model Selection
- Analyze and Prediction
- Accuracy on test set
- Saving the Trained Model

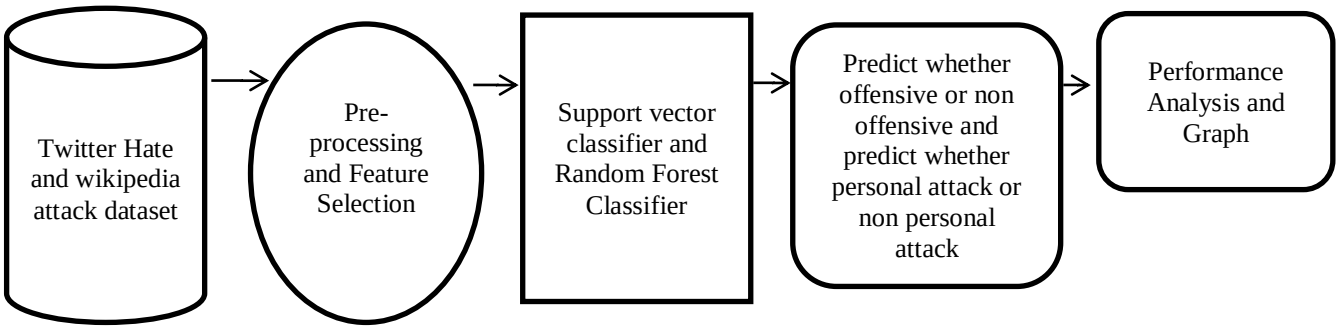

Fig.1.system model based on SVM and Random Forest Classifier

MODULES DESCSRIPTION FOR TWITTER HATE DETECTION

Data Collection:

This is the first real step towards the real development of a machine learning model, collecting data. This is a critical step that will cascade in how good the model will be, the more and better data that we get, the better our model will perform. There are several techniques to collect the data, like web scraping, manual interventions and etc.Detection of Cyberbullying on Social Media Using Machine learning We given the Twitter Hate data set in the project folder.

Dataset:

The dataset consists of 31962 individual data. There are 3 columns in the dataset, which are described below

1. Id: unique id
2. Labels :
   1: offensive
   0: non offensive
3. Tweet : comment

Data Preparation:

We will transform the data. by getting rid of missing data and removing some columns. First we will create a list of column names that we want to keep or retain.

Next we drop or remove all columns except for the columns that we want to retain.

Finally we drop or remove the rows that have missing values from the data set.

Steps to follow:

Removing extra symbols
Removing punctuations
Removing the Stopwords
Stemming
Tokenization
Feature extractions
TF-IDF vectorizer
Counter vectorizer with TF-IDF transformer

Model Selection:

We used SVC algorithms.

Support Vector Machine algorithm:

SVM offers very high accuracy compared to other classifiers such as logistic regression, and decision trees. It is known for its kernel trick to handle nonlinear input spaces. It is used in a variety of applications such as face detection, intrusion detection, classification of emails, news articles and web pages, classification of genes, and handwriting recognition.

In this tutorial, you will be using scikit-learn in Python. If you would like to learn more about this Python package, I recommend you take a look at our Supervised Learning with scikit-learn course.

SVM is an exciting algorithm and the concepts are relatively simple. The classifier separates data points using a hyperplane with the largest amount of margin. That's why an SVM classifier is also known as a discriminative classifier. SVM finds an optimal hyperplane which helps in classifying new data points.

SUPPORT VECTOR MACHINES

Generally, Support Vector Machines is considered to be a classification approach, it but can be employed in both types of classification and regression problems. It can easily handle multiple continuous and categorical variables. SVM constructs a hyperplane in multidimensional space to separate different classes. SVM generates optimal hyperplane in an iterative manner, which is used to minimize an error. The core idea of SVM is to find a maximum marginal hyperplane(MMH) that best divides the dataset into classes.

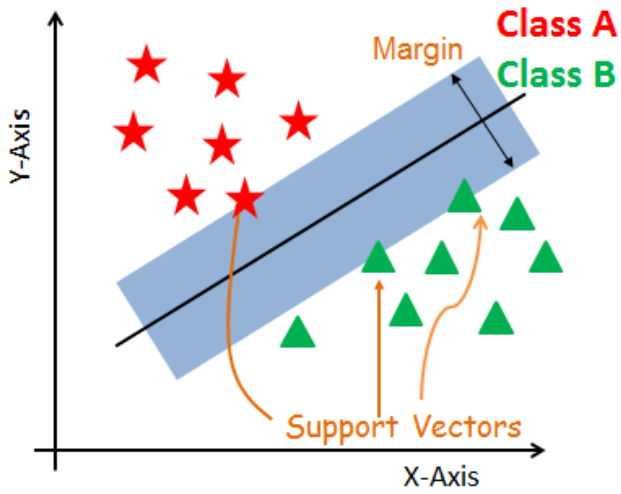

Fig.2.Support Vector Machine Approach

Support Vectors

Support vectors are the data points, which are closest to the hyperplane. These points will define the separating line better by calculating margins. These points are more relevant to the construction of the classifier.

Hyperplane

A hyperplane is a decision plane which separates between a set of objects having different class memberships.

Margin

A margin is a gap between the two lines on the closest class points. This is calculated as the perpendicular distance from the line to support vectors or closest points. If the margin is larger in between the classes, then it is considered a good margin, a smaller margin is a bad margin.

How does SVM work?

The main objective is to segregate the given dataset in the best possible way. The distance between the either nearest points is known as the margin. The objective is to select a hyperplane with the maximum possible margin between support vectors in the given dataset. SVM searches for the maximum marginal hyperplane in the following steps:

Generate hyperplanes which segregates the classes in the best way. Left-hand side figure showing three hyperplanes black, blue and orange. Here, the blue and orange have higher classification error, but the black is separating the two classes correctly.

Select the right hyperplane with the maximum segregation from the either nearest data points as shown in the right-hand side figure.

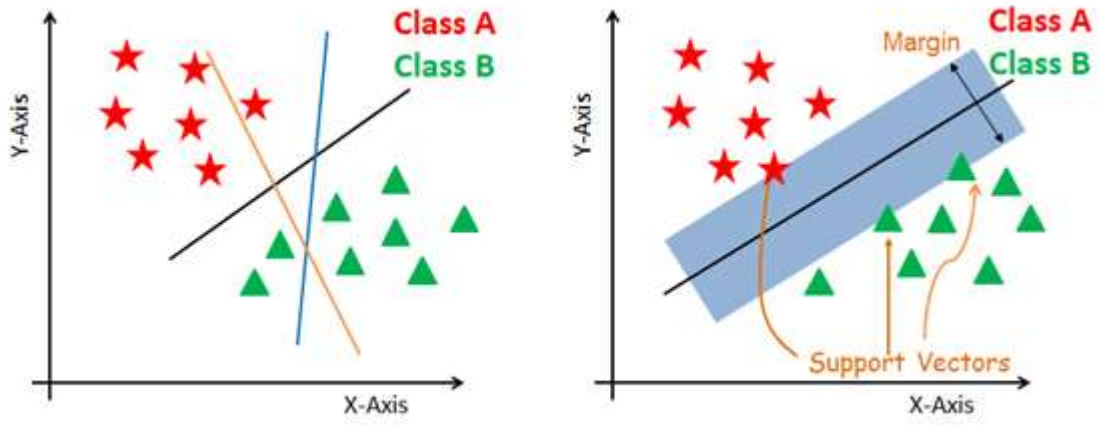

Fig.3. SVM Hyper planes

Dealing with non-linear and inseparable planes

Some problems can't be solved using linear hyperplane, as shown in the figure below (left-hand side).

In such situation, SVM uses a kernel trick to transform the input space to a higher dimensional space as shown on the right. The data points are plotted on the x-axis and z-axis (Z is the squared sum of both x and y: z=x^2=y^2). Now you can easily segregate these points using linear separation.

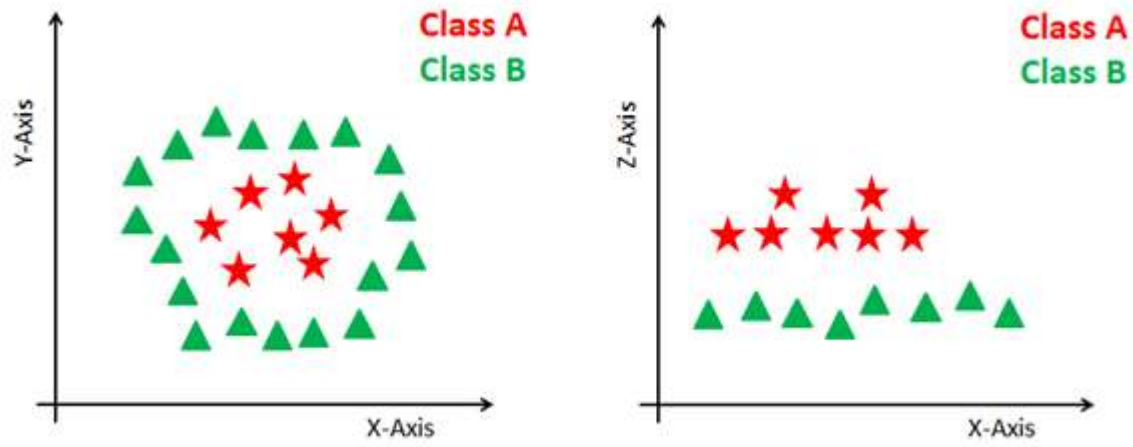

Fig.4. SVM Non linear Hyper planes

SVM Kernels

The SVM algorithm is implemented in practice using a kernel. A kernel transforms an input data space into the required form. SVM uses a technique called the kernel trick. Here, the kernel takes a low-dimensional input space and transforms it into a higher dimensional space. In other words, you can say that it converts nonseparable problem to separable problems by adding more dimension to it. It is most useful in non-linear separation problem. Kernel trick helps you to build a more accurate classifier.

Linear Kernel A linear kernel can be used as normal dot product any two given observations. The product between two vectors is the sum of the multiplication of each pair of input values.

```
K(x, xi) = sum(x * xi)
```

Polynomial Kernel A polynomial kernel is a more generalized form of the linear kernel. The polynomial kernel can distinguish curved or nonlinear input space.

```
K(x,xi) = 1 + sum(x * xi)^d
```

Where d is the degree of the polynomial. d=1 is similar to the linear transformation. The degree needs to be manually specified in the learning algorithm.

Radial Basis Function Kernel The Radial basis function kernel is a popular kernel function commonly used in support vector machine classification. RBF can map an input space in infinite dimensional space.

```
K(x,xi) = exp(-gamma * sum((x - xi^2))
```

Here gamma is a parameter, which ranges from 0 to 1. A higher value of gamma will perfectly fit the training dataset, which causes over-fitting. Gamma=0.1 is considered to be a good default value. The value of gamma needs to be manually specified in the learning algorithm.

MODULES DESCRIPTION FOR WIKIPEDIA ATTACK

Data Collection:

This is the first real step towards the real development of a machine learning model, collecting data. This is a critical step that will cascade in how good the model will be, the more and better data that we get, the better our model will perform.

There are several techniques to collect the data, like web scraping, manual interventions and etc.

Detection of Cyberbullying on Social Media Using Machine learning

We given the wikipedia attack data set in the project folder

Dataset:

The dataset consists of 115864 individual data. There are 4 columns in the dataset, which are described below

1. Review Id: unique id
2. comment : comment about wikipedia titles
3. year : year of comment
4. attack : Personal attack or non personal attack

Data Preparation:

We will transform the data. by getting rid of missing data and removing some columns. First we will create a list of column names that we want to keep or retain.

Next we drop or remove all columns except for the columns that we want to retain.

Finally we drop or remove the rows that have missing values from the data set.

Steps to follow:

- Removing extra symbols
- Removing punctuations
- Removing the Stopwords
- Stemming
- Tokenization
- Feature extractions
- TF-IDF vectorize
- Counter vectorizer with TF-IDF transformer

Model Selection:

We used Random Forest Classifier algorithms

Random Forests Classifiers:

Random forests is a supervised learning algorithm. It can be used both for classification and regression. It is also the most flexible and easy to use algorithm. A forest is comprised of trees. It is said that the more trees it has, the more robust a forest is. Random forests creates decision trees on randomly selected data samples, gets prediction from each tree and selects the best solution by means of voting. It also provides a pretty good indicator of the feature importance.

The Random Forests Algorithm

Let's understand the algorithm in layman's terms. Suppose you want to go on a trip and you would like to travel to a place which you will enjoy.

So what do you do to find a place that you will like? You can search online, read reviews on travel blogs and portals, or you can also ask your friends.

Let's suppose you have decided to ask your friends, and talked with them about their past travel experience to various places. You will get some recommendations from every friend. Now you have to make a list of those recommended places. Then, you ask them to vote (or select one best place for the trip) from the list of recommended places you made. The place with the highest number of votes will be your final choice for the trip.

In the above decision process, there are two parts. First, asking your friends about their individual travel experience and getting one recommendation out of multiple places they have visited. This part is like using the decision tree algorithm. Here, each friend makes a selection of the places he or she has visited so far.

The second part, after collecting all the recommendations, is the voting procedure for selecting the best place in the list of recommendations. This whole process of getting recommendations from friends and voting on them to find the best place is known as the random forests algorithm.

It technically is an ensemble method (based on the divide-and-conquer approach) of decision trees generated on a randomly split dataset. This collection of decision tree classifiers is also known as the forest. The individual decision trees are generated using an attribute selection indicator such as information gain, gain ratio, and Gini index for each attribute. Each tree depends on an independent random sample. In a classification problem, each tree votes and the most popular class is chosen as the final result. In the case of regression, the average of all the tree outputs is considered as the final result. It is simpler and more powerful compared to the other non-linear classification algorithms.

How does the algorithm work?

It works in four steps:

Select random samples from a given dataset.

Construct a decision tree for each sample and get a prediction result from each decision tree.

Perform a vote for each predicted result.

Select the prediction result with the most votes as the final prediction.

IV. RESULT ANALYSIS

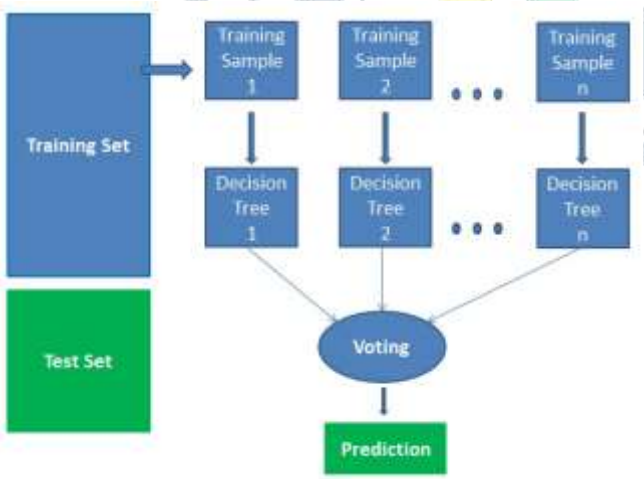

Fig.5. Prediction Result Analysis

Fig.6.Performance Analysis.

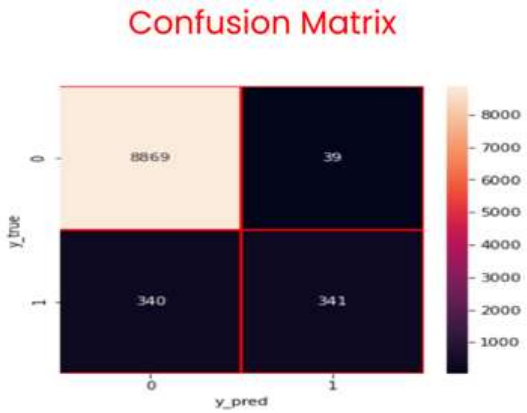

Fig.7. Confusion Matrix

```
# transforming input
tfidf_test = tfid_wikis.transform(input_data)
# predicting the input
y_pred = wikis.predict(tfidf_test)
if y_pred[0] == 1:
  label='Personal attack'
elif y_pred[0] == 0:
   label='Non Personal attack'
return render_template('predictions.html', prediction_text=label)
```

Fig.8. Implementation Results

Analyze and Prediction:
In the actual dataset, we chose only 2 features :

1 Text: the tweets
2 Labels :
1: offensive
0: non offensive

Accuracy on test set:
We got an accuracy of 96.02% on test set.

V. CONCLUSION

Cyber bullying across internet is dangerous and leads to mishappenings like suicides, depression etc and therefore there is a need to control its spread. Therefore cyber bullying detection is vital on social media platforms. With avaibility of more data and better classified user information for various other forms of cyber attacks Cyberbullying detection can be used on social media websites to ban users trying to take part in such activity In this paper we proposed an architecture for detection of cyber bullying to combat the situation. We discussed the architecture for two types of data: Hate speech Data on Twitter and Personal attacks on Wikipedia. For Hate speech Natural Language Processing techniques proved effective with accuracies of over 90 percent using basic Machine learning algorithms because tweets containing Hate speech consisted of profanity which made it easily detectable. Due to this it gives better results with BoW and Tf-Idf models rather than Word2Vec models However, Personal attacks were difficult to detect through the same model because the comments generally did not use any common sentiment that could be learned however the three feature selection methods performed similarly.Word2Vec models that use context of features proved effective in both datasets giving similar results in comparatively less features when combined with Multi Layered Perceptrons.

SCOPE FOR FEATURE ENHANCEMENT

Multi Layered Perceptron's are the Artificial Neural Networks containing at least 3 layers: one input, one output and at least one hidden layer. Each node has a activation value calculated using an activation function in a process called forward propagation and back propagation is used to train the weight used in the neural networks. It is generally used when data is linearly non separable. Activation functions used can be relu or sigmoid. Sigmoid function is similar to the tan function which is hyperbolic in nature between -1 and 1. Relu is defined as f(x)=max(0,x). Multi Layered Perceptron's can be created and trained using Keras Framework.

The classifiers were loaded through sklearn library except the Multi Layer Perceptrons which were made in Keras. Two MLPs were used: one for Bag of Words and tfidf for 10000 feature input and other for 400 feature input of Word2vec. The architectures of both neural networks . The Classifiers used are Linear SVM (SVC), Random Forest Classifier (RF), Logistic Regression (LR) and Multi Layered Perceptron (MLP). Tables 1 and 2 show results for Twitter and Wikipedia dataset respectively. The Twitter dataset which contained tweets related to Hate speech show F- measures above 0.9 for all three feature selection methods. The values for Word2Vec model are a bit less but are ideal considering it used 400 features instead of other methods using

10000. TF-IDF method combined with Linear SVM gives best recall and F-measure. For the Wikipedia dataset which contained comments with Personal attacks it shows F-measures only around 0.8 for all models. The TF-IDF with Linear SVM still get the best Fmeasure but Word2Vec with Multi Layered Perceptron
gives better recall.

REFERENCES

[1] I. H. Ting, W. S. Liou, D. Liberona, S. L. Wang, and G. M. T. Bermudez, "Towards the detection of cyberbullying based on social network mining techniques," in Proceedings of 4th International Conference on Behavioral, Economic, and Socio Cultural Computing, BESC 2017, 2017, vol. 2018-January, doi: 10.1109/BESC.2017.8256403.

[2] P. Galán-García, J. G. de la Puerta, C. L. Gómez, I. Santos, and P. G. Bringas, "Supervised machine learning for the detection of troll profiles in twitter social network: Application to a real case of cyberbullying," 2014, doi: 10.1007/978-3-319-01854-6_43.

[3] A. Mangaonkar, A. Hayrapetian, and R. Raje, "Collaborative detection of cyberbullying behavior in Twitter data," 2015, doi: 10.1109/EIT.2015.7293405.

[4] R. Zhao, A. Zhou, and K. Mao, "Automatic detection of cyberbullying on social networks based on bullying features," 2016, doi: 10.1145/2833312.2849567.

[5] V. Banerjee, J. Telavane, P. Gaikwad, and P. Vartak, "Detection of Cyberbullying Using Deep Neural Network," 2019, doi: 10.1109/ICACCS.2019.8728378.

[6] K. Reynolds, A. Kontostathis, and L. Edwards, "Using machine learning to detect cyberbullying," 2011, doi: 10.1109/ICMLA.2011.152.

[7] J. Yadav, D. Kumar, and D. Chauhan, "Cyberbullying Detection using Pre-Trained BERT Model," 2020, doi: 10.1109/ICESC48915.2020.9155700.

[8] M. Dadvar and K. Eckert, "Cyberbullying Detection in Social Networks Using Deep Learning Based Models; A Reproducibility Study," arXiv. 2018.

[9] S. Agrawal and A. Awekar, "Deep learning for detecting cyberbullying across multiple social media platforms," arXiv. 2018.

[10] Y. N. Silva, C. Rich, and D. Hall, "BullyBlocker: Towards the identification of cyberbullying in social networking sites," 2016, doi: 10.1109/ASONAM.2016.7752420.

[11] Z. Waseem and D. Hovy, "Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter," 2016, doi: 10.18653/v1/n16-2013.

[12] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," 2017.

[13] E. Wulczyn, N. Thain, and L. Dixon, "Ex machina: Personal attacks seen at scale," 2017, doi: 10.1145/3038912.3052591.

[14] A. Yadav and D. K. Vishwakarma, "Sentiment analysis using deep learning architectures: a review," Artif. Intell. Rev., vol. 53, no. 6, 2020, doi: 10.1007/s10462-019-09794-5.

[15] T. MIKOLOV, K. CHEN, G. CORRADO, AND J. DEAN, "EFFICIENT ESTIMATION OF WORD REPRESENTATIONS IN VECTOR SPACE," 2013.