DETECTING ABUSIVE AND INSULTING COMMENTS ON SOCIAL MEDIA

Joshi Padma. N Associate Professor
CSE, Sreyas institute of engineering and Technology, Telangana, India
padmajoshi@sreyas.ac.in

Peddada Anvitha
B.Tech Student of CSE,
Sreyas institute of engineering and Technology, Telangana, India
anvithapeddada@gmail.com

Poddaturu Sushmitha Reddy
B.Tech Student of CSE ,
Sreyas institute of engineering and Technology, Telangana, India
psreddie@gmail.com

Raperthi Tarun Chandra Bhoopathi
B.Tech Student of CSE,
Sreyas institute of engineering and Technology, Telangana, India.
Tarunchandra001@gmail.com

Chevity Nagaraju
B.Tech Student of CSE,
Sreyas institute of engineering and technology,Telangana,India
nagaraju.chevity@gmail.com

Abstract

Now a days, Social media has become a very important mode of communication. Billions of people communicate through various social networking sites. People are using these sites as a medium to express their views and opinions. As the faces of the coin, there are both pros and cons by using this medium. One of the major problems is the insulting and the abusive comments that we get on the social media. Hence, the need for good quality automated abusive language classifiers becomes important. Our project "Detecting Abusive and Insulting Comments on Social Media" aims at designing a system for automatically detecting comments on social media as abusive or non-abusive using machine learning algorithms such as Random Forest classifier and Artificial Neural Network techniques (ANN).

Keywords— Abusive, Insulting, Comments, Social Media, Random forest classifier, ANN.

I. INTRODUCTION

These days most of the people are voicing out their opinions through social media. While most of the information is useful and bringing necessary changes in society while some of this information is leading to negative impacts in the society. Most of the data we deal with are grouped towards aspects like prejudice, sexism, hate speech, animosity and individual assaults. The comments may be positive or negative. The positive feedback motivates people and encourages them to overcome their fears. Whereas the negative comments not only hurt the feelings of individual but also demotivates them which triggers to many health-related issues of depression, anxiety etc. Hateful and abusive language and verbal aggression bring unnecessary disturbances in the society. Using today's advanced digital communication technologies, it has become very easy for people to make abusive comments without any fear. We live in a time where online communication has become the most important thing and it is used frequently, this situation makes it important to address the key problem (i.e., online hate speech). To overcome this problem, we designed an automated system utilizing a few Text Classification Algorithms for detecting such abusive and insulting comments on social media. We are using ANN (Artificial Neural Networks) and Random Forest Classification algorithm to address this problem. When we compare accuracy of both these algorithms, we can say that Random Forest Classifier works more efficiently as it gives more accuracy.

II. LITERATURE SURVEY

Comments and statements of hateful type seriously hamper a constructive private discussion or public debate[1]. In recent years, the topic has received an increasing amount of attention from multiple stakeholders [2]. Among these are social scientists who want to analyse this phenomenon and reasons for abusive online behaviour and politicians who realise that major parts of public debates and social discourse are carried out online [3].

In addition, we have seen that not only such online discussions but also the perception of concepts, politicians, elections and civil rights movements can be influenced using highly targeted social media marketing campaigns[4]. We live in a time in which online media, including online news and online communication, have an unprecedented level of social, political and also economic relevance[5]. This situation creates a plethora of challenges with regard to the key question how best to address the importance and relevance of online media and online content with technological means while at the same time not putting in place a centralised infrastructure that can be misused for the purpose of censorship or surveillance[6]. One challenge is to separate high-quality content from offensive, hateful, abusive or massively biased content[7]. While these tasks have been mostly in the realm of journalism, they are getting more and more transferred to the end user of online content, i.e., the analysis, curation and assessment of information is no longer carried out by professional news editors or journalists exclusively – the burden of fact checking is more and more left to the reader[8].

There are many algorithms which are specially designed for detection of abusive or insulting content with different software applications[9]. There is no clarity in their uses under different scenarios hence there is need of comparative study for discussion of such algorithms which can

solve many the problem of finding abusive and insulting content from data[10]. Older and existing algorithms need much amount of data to train the classifier as well as one of the major drawback is in what scenario which algorithm we can use there is no proper clarity[11]. In proposed work we designed system of deep learning algorithm which uses python software for implementing classification of abusing as well as insulting content[12]. Further these all algorithms are compared to understand the best suitable algorithms among all. This algorithms which are machine learning algorithms provides more accurate and reliable results than existing state of art techniques[13].

III. PROPOSED METHOD

With the level of advancements in technology, we can see that many libraries have been developed. So, with those libraries, we can create many applications and APIs which are user-friendly. In this we are developing a model which takes the data from the huge repositories containing the abusive and insulting comments collected over various social networking sites. We have many machine learning algorithms to detect the comments, but there is no proper clarity in their usuage under different scenarios. And the other major issue with the existing system is that it requires huge amount of training data. To overcome this we are using proper machine learning algorithms that makes our system more accurate and reliable.So this to happen we are making a comparative study of such algorithms which can solve the problem of finding abusive and insulting content from data.

A. PROPOSED ARCHITECTURE

The architecture diagram consists of data for classifying the insults and abusive comments. The data collection is the data acquisition step. The next step is the data analysis. The data analysis is used to analize the type of the data collected and to understand the modifications that are to be done to the data to make it machine understandable and to reduce the complexities. The data thus collected is highly unstructured data, highly practical and very difficult to predict the output because most part of the text matches with normal, non-insulting text with a very little or minute difference. So the text cleaning needs to be performed for removing the unnecessary symbols, numbers, useless usernames extra spaces, stop words and then dimensionality reduction by stemming is done. Count vectorizer converts the text data into integer type data using the frequency of the words in the sentence. The data thus obtained is the vector data. Thus we use this transformed data to train the models. The models we are using here are random forest classifier and artificial neural networks. The testing is thus performed to evaluate the accuracy. This is the comparitive study technique where we are comparing the accuracy of both the models and choosing the best model.

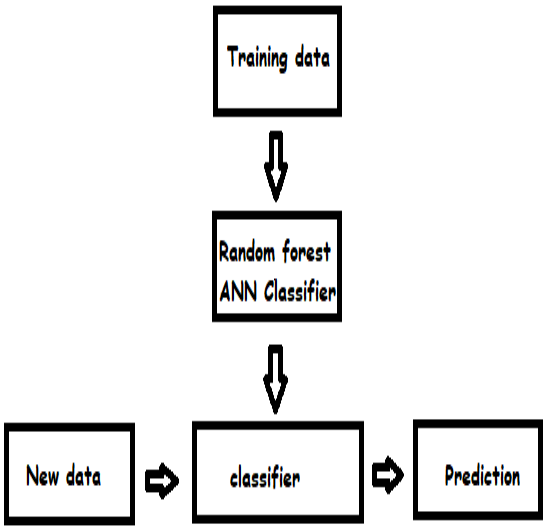

Fig 1: Basic Architecture

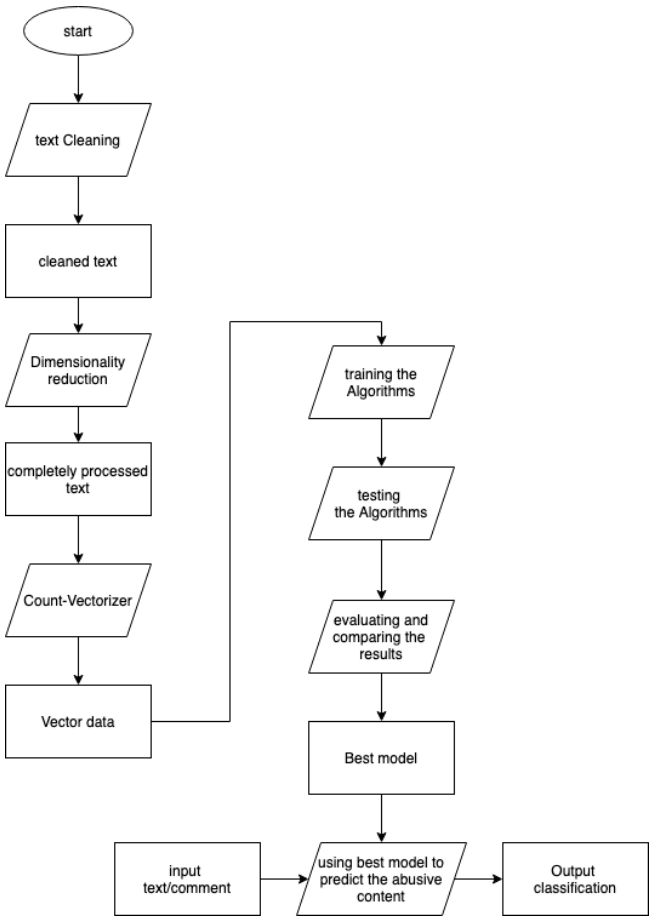

Fig 2: System architecture

B. IMPLEMENTATION

According to the literature survey this model description provides the overall information about its supposed components, which is accessible in various manners.The comment detection is performed using various algorithms based on the accuracy of the models the best model is choosed.The models used here are

1. Random Forest classifier

In random forest classifier the decision trees are built. It is an ensemble algorithm which uses weak worker and group them and become stronger as whole. Random forest, like its name implies, consists of a large number of individual decision trees forest spits out a class prediction and the class with the most votes become our model's prediction.The fundamental concept behind random forest is a simple but powerful one — the wisdom of crowds. A large number of relatively uncorrelated models (trees) operating as a committee will outperform any of the individual constituent models.The low correlation between models is the key,uncorrelated models can produce ensemble predictions that are more accurate than any of the individual predictions. The reason is that the trees protect each other from their individual errors. While some trees may be wrong, many other trees will be right, so as a group the trees are able to move in the correct direction.

2. Artificial Neural Network(ANN)

An Artificial neural network is usually a computational network based on biological neural networks. Similar to a human brain has neurons interconnected to each other, artificial neural networks also have neurons that are linked to each other in various layers of the networks. These neurons are known as nodes.There are various types of layers in ANN

Input layer - This layer takes the input from the users.

Hidden layer - It performs the calculations to find hidden features and patterns.

Output layer - This layer is responsible for producing final result.

The Ann can be implemented using 2 propagations

Forward propagation - It refers to the storage and calculation of the intermediate variables in order from input layer to output layer.

Backward propagation -The process of moving backward i.e;from output layer to input layer.

This system thus developed helps in the detection of the abusive and insulting comments on the social media thus helping the user to take the necessary action against this cyber bullying.

IV. RESULTS

The proposed system uses GUI as the interface to interact with the client. The model which we developed gave an accuracy of 83%. The remaining 17% is failed due to the system requirements and the huge amounts of the training data required. Some failed cases are -The comments which may be abusive or insulting may not be actually intended to a particular person but our model cannot differentiate them. Some of the comments may be insulting but cannot be identified by the model as these may not cantain abusive text.

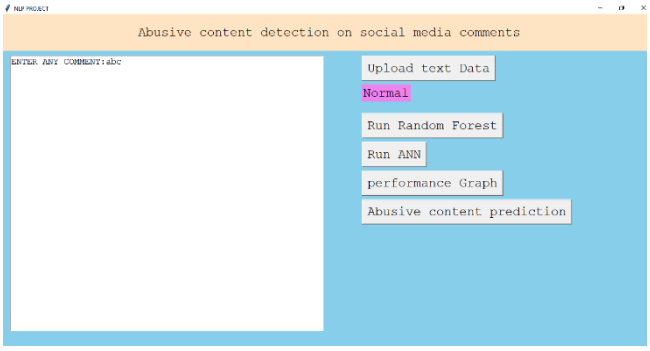

Fig 3 : Normal text

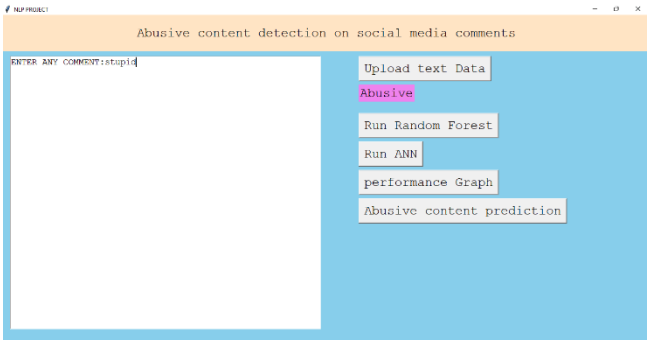

Fig 4 : Abusive text

V. CONCLUSION

The model thus developed helps in the cyber bullying detection. The project is developed with the aim of providing security in social media. By this model people can identify the insulting and abusive comments received and can take the necessary actions accordingly by reporting or by blocking the sender. This is an effective tool which helps people to express their views and opinions.

VI. REFERENCES

[1] Alkula, R. From plain character strings to meaningful words: Producing better full text databases for inflectional and compounding languages with morphological analysis software. Information Retrieval, 4, (2001), 195-208.

[2] Krovetz, R. Viewing morphology as an inference process. In Proceedings of the Sixteenth Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval (SIGIR'03) (Pittsburg, PA, 27 June - 1 July 1993). ACM Press, New York, NY, 1993, 191-202.

[3] Pirkola, A. Morphological typology of languages for information retrieval. Journal of Documentation, 57, 3 (2001), 330-348.

[4] Harman D. How effective is suffixing? Journal of the American Society for Information Science, 42, 1 (1991), 7-15.

[5] Hull, D. Stemming algorithms: a case study for detailed evaluation. Journal of the American Society for Information Science, 47, 1 (1996), 70-84.

[6] Popovic, M., and Willett, P. The effectiveness of stemming for natural-language access to Slovene textual data. Journal of the American Society for Information Science, 43, 1 (1992), 384-390.
[7] Savoy, J. A stemming procedure and stopword list for general French corpora. Journal of the American Society for Information Science, 50, 10 (1999), 944-952.
[8] Kalamboukis, T. Z. Suffix stripping with modern Greek. Program, 29, 3 (1995), 313-321.
[9] Abu-Salem, H., Al-Omari, M., and Evens, M. W. Stemming methodologies over individual query words for an Arabic information retrieval system. Journal of the American Society for Information Science, 50, 6 (1999), 524-529.
[10] Matthews, P. H. The Concise Oxford Dictionary of Linguistics. Oxford University Press, Oxford - New York, NY, 1997.
[11] Karlsson, F. Finnish grammar. WSOY, Porvoo, 1987.
[12] Koskenniemi, K. An application of the two-level model to Finnish. In Computational morphosyntax: Report on research 1981-84. Publications 13, University of Helsinki, Department of General Linguistics, Helsinki, 1985, 19-41.
[13] Koskenniemi, K. Two-level morphology: A general computational model for word-form recognition and production. Publications 11, University of Helsinki, Department of General Linguistics, Helsinki, 1983.