International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

COPY RIGHT

ELSEVIER
SSRN

IJIEMR Transactions, online available on 25th Jun 2022. Link
:http://www.ijiemr.org/downloads.php?vol=Volume-11&issue= Spl Issue 05

DOI: 10.48047/IJIEMR/V11/SPL ISSUE 05/17

Title **DETECTING ABUSIVE AND INSULTING COMMENTS ON SOCIAL MEDIA**

Volume 11, SPL ISSUE 05, Pages: 111-116

Paper Authors
Dr.BV Ramakrishna , M.N.L Shivani , V.Hema Ratna , P.Sailaja, T. Meghana

USE THIS BARCODE TO ACCESS YOUR ONLINE PAPER

To Secure Your Paper As Per UGC Guidelines We Are Providing A Electronic Bar Code

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

DETECTING ABUSIVE AND INSULTING COMMENTS ON SOCIAL MEDIA

Dr.BV Ramakrishna[1] , M.N.L Shivani[2] , V.Hema Ratna[3] , P.Sailaja[4], T. Meghana[5].

[1]Professor, Dept. of CSE, [2]18ME1A0562, [3]18ME1A05B7, [4]18ME1A0577 , [5]18ME1A05B0
Ramachandra College of Engineering, A.P, India
shivanichocolaty21@gmail.com , vallabhaneni.731@gmail.com, parepallisailaja577@gmail.com, meghanathota1011@gmail.com

Abstract
Social media is becoming increasingly exposed to issues of harmful behaviour, such as private assaults and cyberbullying. Manually checking which comments need to be blocked is inconvenient and time-consuming. As a result, automating the process of recognizing and blocking abusive comments will not only save time but also ensure user safety. This paper focuses on employing machine learning and deep learning techniques to solve this challenge. The model is trained using the Twitter dataset. There are two types of comments: abusive and non-abusive. Our project is used to identify and regulate harmful social media remarks.

Keywords: Deep Learning, ANN, RF, Abusive, Non-Abusive.

Introduction

In the recent decade, there has been a significant increase in research on social media safety and security. Detecting and blocking the use of various forms of abusive language in blogs, microblogs, and social networks is a particularly important facet of this effort. Several recent studies have been published on this topic, on detecting cyberbullying, hate speech detection and racism detection in user generated content. Artificial neural networks are algorithms inspired by the structure and function of the brain. Deep learning is an area of machine learning dealing with algorithms inspired by the structure and function of the brain. Deep learning algorithms use models with several processing layers to learn input and translate it into numerous degrees of abstraction. Unstructured data can be learned using deep learning algorithms. The Artificial neural network (ANN) is a type of deep neural network that is powered by the biological processes of neurons. There are numerous hidden layers, as well as an input and output layer.

The existence of profane content does not automatically imply hate speech. General profanity is not always directed at a specific person and can be used for emphasis or artistic purposes.

Hate speech, on the other hand, can be used to disparage or threaten an individual or a group of people without using profanity. On this standard dataset, the major goal of this paper is to construct a lexical baseline for differentiating between hate speech and profanity. The corpus we're using here gives us an interesting opportunity to see how well a system can distinguish hate speech from other profane stuff.

Online hate, described as abusive language, aggression, cyberbullying, hatefulness, insults, personal attacks, provocation, racism, sexism, threats, or toxicity, has been identified as a major threat on online social media platforms. Pew Research Center reports that among 4248 adults in the United States, 41% have personally experienced harassing behavior online, whereas 66% witnessed harassment directed towards others. Among other sorts of harassment, 22 percent of adults have experienced harsh name-calling, intended embarrassment (22 percent), physical threats (10 percent), and sexual harassment (6 percent). The most common venues for such toxic behaviour are social media platforms. Despite the fact that companies frequently provide means to report offensive or bigoted content, just 17% of all adults have highlighted harassing dialogue, and only 12% of adults have reported someone for such behaviour. Manual procedures such as flagging are neither effective nor scalable, and they risk discrimination due to subjective human annotator judgments. Machine learning models to automatically detect online hate are gaining popularity and bringing academics from several domains together since an automated technique can be faster than human annotation. Despite the fact that hate has been identified as a problem on multiple online social media platforms, such as Reddit, YouTube, Wikipedia, Twitter, and others, there has been little development and testing of models using data from multiple social media platforms, aside from a few exploratory studies. Rather, research tend to concentrate on a single platform. This singular concentration on a single platform is problematic since there is no guarantee that the models developed by academics will generalize well across platforms. It is fair to assume that information gleaned from multiple training sets and circumstances could be useful in constructing a universal hate classifier. The mono-platform approach is particularly frustrating since the lack of a general hate classifier forces researchers and practitioners to "reinvent the wheel," which means that a new classifier must be constructed each time online hate research is conducted on a specific social media platform. This not only results in repetitive mental strain, but it also creates "barriers to entrance" for researchers who lack model creation skills yet are interested in interpretative OHR. Furthermore, the lack of universal classifiers makes it difficult to compare results between studies and social media platforms. To summarize, the fragmentation of models and feature representations makes hate detection across platforms and settings unnecessarily difficult. To address these concerns, we are working on the creation of a cross-platform online hate classifier. Our approach, which employs advanced linguistic features, is effective at recognizing nasty comments across different social media networks. While we do not claim to have created the universal classifier that addresses all problems in online hate detection, our findings show that this area of research has promise for the greater community and can be further developed.

Related Work

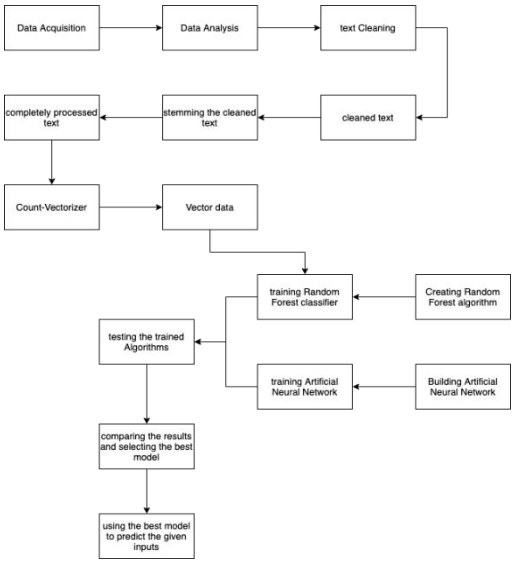
In recent years, various papers on computational approaches for detecting abusive language have been published. Xu et al. (2012), for example, used sentiment analysis to detect bullying in tweets and Latent Dirichlet Allocation (LDA) topic models (Blei et al., 2003) to find significant subjects in these texts. A lot of

studies on hate speech detection have been published. To the best of our knowledge, they all rely on binary classification (for example, hate speech vs. non-hate speech). Kwok and Wang (2013), Djuric et al. (2015), and Nobata et al. (2015) are examples of similar investigations (2016).

A lot of studies on hate speech detection have been published. To the best of our knowledge, they all rely on binary classification (for example, hate speech vs. non-hate speech). Kwok and Wang (2013), Djuric et al. (2015), and Nobata et al. (2015) are examples of similar investigations (2016).

Architecture of Proposed System

The tweets data for classifying insults and abusive content has a very interesting set of features covered by a highly imbalanced data, and overall, the data is highly practical and very difficult to predict the output because the majority of the tweets match with normal, non-insulting tweets with only a minor difference.



The technique of identifying abusive and insult tweets is depicted in the diagram above. The first step is to extract data from a ".csv" file and convert it to a python-friendly object called a "DataFrame" in the pandas package. It's a Data type that works with tabular data, and the tweets are the values of rows in a pandas DataFrame with two columns. The most delicate procedure is to clean the text by removing extraneous symbols, numbers, irrelevant excess spaces, and, last but not least, stopwords. Only one stage remains in the cleaning process, which aids in dimensionality reduction.

Stemming is the next phase. The count-vectorization process uses the frequency of the words in the phrase to turn text data into integer type data. We use this to convert the data, which we then use to train and assess other types of algorithms, such as "Random Forest" and "Artificial Neural Networks." While each algorithm's training process is running on the processing line, when one algorithm's training is complete, the testing process is completed and the results are placed in the specified data structure.

Methods

Random Forest

Random forest is a strong machine learning method that may be used for a range of tasks such as regression and classification, each of which generates its own predictions. The random forest model combines the estimators' predictions to get a more precise prediction**.**

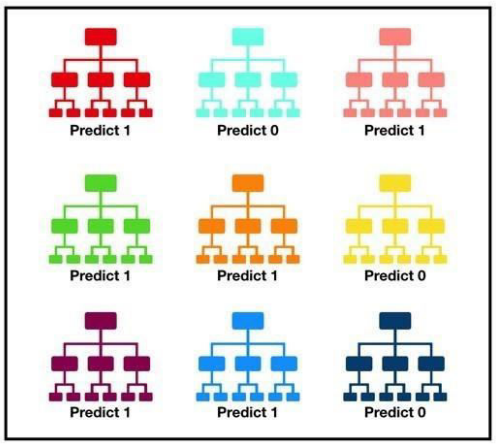
The Random Forest Classifier

The random forest, as the name suggests, is made up of a huge number of individual decision trees that work together as an ensemble. The random forest's individual trees spit out class predictions, and the class with the highest votes becomes our model's prediction (see figure below).

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

Random forest is based on a simple yet powerful concept: the wisdom of crowds. The following are the reasons why the random forest model performs so well in data science:

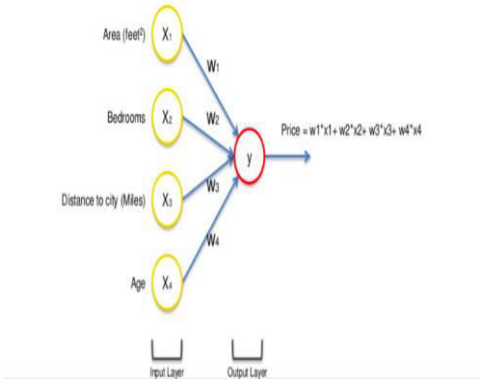
Any of the individual constituent models will outperform a large number of reasonably uncorrelated models (trees) working as a committee.



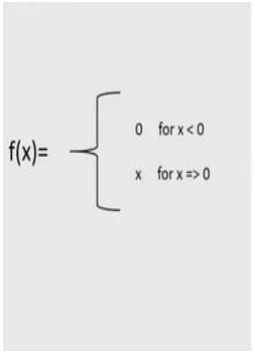
Tally: Six 1s and Three 0s
Prediction: 1

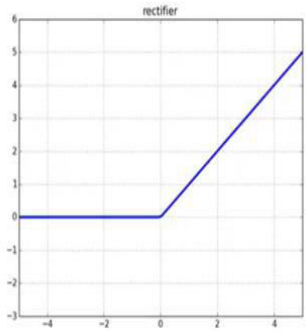
Artificial Neural Networks

Artificial Neural Networks is a computational model that mimics the way nerve cells work in the human brain. Artificial Neural Networks use learning algorithms that can independently make adjustments or learn, in a sense -as they receive new input.



Rectified Linear Units — (ReLu)
ReLu is the most used activation function in CNN and ANN which ranges from zero to infinity.[0,∞).





If x is positive, it returns 'x'; otherwise, it returns 0. It appears to have the same linear function difficulty as it is linear in the positive axis. Relu is non-linear in nature, and a ReLu combination is non-linear as well. In reality, it's a good approximator, and any function may be approximated using Relu and other functions.
ReLu outperforms hyperbolic tangent function by 6 times**.**

Implementation

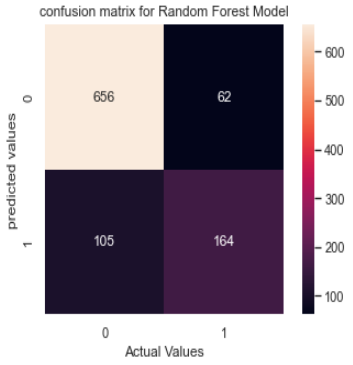
The methodology provides a description of the steps involved in the system's operation. The input, output, and other components of the system are described in methodology.

In this project, we are using two models as earlier mentioned.

One is random forest classifier and other one is artificial neural networks.

The confusion matrix for random forest classifier is shown below.

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org



The model for ANN is shown Below

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 flatten (Flatten)           (None, 14313)             0

 dense (Dense)               (None, 128)               1832192

 dropout (Dropout)           (None, 128)               0

 dense_1 (Dense)             (None, 2)                 258

=================================================================
Total params: 1,832,450
Trainable params: 1,832,450
Non-trainable params: 0
_________________________________________________________________
```

1. The user creates an account on the system.
2. The user enters his or her email address and password.
3. The user is verified by the system. The user is logged in if the validation is successful; otherwise, an error message is displayed.
4. The user types his or her comment on the screen and presses the submit button.
5. The training dataset is used to train the system, and features are extracted.
6. The user's comment is used as testing data by the system.
7. The user comment is processed beforehand.
8. The system takes the processed comment and extract the features that are necessary.
9. Once the system predicts whether the comment is abusive or not it displays the output.
10. If the comment is abusive the respective authorities will take action accordingly.
11. The User logs out of the system.

Result

This section delves into the numerous test scenarios that are used to put the system through its paces. The model was trained using the Twitter dataset. It is divided into three categories: positive, negative, and abusive. The model was tested using user input. Below are the test cases and findings.

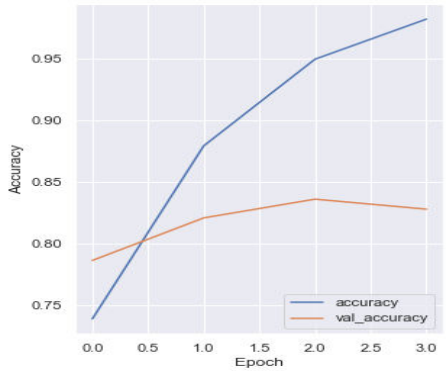


Fig 1: Graph of Accuracy V/S Epoch

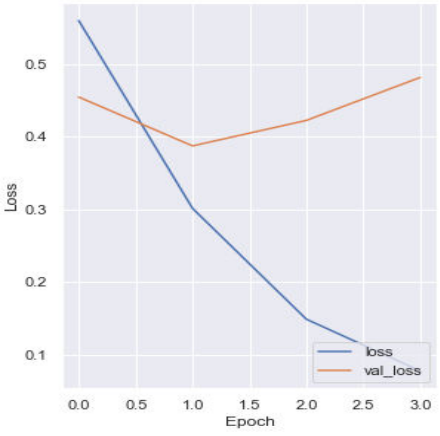


Fig 2: Graph Of Loss V/S Epoch

Conclusion

The initiative is being created with the goal of offering social media security. This goal is achieved by employing a deep learning technology that detects abusive remarks on social media automatically. The algorithm then moderates the discovered abusive comments by immediately blocking them. The project creates a comprehensive system that lowers online abuse and harassment without the need for human intervention.

With some training delay, the proposed approach provides improved accuracy. Future work will concentrate on detecting the many classes contained in a single comment, displaying only those sections of the comment that are judged to be non-abusive, and alerting the user to any abusive content.

References

1.Sidharth mehra,Detecting of offensive language in social media posts on may 2020

2. Peddada Anitha,Detecting Abusive Comments On Social Media. poddatare sushmitha teddy,Reperthi tarun,charity nagaraju,on may- 2022 p(129 to 135)

3. Simart kaur,sarbjeet Singh,sakshi kaushal Abusive content detection in online user- generated data: sct on 2021

4. D.R. Janardhana, Asha B. Shetty, Madhura N. Hegde, Jayapadmini kanchan, and anjana Hegde, Abusive comments Classification in Social Media using Neural Network on May (2021).

5. Jayadev Bhaskaran, Amita kamath, Suvadip paul, Detecting Insults in Social Commentary on (2019).

6. S. H. Yadav, P.M. Manwatkar, An approach for Offensive text Detection and Preventing in social networks on (2015).

7. Ching Seh Wu, Unnathi Bhandary, Detection of Hate Speech in Videos Using Machine learning on (2020)

8. Marta Navarron Garcia and Isabel Segura Bedmar, Detecting Offensiveness in Social Networn Comments on (2021).

9. H.Chen, S. McKeever, S.J. Delany, Abusive text detection using neural network, Elsevier,(2017).

10.S.B.Shende, L.Deshpande, A computational framework for detecting offensive language with support vector machine in social communities, on (2017).

11. Jun-Ming Xu, Kwang-Sung Jun, Xiaojin Zhu, and Amy Bellmore. 2012. Learning from bullying traces in social media. In Proceedings of the 2012 conference of the North American chapter of the association for computational linguistics: Human language technologies. Association for Computational Linguistics, pages 656–666.

12. David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent Dirichlet Allocation. Journal of machine Learning research 3(Jan):993–1022.

13. Irene Kwok and Yuzhou Wang. 2013. Locate the hate: Detecting tweets against blacks. In Twenty-Seventh AAAI Conference on Artificial Intelligence.

14. Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati. 2015. Hate speech detection with comment embeddings. In Proceedings of the 24th International Conference on World Wide Web Companion. International World Wide Web Conferences Steering Committee, pages 29–30.

15. Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive Language Detection in Online User Content. In Proceedings of the 25th International Conference on World Wide Web. International World Wide Web Conferences Steering Committee, pages 145–153.